



Zhu, Qiangqiang (2026) *Bayesian spatio-temporal modelling of air pollution and disease risk with spatially misaligned data*. PhD thesis.

<https://theses.gla.ac.uk/86158/>

Copyright and moral rights for this work are retained by the author

A copy can be downloaded for personal non-commercial research or study, without prior permission or charge

This work cannot be reproduced or quoted extensively from without first obtaining permission from the author

The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author

When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given

Enlighten: Theses

<https://theses.gla.ac.uk/>  
[research-enlighten@glasgow.ac.uk](mailto:research-enlighten@glasgow.ac.uk)

# **Bayesian spatio-temporal modelling of air pollution and disease risk with spatially misaligned data**

Qiangqiang Zhu

Submitted in fulfilment of the requirements for the  
Degree of Doctor of Philosophy

School of Mathematics and Statistics  
College of Science and Engineering  
University of Glasgow



April 2026



# Abstract

The spatial support at which environmental and health data are observed often does not coincide with the support at which inference is required. In Scotland, air quality is measured at a sparse monitoring network, respiratory prescribing is recorded at general practitioner (GP) surgery level, and small-area population and socio-economic covariates are available on Data Zones (DZs). Assigning exposures to populations, realigning GP-level outcomes, and mapping small-area disease risk each require the analyst to bridge supports that do not align by design. This thesis develops statistical methodology for the combined predictions of air pollution exposures, estimation of the effects of air pollution on respiratory ill health treated in primary care, and the small-area risk estimation of respiratory ill health treated in primary care where the supports of the underlying data are partially unknown or misaligned.

The first methodological strand addresses exposure prediction. Statistical and machine learning candidate approaches are compared under spatial cross-validation, with random forests selected for  $\text{NO}_2$  and  $\text{PM}_{2.5}$  and a linear model for  $\text{PM}_{10}$ . The resulting monthly  $1 \text{ km}^2$  prediction surfaces cover Scotland from 2016 to 2020 with associated 95% prediction intervals, and provide the exposure resource used in the subsequent chapters.

The second study concerns the estimation of the health effect that air pollution has at the GP surgery level. Two population informed neighbourhood matrices are introduced to model the spatial autocorrelation in the data, which define spatial closeness between GP patient catchments through the overlap of their registered patient populations rather than the geographic distance between surgery locations. Both substantially reduce residual spatial autocorrelation relative to a 5-nearest-neighbours baseline definition based on surgery locations. Air pollution exposures are also assigned to each GP surgery from the residential distribution of its registered patients rather than using concentrations at the surgery location. Under this specification, a consistent positive association is identified between  $\text{PM}_{10}$  and preventer medication prescribing, with an estimated 0.5% increase in prescribing rates per one

standard deviation increase in  $PM_{10}$ . This finding is robust to full propagation of the predictive uncertainty from the exposure model.

The third study reverses the direction of the support problem, estimating relative risks on Data Zones from counts observed on GP surgeries. A Bayesian spatio-temporal disease mapping framework links GP-level observations to a latent DZ-level risk surface through a population-weighted change-of-support mechanism, adjusting an existing multiple-membership formulation to the relative-risk scale. Two simulation studies show that this alignment model consistently lowers the root mean squared error of DZ-level risk estimates relative to naive interpolation across plausible data generating scenarios. Applied to Glasgow City over 2019–2020 with a socio-economic covariate, the framework recovers a persistent and monotonic deprivation gradient in respiratory prescribing risk, with time-averaged posterior median relative risks in the most-deprived SIMD 2020 decile approximately four times those in the least-deprived decile. The estimated surface also identifies a city-wide spike in March 2020 coinciding with the onset of the Covid-19 pandemic.

Taken together, the thesis shows that spatial support mismatch is a primary modelling concern that shapes inference at each stage from exposure reconstruction to small-area risk estimation. Modelling the supports explicitly yields credible exposure predictions, pollution–health effect estimates, and small-area risk maps of respiratory ill health treated in primary care resolved below the level at which prescribing counts are recorded. The resulting estimates support more targeted environmental and public health action on respiratory burden in Scotland at the small-area scale relevant to policy.

# Contents

|                                                                           |              |
|---------------------------------------------------------------------------|--------------|
| <b>Abstract</b>                                                           | <b>i</b>     |
| <b>Acknowledgements</b>                                                   | <b>xvii</b>  |
| <b>Declaration</b>                                                        | <b>xviii</b> |
| <b>1 Introduction</b>                                                     | <b>1</b>     |
| 1.1 Respiratory ill health in primary care . . . . .                      | 3            |
| 1.2 Small-area disease mapping . . . . .                                  | 4            |
| 1.3 The unifying challenge . . . . .                                      | 5            |
| 1.4 Study region and data . . . . .                                       | 7            |
| <b>2 Statistical background</b>                                           | <b>9</b>     |
| 2.1 Classic regression and GLMs . . . . .                                 | 9            |
| 2.1.1 Linear regression . . . . .                                         | 10           |
| 2.1.2 Generalised linear models . . . . .                                 | 11           |
| 2.1.3 Multicollinearity and the variance inflation factor . . . . .       | 12           |
| 2.2 Smooth functions and additive models . . . . .                        | 13           |
| 2.2.1 Regression splines . . . . .                                        | 14           |
| 2.2.2 Generalised additive models . . . . .                               | 17           |
| 2.2.3 Generalised additive models for location, scale and shape . . . . . | 18           |
| 2.3 Tree-based machine learning methods . . . . .                         | 19           |
| 2.3.1 Regression trees . . . . .                                          | 20           |
| 2.3.2 Bagging of regression trees . . . . .                               | 21           |
| 2.3.3 Random forests . . . . .                                            | 21           |
| 2.4 Bayesian hierarchical modelling . . . . .                             | 23           |
| 2.4.1 Introduction . . . . .                                              | 23           |
| 2.4.2 Prior distributions . . . . .                                       | 24           |
| 2.4.3 Bayesian inference . . . . .                                        | 25           |

|          |                                                                         |           |
|----------|-------------------------------------------------------------------------|-----------|
| 2.5      | Model comparison . . . . .                                              | 31        |
| 2.5.1    | Akaike information criterion . . . . .                                  | 31        |
| 2.5.2    | Bayesian information criterion . . . . .                                | 32        |
| 2.5.3    | Deviance information criterion . . . . .                                | 32        |
| 2.5.4    | Watanabe-Akaike information criterion . . . . .                         | 33        |
| 2.6      | Spatial and spatio-temporal modelling . . . . .                         | 33        |
| 2.6.1    | Defining spatial closeness . . . . .                                    | 34        |
| 2.6.2    | Spatial autocorrelation . . . . .                                       | 36        |
| 2.6.3    | Spatio-temporal Gaussian process models . . . . .                       | 36        |
| 2.6.4    | Disease mapping . . . . .                                               | 38        |
| 2.6.5    | Spatial models for areal unit data . . . . .                            | 40        |
| 2.6.6    | Spatio-temporal models for areal unit data . . . . .                    | 44        |
| <b>3</b> | <b>Air pollution modelling</b>                                          | <b>47</b> |
| 3.1      | Introduction . . . . .                                                  | 47        |
| 3.2      | Literature review of modelling techniques in air pollution prediction . | 50        |
| 3.2.1    | Linear models . . . . .                                                 | 51        |
| 3.2.2    | Generalised additive models . . . . .                                   | 52        |
| 3.2.3    | Hierarchical spatio-temporal models . . . . .                           | 53        |
| 3.2.4    | Machine learning . . . . .                                              | 54        |
| 3.3      | Data and exploratory analysis . . . . .                                 | 55        |
| 3.3.1    | Study region and data collection . . . . .                              | 55        |
| 3.3.2    | Monitoring pollution concentrations . . . . .                           | 56        |
| 3.3.3    | Modelled pollution concentrations . . . . .                             | 61        |
| 3.3.4    | Meteorological predictors . . . . .                                     | 64        |
| 3.3.5    | NDVI and land use data . . . . .                                        | 65        |
| 3.3.6    | Population size and traffic conditions . . . . .                        | 68        |
| 3.4      | Methods for predicting ambient air pollutant concentrations . . . . .   | 70        |
| 3.4.1    | Normal linear models . . . . .                                          | 71        |
| 3.4.2    | Additive models . . . . .                                               | 72        |
| 3.4.3    | Hierarchical spatio-temporal models . . . . .                           | 73        |
| 3.4.4    | Random forests . . . . .                                                | 74        |
| 3.5      | Study design . . . . .                                                  | 76        |
| 3.5.1    | Spatial cross-validation . . . . .                                      | 76        |
| 3.5.2    | Performance metrics . . . . .                                           | 76        |
| 3.5.3    | Implementation details . . . . .                                        | 77        |

|          |                                                                           |           |
|----------|---------------------------------------------------------------------------|-----------|
| 3.5.4    | Construction of predictive uncertainty . . . . .                          | 78        |
| 3.5.5    | Hyperparameter tuning for random forests . . . . .                        | 79        |
| 3.6      | Model comparison for the three air pollutants . . . . .                   | 80        |
| 3.7      | Results from the best performing prediction models . . . . .              | 81        |
| 3.7.1    | Spatio-temporal trends in predicted concentrations . . . . .              | 81        |
| 3.7.2    | Long-term changes in concentrations . . . . .                             | 88        |
| 3.7.3    | Predictive uncertainty quantification . . . . .                           | 89        |
| 3.8      | Discussion . . . . .                                                      | 89        |
| <b>4</b> | <b>Health effects in primary care</b>                                     | <b>92</b> |
| 4.1      | Introduction . . . . .                                                    | 92        |
| 4.2      | Study region and data . . . . .                                           | 96        |
| 4.2.1    | Data sources and spatial supports . . . . .                               | 97        |
| 4.2.2    | GP data . . . . .                                                         | 99        |
| 4.2.3    | Air pollution data . . . . .                                              | 102       |
| 4.2.4    | Socio-economic deprivation . . . . .                                      | 105       |
| 4.2.5    | Meteorological data . . . . .                                             | 107       |
| 4.2.6    | Ethnicity . . . . .                                                       | 108       |
| 4.3      | Spatial realignment . . . . .                                             | 109       |
| 4.4      | Exploratory analysis using a negative binomial model . . . . .            | 111       |
| 4.4.1    | Data likelihood model . . . . .                                           | 111       |
| 4.4.2    | Selecting the confounders . . . . .                                       | 112       |
| 4.4.3    | Measuring residual spatial correlation . . . . .                          | 113       |
| 4.4.4    | Results from the negative binomial models . . . . .                       | 115       |
| 4.5      | Spatio-temporal smoothing model with the CAR-AR(1) structure . . .        | 116       |
| 4.5.1    | Data likelihood model . . . . .                                           | 116       |
| 4.5.2    | Spatio-temporal random effects model . . . . .                            | 118       |
| 4.5.3    | Neighbourhood matrix . . . . .                                            | 119       |
| 4.5.4    | Selecting priors for the spatio-temporal CAR-AR(1) component              | 124       |
| 4.6      | Results of the motivating study . . . . .                                 | 128       |
| 4.6.1    | Assessment of the residual spatial correlation . . . . .                  | 128       |
| 4.6.2    | Results of the non-binary neighbourhood matrices . . . . .                | 129       |
| 4.6.3    | Pollution-health associations . . . . .                                   | 131       |
| 4.6.4    | Results for PM <sub>10</sub> lagged by one month . . . . .                | 137       |
| 4.6.5    | Results for PM <sub>10</sub> incorporating exposure uncertainty . . . . . | 138       |
| 4.7      | Discussion . . . . .                                                      | 139       |

|          |                                                                |            |
|----------|----------------------------------------------------------------|------------|
| <b>5</b> | <b>Bayesian disease mapping</b>                                | <b>144</b> |
| 5.1      | Introduction . . . . .                                         | 144        |
| 5.2      | Motivating study . . . . .                                     | 147        |
| 5.2.1    | Study data . . . . .                                           | 147        |
| 5.2.2    | Aims and motivating questions . . . . .                        | 149        |
| 5.3      | Methodology . . . . .                                          | 151        |
| 5.3.1    | Data model and change of support . . . . .                     | 151        |
| 5.3.2    | Spatio-temporal random effects . . . . .                       | 153        |
| 5.3.3    | Prior specification and posterior inference . . . . .          | 154        |
| 5.3.4    | Computational considerations . . . . .                         | 156        |
| 5.4      | Simulation study 1: spatial realignment . . . . .              | 158        |
| 5.4.1    | Model specification . . . . .                                  | 158        |
| 5.4.2    | Competitor models . . . . .                                    | 159        |
| 5.4.3    | Data generation and simulation scenarios . . . . .             | 160        |
| 5.4.4    | Performance metrics . . . . .                                  | 163        |
| 5.4.5    | Results of simulation study 1 . . . . .                        | 165        |
| 5.5      | Simulation study 2: spatio-temporal realignment . . . . .      | 166        |
| 5.5.1    | Model specification . . . . .                                  | 167        |
| 5.5.2    | Competitor models . . . . .                                    | 169        |
| 5.5.3    | Data generation and simulation scenarios . . . . .             | 171        |
| 5.5.4    | Results of simulation study 2 . . . . .                        | 174        |
| 5.6      | Motivating study results . . . . .                             | 177        |
| 5.6.1    | Model specification . . . . .                                  | 177        |
| 5.6.2    | Model checking . . . . .                                       | 178        |
| 5.6.3    | Spatial distribution of relative risk . . . . .                | 179        |
| 5.6.4    | Temporal patterns and the Covid-19 prescribing spike . . . . . | 180        |
| 5.6.5    | Health inequalities by deprivation . . . . .                   | 182        |
| 5.7      | Discussion . . . . .                                           | 184        |
| <b>6</b> | <b>Conclusions</b>                                             | <b>187</b> |
| 6.1      | Summary of contributions . . . . .                             | 187        |
| 6.2      | Common methodological themes . . . . .                         | 190        |
| 6.3      | Limitations . . . . .                                          | 191        |
| 6.4      | Future work . . . . .                                          | 193        |

|          |                                                          |            |
|----------|----------------------------------------------------------|------------|
| <b>A</b> | <b>CW-MH sampler for the ICAR-AR(1) model</b>            | <b>195</b> |
| A.1      | Innovation decomposition and sampling targets . . . . .  | 195        |
| A.2      | Parameter blocking and sampler assignment . . . . .      | 196        |
| A.3      | Component-wise Metropolis–Hastings algorithm . . . . .   | 197        |
| A.4      | Comparison with the default CAR_normal sampler . . . . . | 200        |
| A.5      | Proposal adaptation and identifiability . . . . .        | 200        |

# List of Tables

|     |                                                                                                                                                                                                                                                                                                                                                           |     |
|-----|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|
| 3.1 | Summary statistics of the monitoring site (measured) data and the modelled concentration data for NO <sub>2</sub> , PM <sub>10</sub> and PM <sub>2.5</sub> . The measured data (left) are daily averages, while the modelled data (right) are yearly averages for those 1 km <sup>2</sup> grid squares that contain at least one monitoring site. . . . . | 58  |
| 3.2 | Meteorological predictors derived from HadUK-Grid and used in the modelling. . . . .                                                                                                                                                                                                                                                                      | 65  |
| 3.3 | Aggregated (three-class) Scottish Urban Rural Classification used in this study. . . . .                                                                                                                                                                                                                                                                  | 68  |
| 3.4 | Summary of the candidate models compared in this study. . . . .                                                                                                                                                                                                                                                                                           | 78  |
| 3.5 | Optimal hyperparameters selected via 10-fold cross-validation for the random forest models. . . . .                                                                                                                                                                                                                                                       | 80  |
| 3.6 | Predictive model comparison for NO <sub>2</sub> , PM <sub>10</sub> and PM <sub>2.5</sub> , averaged across ten spatial cross-validation splits. . . . .                                                                                                                                                                                                   | 82  |
| 4.1 | The number of GP surgeries in each health board. . . . .                                                                                                                                                                                                                                                                                                  | 98  |
| 4.2 | Datasets used in the health analysis, with their spatial support and temporal resolution. All variables are realigned onto the GP-catchment support and to monthly resolution before modelling (Section 4.3). . .                                                                                                                                         | 99  |
| 4.3 | Summary of the indicators included in the 2020 Scottish Index of Multiple Deprivation (SIMD) that are used in this study. . . . .                                                                                                                                                                                                                         | 106 |
| 4.4 | Summary statistics for the number of non-zero weights in the neighbourhood matrix $\mathbf{W}^{(o)}$ at different threshold values of $\epsilon$ for the data spans 2016–2020 and 2016–2019. . . . .                                                                                                                                                      | 124 |
| 4.5 | Summary statistics for the number of neighbours for each GP surgery in the binary matrix $\mathbf{W}^{(s)}$ at different threshold values of $\omega$ for the data spans 2016–2020 and 2016–2019. . . . .                                                                                                                                                 | 124 |



|      |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  |     |
|------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|
| 4.6  | Posterior median relative risk estimates, 95% credible intervals and exceedance probabilities from models based on $\{\mathbf{W}^{(5nn)}, \mathbf{W}^{(O)}, \mathbf{W}^{(S)}\}$ quantifying the impacts of $\text{NO}_2$ (based on an increase of $6.34 \mu\text{g}/\text{m}^3$ ) on respiratory prescription rates for all three medication types and both study periods. The thresholds used for each of $\{\mathbf{W}^{(O)}, \mathbf{W}^{(S)}\}$ are also given. . . . .      | 135 |
| 4.7  | Posterior median relative risk estimates, 95% credible intervals and exceedance probabilities from models based on $\{\mathbf{W}^{(5nn)}, \mathbf{W}^{(O)}, \mathbf{W}^{(S)}\}$ quantifying the impacts of $\text{PM}_{10}$ (based on an increase of $2.77 \mu\text{g}/\text{m}^3$ ) on respiratory prescription rates for all three medication types and both study periods. The thresholds used for each of $\{\mathbf{W}^{(O)}, \mathbf{W}^{(S)}\}$ are also given. . . . .   | 136 |
| 4.8  | Posterior median relative risk estimates, 95% credible intervals and exceedance probabilities from models based on $\{\mathbf{W}^{(5nn)}, \mathbf{W}^{(O)}, \mathbf{W}^{(S)}\}$ quantifying the impacts of $\text{PM}_{2.5}$ (based on an increase of $1.68 \mu\text{g}/\text{m}^3$ ) on respiratory prescription rates for all three medication types and both study periods. The thresholds used for each of $\{\mathbf{W}^{(O)}, \mathbf{W}^{(S)}\}$ are also given. . . . .  | 137 |
| 4.9  | Posterior median relative risk estimates, 95% credible intervals and exceedance probabilities from models based on $\{\mathbf{W}^{(5nn)}, \mathbf{W}^{(O)}, \mathbf{W}^{(S)}\}$ quantifying the impacts of Lag-1 $\text{PM}_{10}$ (based on an increase of $2.77 \mu\text{g}/\text{m}^3$ ) on respiratory prescription rates for all 3 medication types and both study periods. The thresholds used for each of $\{\mathbf{W}^{(O)}, \mathbf{W}^{(S)}\}$ are also given. . . . . | 138 |
| 4.10 | Posterior median relative risk estimates, 95% credible intervals and exceedance probabilities from models quantifying the effect of $\text{PM}_{10}$ (based on an increase of $2.77 \mu\text{g}/\text{m}^3$ ) on preventer medications while incorporating uncertainty in the $\text{PM}_{10}$ concentrations. . . . .                                                                                                                                                           | 139 |

|     |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           |     |
|-----|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|
| 5.1 | Mean performance metrics across 50 replicates for Simulation Study 1, Setting 1 (no covariates). Metrics are averaged over all $M = 746$ Data Zones. CP and MIW are not applicable for Naive as it does not produce posterior distributions. The rows compare the Naive model (deterministic population-weighted interpolation), Realign (the proposed change-of-support model), and Benchmark (an idealised model fitted directly to DZ-level data). Prevalence is either Common or Rare. . . . .                                                                                                                                                        | 167 |
| 5.2 | Mean performance metrics across 50 replicates for Simulation Study 1, Setting 2 (with spatially structured covariate). Metrics are averaged (mean) over all $M = 746$ Data Zones. CP and MIW are not applicable for Naive as it does not produce posterior distributions. The rows compare the Naive model (deterministic population-weighted interpolation), Realign (the proposed change-of-support model), and Benchmark (an idealised model fitted directly to DZ-level data). . . . .                                                                                                                                                                | 168 |
| 5.3 | Mean performance metrics across 50 replicates for Simulation Study 2. Metrics are averaged (mean) over all $M = 746$ Data Zones and $T = 12$ months. CP and MIW are not applicable for Naive0 and Naive1 as they do not produce posterior distributions. The temporal scenarios are T1 (global linear trend), T2 (AR(1) persistence) and T3 (widening inequality). The rows compare Naive0 and Naive1 (deterministic interpolation using the first-year and time-varying registration weights respectively), Realign (the proposed spatio-temporal change-of-support model), and Benchmark (an idealised model fitted directly to DZ-level data). . . . . | 175 |
| 5.4 | Scalar posterior summaries for the ICAR-AR(1) model with IncRate covariate in the motivating study. Point estimates are posterior medians, and 95% credible intervals are the 2.5th and 97.5th posterior percentiles. . . . .                                                                                                                                                                                                                                                                                                                                                                                                                             | 177 |

# List of Figures

|      |                                                                                                                                                                                                                                                                                              |    |
|------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| 3.1  | Map showing the air pollution monitoring sites in Scotland, which are colour-coded according to the type of local environment in which they reside. . . . .                                                                                                                                  | 57 |
| 3.2  | Heat map summarising the presence or absence of NO <sub>2</sub> concentration data by day and site. . . . .                                                                                                                                                                                  | 59 |
| 3.3  | Heat map summarising the presence or absence of PM <sub>10</sub> concentration data by day and site. . . . .                                                                                                                                                                                 | 59 |
| 3.4  | Heat map summarising the presence or absence of PM <sub>2.5</sub> concentration data by day and site. . . . .                                                                                                                                                                                | 60 |
| 3.5  | Monthly temporal trends in the monitored concentrations of NO <sub>2</sub> , PM <sub>10</sub> and PM <sub>2.5</sub> via box plots and medians (across sites). . . . .                                                                                                                        | 61 |
| 3.6  | Line charts showing the seasonality of monthly mean concentrations of NO <sub>2</sub> , PM <sub>10</sub> and PM <sub>2.5</sub> by year. . . . .                                                                                                                                              | 62 |
| 3.7  | Box plots of annual PCM modelled concentrations (NO <sub>2</sub> , PM <sub>10</sub> and PM <sub>2.5</sub> ) for grid cells that contain at least one monitoring site (2016–2020). . . . .                                                                                                    | 63 |
| 3.8  | Spatial distributions of the temporal mean (between 2016 and 2020) of the 1 km <sup>2</sup> estimates of NO <sub>2</sub> , PM <sub>10</sub> and PM <sub>2.5</sub> from the Pollution Climate (PCM) Mapping model. . . . .                                                                    | 64 |
| 3.9  | Box plots of monthly meteorological predictors (relative humidity, sea-level pressure, 10 m wind speed, sunshine hours, total precipitation and air temperature) at the monitoring-site locations over 2016–2020. Source: HadUK-Grid. . . . .                                                | 66 |
| 3.10 | Spatial distribution of the five-year mean (2016–2020) of each monthly meteorological predictor (relative humidity, sea-level pressure, 10 m wind speed, sunshine hours, total precipitation and air temperature) across Scotland on the 1 km <sup>2</sup> grid. Source: HadUK-Grid. . . . . | 67 |

|                                                                                                                                                                                                                                                                                                                                                                                                     |     |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|
| 3.11 (a) The aggregated three-class Scottish Urban Rural Classification and (b) the five-year mean (2016–2020) NDVI, both mapped on the 1 km <sup>2</sup> grid. . . . .                                                                                                                                                                                                                             | 68  |
| 3.12 Spatial distribution of population density (normalised) across Scotland, derived from National Records of Scotland mid-year population estimates and harmonised to the 1 km <sup>2</sup> prediction grid (left panel), together with the gridded distance from each 1 km <sup>2</sup> grid cell to the nearest major road (in km), derived from OpenStreetMap road data (right panel). . . . . | 70  |
| 3.13 Scatter plots of the observed versus predicted NO <sub>2</sub> concentrations across ten test sets. . . . .                                                                                                                                                                                                                                                                                    | 83  |
| 3.14 Scatter plots of the observed versus predicted PM <sub>10</sub> concentrations across ten test sets. . . . .                                                                                                                                                                                                                                                                                   | 84  |
| 3.15 Scatter plots of the observed versus predicted PM <sub>2.5</sub> concentrations across ten test sets. . . . .                                                                                                                                                                                                                                                                                  | 85  |
| 3.16 Line plots of the mean (over all 1 km <sup>2</sup> grid squares in Scotland) monthly predicted concentrations of NO <sub>2</sub> , PM <sub>10</sub> and PM <sub>2.5</sub> between 2016 and 2020. . . . .                                                                                                                                                                                       | 86  |
| 3.17 Predicted seasonal mean concentrations on the 1 km <sup>2</sup> grid over 2016–2020 for (a) NO <sub>2</sub> , (b) PM <sub>10</sub> and (c) PM <sub>2.5</sub> . Spring: {March, April, May}; summer: {June, July, August}; autumn: {September, October, November}; and winter: {December, January, February}. . . . .                                                                           | 87  |
| 3.18 Estimated slopes of the temporal trend for NO <sub>2</sub> , PM <sub>10</sub> and PM <sub>2.5</sub> , obtained by regressing annual mean predicted concentration on normalised calendar year at each 1 km <sup>2</sup> grid square. Negative values indicate decreasing concentrations. . . . .                                                                                                | 88  |
| 3.19 Average widths of the monthly 95% prediction intervals on the 1 km <sup>2</sup> grid over the period 2016–2020 for NO <sub>2</sub> , PM <sub>10</sub> and PM <sub>2.5</sub> . . . . .                                                                                                                                                                                                          | 89  |
| 4.1 Map showing the distribution of GP surgeries across Scotland that were operational in any month between 2016 and 2020, colour-coded by Health Board. . . . .                                                                                                                                                                                                                                    | 97  |
| 4.2 National average prevalence $\{r_j\}$ of asthma and COPD by age and sex in Scotland, obtained from the SPIRE database. . . . .                                                                                                                                                                                                                                                                  | 101 |

|      |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    |     |
|------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|
| 4.3  | Box plots of standardised prescription ratios (SPRs) across all GP surgeries in Scotland for each month between January 2016 and December 2020. . . . .                                                                                                                                                                                                                                                                                                                                                                                                                                                                            | 103 |
| 4.4  | Line plot showing the monthly variation in the mean standardised prescription ratio (SPR) over all GP surgeries in Scotland, 2016–2020. . . . .                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    | 103 |
| 4.5  | Spatial distribution of mean standardised prescription ratios (SPRs) for chronic respiratory medication prescriptions across GP surgeries in Scotland, averaged over the period 2016–2020. . . . .                                                                                                                                                                                                                                                                                                                                                                                                                                 | 104 |
| 4.6  | Histograms and density plots of the SIMD indicators at the DZ scale. . . . .                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       | 107 |
| 4.7  | Correlation plot of the SIMD indicators at the DZ scale. . . . .                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   | 108 |
| 4.8  | Correlation plot of the ethnic groups: White, South Asian descent, Chinese descent, and Black at the DZ scale. . . . .                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             | 109 |
| 4.9  | Moran’s I values for the standardised residuals from the negative binomial models with $\text{NO}_2$ as the pollutant, by medication type (rows) and time period (blue: 2016–2020; red: 2016–2019). A 5-nearest-neighbour spatial weights matrix based on GP surgery locations is used to compute Moran’s I. . . . .                                                                                                                                                                                                                                                                                                               | 116 |
| 4.10 | Lag-1 autocorrelation ( $\text{ACF}(1)$ ) of the negative binomial model residuals for $\text{NO}_2$ , by medication type (rows) and time period (columns). Each point represents $\text{ACF}(1)$ for a single GP surgery. The dashed red lines indicate approximate 95% significance bounds under the null of no temporal autocorrelation, and the red text gives the proportion of surgeries with significant $\text{ACF}(1)$ values. . . . .                                                                                                                                                                                    | 117 |
| 4.11 | Schematic comparison of the three neighbourhood constructions. Patients live across the Data Zones that form each catchment rather than at the surgery buildings, and the catchments of $S_i$ and $S_j$ share the Data Zone outlined in dashed purple. The distance-based $\mathbf{W}^{(5nn)}$ links $S_i$ and $S_k$ because their buildings are close together (a), whereas the population-overlap $\mathbf{W}^{(o)}$ instead links $S_i$ and $S_j$ because their patient populations share Data Zones (b), and $\mathbf{W}^{(s)}$ links both pairs since it also counts catchments lying in neighbouring Data Zones (c). . . . . | 120 |
| 4.12 | Probability density function (PDF) curves for the prior distributions assumed for the following parameters: (a) the standard deviation $\sigma$ ; (b) the spatial dependence $\rho$ ; and (c) the temporal correlation $\alpha$ . . . . .                                                                                                                                                                                                                                                                                                                                                                                          | 125 |

|      |                                                                                                                                                                                                                                                                                                                                                                                                    |     |
|------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|
| 4.13 | Moran's I statistics quantifying the level of spatial correlation in the residuals for each month when using $\mathbf{W}^{(5nn)}$ as the neighbourhood matrix. Each panel represents a different disease outcome, while red lines relate to the 2016–2019 study period and the blue lines relate to 2016–2020. PSAI, GSAI and MAI indicators are listed in each case. .                            | 130 |
| 4.14 | Moran's I statistics quantifying the level of spatial correlation in the residuals for each month when using $\mathbf{W}^{(O)}$ as the neighbourhood matrix. Each panel represents a different disease outcome and time period, while the coloured lines relate to the different thresholds $\varepsilon$ . PSAI, GSAI and MAI indicators are listed in each case. . . . .                         | 131 |
| 4.15 | Moran's I statistics quantifying the level of spatial correlation in the residuals for each month when using $\mathbf{W}^{(S)}$ as the neighbourhood matrix. Each panel represents a different disease outcome and time period, while the coloured lines relate to the different thresholds $\omega$ . PSAI, GSAI and MAI indicators are listed in each case. . . . .                              | 132 |
| 4.16 | Moran's I statistics quantifying the level of spatial correlation in the residuals for each month when using a non-binary version of $\mathbf{W}^{(O)}$ as the neighbourhood matrix. Each panel represents a different disease outcome and time period, while the coloured lines relate to the different thresholds $\varepsilon$ . PSAI, GSAI and MAI indicators are listed in each case. . . . . | 133 |
| 4.17 | Moran's I statistics quantifying the level of spatial correlation in the residuals for each month when using a non-binary version of $\mathbf{W}^{(S)}$ as the neighbourhood matrix. Each panel represents a different disease outcome and time period, while the coloured lines relate to the different thresholds $\omega$ . PSAI, GSAI and MAI indicators are listed in each case. . . . .      | 134 |
| 5.1  | Socio-economic context of the Glasgow City study region. Panel (A) maps SIMD 2020 deprivation deciles across the $M = 746$ Data Zones (decile 1 denotes the most deprived 10% of Scottish Data Zones); Panel (B) shows the distribution of Glasgow City Data Zones by SIMD decile. . . . .                                                                                                         | 148 |

|     |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   |     |
|-----|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|
| 5.2 | Summary of GP-level prescription SPRs in Glasgow City, January 2019 to December 2020. Panel (A) shows box plots of GP-level SPRs over time for the 141 surgeries retained in 2019 and 139 retained in 2020 by the 75% city-fraction rule, with the line tracking the monthly median. Panel (B) shows the distribution of time-averaged GP-level SPRs across the 142 distinct surgeries that satisfy the 75% rule in at least one of the two study years, with the vertical line marking the median. The union count (142) slightly exceeds each within-year count because a small number of surgeries opened, closed, or crossed the 75% threshold between the two years. . . . . | 150 |
| 5.3 | Registration-weight structure between GP surgeries and Data Zones, evaluated on the $K = 260$ surgeries with any registered Glasgow patient using the 2020 registration snapshot and the sparsification threshold $\varepsilon = 0.01$ used in the simulation studies. Panel (A) shows the distribution of the number of Data Zones with a non-zero registration weight for each surgery before (grey) and after (blue) applying the $\eta_{km} \geq 0.01$ threshold. Panel (B) shows the distribution across surgeries of the share of each surgery's total registration weight that is retained after sparsification. . . . .                                                   | 157 |
| 5.4 | Between-replicate distribution of root mean squared error (RMSE) of posterior median DZ-level relative risks, Simulation Study 1 Setting 1 (no covariate). Panels are arranged by spatial correlation range ( $r$ ) and marginal variability ( $\sigma_\phi$ ), with rows separating Rare and Common disease prevalence. . . . .                                                                                                                                                                                                                                                                                                                                                  | 169 |
| 5.5 | Scatter plot of true simulated versus estimated DZ-level relative risks under a rougher ( $r = 0.5$ ), higher-variance ( $\sigma_\phi = 0.5$ ) spatial surface. Panel (A) shows relative risks on the original scale and Panel (B) shows the same comparison on the log scale. The mean bias of the posterior median estimates across the 746 Data Zones for this replicate is $-0.07$ .170                                                                                                                                                                                                                                                                                       |     |
| 5.6 | Between-replicate distribution of RMSE of posterior median DZ-level relative risks, Simulation Study 1 Setting 2 (with spatially structured covariate). Panels are arranged by covariate variance share $p$ and marginal variability $\sigma_\phi$ , with rows separating Rare and Common disease prevalence. . . . .                                                                                                                                                                                                                                                                                                                                                             | 171 |

|      |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   |     |
|------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|
| 5.7  | Between-replicate distribution of RMSE of posterior median DZ-level relative risks across 50 replicates per scenario, Simulation Study 2 (spatio-temporal). Panels are arranged by covariate variance share $p$ (rows) and temporal scenario (columns: T1 global linear trend, T2 AR(1) persistence, T3 widening inequality). . . . .                                                                                                                                                                             | 176 |
| 5.8  | Posterior predictive checks for the ICAR-AR(1) model with IncRate covariate in the motivating study, showing (a) observed versus fitted (posterior median) GP-month prescription counts on logarithmic axes, with the 45° line in grey; (b) randomised PIT histogram, where the dashed line marks the density of the uniform distribution expected under a well-calibrated model. . . . .                                                                                                                         | 179 |
| 5.9  | Estimated DZ-level relative risks and uncertainty under the ICAR-AR(1) model with IncRate covariate. Left column shows posterior median relative risks averaged over 2019 (A), averaged over 2020 (C), and for March 2020 (E). Right column shows the corresponding 95% credible interval widths (B, D, F). Colour scales are shared within each column. . . . .                                                                                                                                                  | 181 |
| 5.10 | City-wide posterior median relative risk (relative to the national level over 2016–2020) over the 24-month study period under the ICAR-AR(1) model with IncRate covariate. Points show the posterior median of the monthly mean relative risk across all $M = 746$ Data Zones, with the shaded band indicating the 95% credible interval. The highlighted window is a one-month band centred on March 2020, when the posterior median relative risk peaks at 1.76. . . . .                                        | 182 |
| 5.11 | Health inequality in respiratory prescribing risk over the study period under the ICAR-AR(1) model with IncRate covariate. Panel (A) shows the posterior median of the city-wide spatial standard deviation of DZ-level relative risks with 95% credible bands. Panel (B) shows the posterior median of the decile-averaged DZ-level relative risk by SIMD 2020 deprivation decile, where decile 1 denotes the most deprived 10% of Scottish Data Zones. The highlighted window is centred on March 2020. . . . . | 183 |



# Acknowledgements

I would like to express my sincere gratitude to my supervisors, Prof Duncan Lee and Dr Oliver Stoner, for their guidance, patience and dedication throughout my PhD. Duncan provided invaluable direction on the broader research vision and nurtured me into an independent researcher. Oliver worked closely with me on the technical details and was always generous with his time when I needed to talk through ideas. Both invested considerable effort in improving my academic skills, and their careful feedback on successive drafts has shaped this thesis in ways I cannot overstate.

I am grateful to my friends and fellow PhD students in the School of Mathematics and Statistics for their companionship and support. Their willingness to listen, offer advice, and simply be there during difficult periods made the experience far less isolating than it might otherwise have been.

I thank my parents for the unconditional support they have given me to pursue my education abroad. Despite having had limited access to higher education themselves, they always encouraged me to follow my own path. I owe everything to their belief in me.

Finally, I thank Mr Jianfeng Li for his companionship throughout the preparation of this thesis. His patience and presence during the most demanding months of this journey gave me the strength to see it through.

# Declaration

I, Qiangqiang Zhu, declare that this thesis titled, “Bayesian spatio-temporal modelling of air pollution and disease risk with spatially misaligned data” and the work presented in it are my own. I confirm that where I have consulted the published work of others, this is always clearly attributed.

The work presented in Chapter 3 has been published in *Environmental and Ecological Statistics* with the title “A comparison of statistical and machine learning models for spatio-temporal prediction of ambient air pollutant concentrations in Scotland” (Zhu et al., 2024), and is jointly authored by Prof Duncan Lee and Dr Oliver Stoner.

The work presented in Chapter 4 has been published in *Statistical Methods in Medical Research* with the title “Quantifying the effects of air pollution on respiratory ill health treated in primary care when the locations of the populations at risk are partially unknown” (Zhu et al., 2026), and is jointly authored by Prof Duncan Lee and Dr Oliver Stoner.

The code supporting the work in Chapter 3 is available at <https://github.com/Qiangqiang-Zhu/air-pollution-prediction>; the code supporting Chapter 4 is available at <https://github.com/Qiangqiang-Zhu/health-effects-primary-care>; and the code supporting Chapter 5 is available at <https://github.com/Qiangqiang-Zhu/disease-mapping>.

# Chapter 1

## Introduction

Air pollution remains one of the most important environmental threats to population health. It is now recognised as the second highest global risk factor for non-communicable disease, after high blood pressure ([World Health Organization, 2025](#)). The combined effects of ambient (outdoor) and household air pollution are associated with 6.7 million premature deaths each year, of which an estimated 4.2 million are attributable to ambient air pollution alone ([World Health Organization, 2024](#)). The World Health Organisation (WHO) estimates that 99% of the global population live in places where its air quality guideline levels are not met ([World Health Organization, 2024](#)). The burden is greatest in low- and middle-income countries, but the health effects remain substantial in high-income countries where air pollution concentrations have declined over time. Air pollution is therefore not only a global environmental problem. It is also a continuing public health problem in countries with established regulatory systems.

In the United Kingdom, poor air quality remains the largest environmental risk to public health. Long-term exposure to ambient air pollution contributes to morbidity, mortality, and major financial costs for the health system and wider society ([Office for Health Improvement and Disparities, 2022](#); [COMEAP, 2023](#)). In Scotland, an estimated 1,800 to 2,700 deaths each year are attributed to long-term exposure to outdoor air pollution ([Public Health Scotland, 2024](#)). The burden is also unequally distributed. More disadvantaged communities are more likely to live in more polluted environments and are more likely to have pre-existing conditions that increase vulnerability to exposure ([Scottish Government, 2021](#); [Chief Medical Officer, 2022](#)). These inequalities strengthen the case for detailed population-level research on the health effects of air pollution.

The severity of these impacts has led to a substantial policy response. The WHO

updated its Global Air Quality Guidelines in 2021 ([World Health Organization, 2021](#)). In Europe, the new European Air Quality Directive (EU) 2024/2881 revised ambient air quality standards that need to be attained by 2030 ([European Parliament and Council, 2024](#)). In the United Kingdom, air quality policy has developed through the Clean Air Acts ([UK Parliament, 1993](#)), the Air Quality Strategy ([Department for Environment, Food and Rural Affairs, 2011](#)), the Clean Air Strategy 2019 ([Department for Environment, Food and Rural Affairs, 2019](#)), and the Environment Act 2021 ([UK Parliament, 2021](#)). In Scotland, concentrations of key outdoor pollutants are regulated by the Air Quality Standards (Scotland) Regulations 2010, as amended in 2016 ([Scottish Government, 2010](#)), while Cleaner Air for Scotland 2 provides the current national policy framework ([Scottish Government, 2021](#)). These measures have contributed to long-term improvements in air quality across the UK. However, they have not removed the need for new evidence on the distribution of exposure and its effects on health.

This thesis focuses on nitrogen dioxide ( $\text{NO}_2$ ), particulate matter with aerodynamic diameter less than or equal to  $10\text{ }\mu\text{m}$  ( $\text{PM}_{10}$ ), and fine particulate matter with aerodynamic diameter less than or equal to  $2.5\text{ }\mu\text{m}$  ( $\text{PM}_{2.5}$ ). These are the pollutants of greatest public health concern in the United Kingdom and in Scotland ([Public Health Scotland, 2024](#); [Department for Environment, Food and Rural Affairs, 2025](#)). The Scottish annual mean targets under the National Air Quality Strategy (NAQS) are  $40\text{ }\mu\text{g}/\text{m}^3$  for  $\text{NO}_2$ ,  $18\text{ }\mu\text{g}/\text{m}^3$  for  $\text{PM}_{10}$ , and  $10\text{ }\mu\text{g}/\text{m}^3$  for  $\text{PM}_{2.5}$ . The more recent EU guidelines (Directive 2008/50/EC) enacted in 2024 set limit values of  $20\text{ }\mu\text{g}/\text{m}^3$  for  $\text{NO}_2$ ,  $20\text{ }\mu\text{g}/\text{m}^3$  for  $\text{PM}_{10}$ , and  $10\text{ }\mu\text{g}/\text{m}^3$  for  $\text{PM}_{2.5}$  to be attained by January 2030. Long-term exposure to these pollutants has been associated with respiratory disease, cardiovascular disease, lung cancer, and premature mortality ([Brunekreef and Holgate, 2002](#); [Rajagopalan et al., 2018](#); [World Health Organization, 2024](#)). There is also growing evidence of links with cognitive decline, dementia, diabetes, and adverse birth outcomes ([Chief Medical Officer, 2022](#); [Public Health Scotland, 2024](#)). For  $\text{NO}_2$  and  $\text{PM}_{2.5}$ , there is no clear evidence of a safe exposure threshold below which adverse health effects do not occur ([World Health Organization, 2021](#); [Public Health Scotland, 2024](#)).

$\text{NO}_2$  is produced mainly by high-temperature combustion, especially from road traffic ([Department for Environment, Food and Rural Affairs, 2025](#)). Exposure can inflame the airways, increase susceptibility to respiratory infection, and worsen existing lung and heart conditions ([Chauhan et al., 2003](#); [Ross et al., 2006](#)). Particulate matter poses a related but broader risk.  $\text{PM}_{10}$  can enter the thoracic region of the

respiratory tract, while  $PM_{2.5}$  can penetrate more deeply into the lungs and enter the bloodstream ([Kampa and Castanas, 2008](#); [World Health Organization, 2024](#)). Long-term exposure to particulate matter has been linked to asthma, chronic obstructive pulmonary disease (COPD), cardiovascular disease, and lung cancer ([International Agency for Research on Cancer, 2013](#); [Rajagopalan et al., 2018](#); [Sin et al., 2023](#); [Chen et al., 2023](#)). These effects are most serious for children, older adults, and people with pre-existing cardiopulmonary conditions ([Anderson, 2009](#); [Public Health Scotland, 2024](#)).

## 1.1 Respiratory ill health in primary care

Most epidemiological research on air pollution and health has focused on severe outcomes such as hospital admissions and mortality ([Rushworth et al., 2014](#); [Lee et al., 2017](#); [Huang et al., 2018](#); [Khomenko et al., 2021](#)). These outcomes are important, but they do not capture the full burden of disease attributable to poor air quality. A much larger number of people live with respiratory conditions such as asthma and COPD that are managed in primary care rather than in hospital. In the United Kingdom, around 5.4 million people receive treatment for asthma and around 3 million live with COPD, with the latter accounting for approximately 1.4 million general practitioner (GP) consultations each year ([NICE, 2025a,b](#)). Thus air pollution may affect less severe respiratory ill health, even if those effects are less substantial than those on hospital admission or mortality.

Only a small number of studies have considered this part of the disease burden. [Blangiardo et al. \(2016\)](#) linked  $NO_2$  concentrations to respiratory prescriptions in GP surgeries in England and found a small positive association, with an estimated 0.07% increase in prescribing rates for a  $10 \mu g/m^3$  increase in  $NO_2$ . [Lee \(2018\)](#) modelled reliever and preventer prescriptions separately in Scotland and found that a  $2 \mu g/m^3$  increase in  $PM_{10}$  and  $PM_{2.5}$  was associated with approximately 1% to 3% increases in prescribing rates. These studies show that primary care data can provide useful evidence on the effects of air pollution on non-hospitalised respiratory ill health. However, the evidence base remains limited and methodologically challenging, for two main reasons. First, air pollution exposure must be estimated rather than directly observed at the population level. Assigning concentrations at the GP surgery location or within arbitrary buffer zones therefore introduces measurement error that would attenuate the estimated health effects towards the null ([Lee, 2018](#)). Second, the spatial dependence among GP surgery populations is poorly captured by conventional

distance-based neighbourhood matrices. Two surgeries that are geographically close may serve non- or less-overlapping populations, while two that are far apart may share patients from the same residential areas. How to capture such spatial closeness of the patient populations registered at two surgeries is one of the main motivations of this thesis.

## 1.2 Small-area disease mapping

Understanding where the burden of respiratory ill health is highest and how it changes over time requires not only evidence on population-health associations, but also reliable small-area disease maps. Disease mapping studies the spatial or spatio-temporal distribution of population-level disease risk ([Lawson, 2018](#)), whose aim is not only to describe where disease risk is high or low. It also supports public health surveillance, public health service allocation, and the study of health inequalities. Disease maps can identify areas with unusually high risk, show how risk changes over time, and reveal whether health burdens are concentrated in disadvantaged communities ([Lawson and Lee, 2017](#); [MacNab, 2022](#)). For these reasons, small-area disease mapping has become an important part of modern public health research. This is particularly relevant in Scotland, where health inequalities are strong and persistent. Areas with high deprivation often experience worse health outcomes than richer areas, although the strength and form of this relationship may vary by disease and region ([Walsh et al., 2010](#); [Scottish Government, 2021](#)). For example, [Whyte et al. \(2021\)](#) showed that there was a 17.6 year gap of male life expectancy in Glasgow ranging from 65.4 years in the most deprived neighbourhoods to around 83 years within a single city.

Bayesian hierarchical models are now commonly used in disease mapping because they provide stable estimates of disease risk and coherent uncertainty quantification ([Besag et al., 1991](#); [Leroux et al., 2000](#)). In these models, health risk is estimated by combining observed disease counts with spatially structured random effects, usually based on conditional autoregressive priors. Extensions to the spatio-temporal setting are well established ([Knorr-Held, 2000](#); [Rushworth et al., 2014](#); [Lee, 2025](#)). A detailed review of these models is given in Chapter 2. The key point is that standard disease mapping methods assume that disease data are observed on a set of non-overlapping areal units with known boundaries, such as census areas or health board regions. This assumption becomes problematic when the available data are recorded on service units such as GP surgeries, because the exact geographical boundaries

of registered patients are unknown and are likely to overlap between neighbouring surgeries.

The long-term effects of air pollution can be studied using either cohort designs or population-level ecological designs. Cohort studies follow individuals over time and relate their health to estimated pollution exposure. Although they are the gold-standard design, they are financially expensive and time consuming to conduct due to the need to follow up a large cohort of people over multiple decades. By contrast, spatio-temporal ecological studies can use routinely collected administrative data and are therefore easier and much faster to implement over large populations and long time periods. This design has been used widely in environmental epidemiology, including in Scotland (Lee et al., 2009, 2016; Huang et al., 2018). The work in this thesis follows that ecological design.

### 1.3 The unifying challenge

A single methodological problem runs through the three studies in this thesis. In each case the support of the data and the support of the inferential target do not coincide, a situation known as spatial support mismatch, or more specifically a change-of-support problem (Bradley et al., 2016; Shaddick et al., 2018). Air pollution is measured at a sparse set of monitoring sites, yet exposure is required on a regular grid covering the whole nation. Respiratory prescriptions are recorded at GP surgeries, yet the populations at risk live across surrounding residential areas whose extent is not directly observed. Small-area disease risk is most useful on fixed administrative units such as Data Zones, while the available outcome data are reported by GP surgeries whose catchments overlap and are partially unknown. The central methodological theme of this thesis is thus how to perform principled spatio-temporal inference when the supports of the data and the supports of the inferential target do not coincide. Therefore, the overarching aim of this thesis is to develop spatio-temporal methods to address these challenges, with application to air pollution exposure estimation and respiratory disease burden quantification in Scotland.

The first instance of this mismatch concerns air pollution exposure, which must be estimated at the spatial and temporal resolutions required for an ecological study from data observed only at monitoring sites. In Scotland, nationwide information on air pollution comes mainly from two sources. The first is the Automatic Urban and Rural Network (AURN), a sparse network of monitoring sites that provides accurate measurements but limited spatial coverage. The second is the Pollution Climate

Mapping (PCM) model, which provides complete spatial coverage on a 1 km<sup>2</sup> grid but only at annual resolution and with lower local accuracy as it is a model and not measurements. A range of quantitative methods have been used to combine or extend these data, including land use regression models (Larkin et al., 2017), geostatistical models (Li et al., 2012), hierarchical spatio-temporal models (Mukhopadhyay and Sahu, 2017) and machine learning methods (Meng et al., 2018). However, no studies have compared these approaches systematically for multiple pollutants within Scotland using a common validation design. This is the first contribution of this thesis. In Chapter 3, I compare a set of statistical and machine learning models for predicting monthly NO<sub>2</sub>, PM<sub>10</sub>, and PM<sub>2.5</sub> concentrations across Scotland between 2016 and 2020, and I use the best performing models to produce a new 1 km<sup>2</sup> monthly pollution data resource with uncertainty quantification (Zhu et al., 2024). This provides the exposure inputs required for the later epidemiological analysis.

A second challenge arises when the health outcomes are reported by GP surgeries, while the registered patient populations for each surgery actually live in the surrounding catchment areas that cannot be accurately observed. Specifically, respiratory medication prescription counts are recorded for each surgery, but the patients who receive these prescriptions do not live at the surgery location. Instead, they are distributed across many residential areas, and the exact GP catchment boundaries are not directly observed. Previous related studies have assigned pollution exposure either at the surgery location or within ad hoc buffer zones around the surgery (Blangiardo et al., 2016; Lee, 2018). Both approaches are unrealistic because they do not reflect where patients actually live, and are therefore likely to introduce exposure measurement error and attenuate the estimated health effects. Similar problems arise for the spatial dependence structure. Distance-based neighbourhood matrices between surgery buildings do not necessarily reflect the spatial relationships between the populations they serve. In Chapter 4, I develop a Bayesian spatio-temporal smoothing model for GP prescribing data that realigns pollution and other covariates to likely patient locations rather than to surgery locations. I also propose new population informed neighbourhood matrices to better capture the spatial dependence between GP patient populations.

The third form of the problem concerns disease mapping, where the outcome is observed at GP surgeries but inference is required on fixed, non-overlapping administrative units. If the aim is to estimate small-area respiratory ill health risks across Scotland, GP surgeries are not ideal reporting units. They are not standard administrative areas, their effective catchments overlap, their number changes over



time, and their boundaries are only partially known. Public health surveillance is more naturally conducted on stable small-area geographies such as Data Zones, which have fixed boundaries and can be linked directly to deprivation and other contextual measures. This creates a change-of-support problem. The outcome is observed at GP surgeries, but inference is desired at the Data Zone level. Existing approaches to spatial misalignment include areal interpolation (Flowerdew and Green, 1992), Bayesian change-of-support models (Bradley et al., 2016; Shaddick et al., 2018), data augmentation methods (Lee and Anderson, 2023), and multiple-membership extensions of disease mapping models (Gramatica et al., 2021, 2024). However, a spatio-temporal Bayesian framework for realigning univariate GP-level count data to non-overlapping areal units using patient registration weights has not been fully developed. This is the third contribution of this thesis. In Chapter 5, I develop a Bayesian disease mapping framework that realigns GP-level prescription counts to Data Zones (official public health and statistical reporting units) and quantifies the resulting uncertainty, while also examining how realignment interacts with spatial smoothing and estimation bias.

The thesis is conducted through three linked research projects. First, I predict monthly concentrations of  $\text{NO}_2$ ,  $\text{PM}_{10}$ , and  $\text{PM}_{2.5}$  on a  $1 \text{ km}^2$  grid across Scotland using routinely available monitoring, modelled, and contextual data. Then I use these predicted exposure surfaces to study the effects of air pollution on respiratory ill health treated in primary care, while accounting for the fact that GP patient populations are only partially observed in space. Finally, I extend this framework to realign GP-level prescription counts to Data Zones to produce interpretable small-area disease maps on a standard administrative geography.

## 1.4 Study region and data

The study region for these works is Scotland, United Kingdom, and the analyses cover the period 2016 to 2020. Scotland provides an ideal and interesting setting for this work because it combines marked spatial contrasts in urbanicity, socio-economic deprivation, and respiratory health with a rich administrative and environmental data infrastructure. Glasgow in particular has long been associated with poor health outcomes relative to comparable cities, a pattern often referred to as the Glasgow effect (Cowley et al., 2016). More generally, Scotland exhibits clear small-area inequalities in health and deprivation, which makes it an informative setting for environmental epidemiology and disease mapping.

Several spatial scales are used in the thesis. Air pollution is measured at approximately 100 monitoring sites in the AURN, most of which are located in the major urban centres of Aberdeen, Dundee, Edinburgh, and Glasgow. Pollution predictions are produced on a regular  $1 \text{ km}^2$  grid covering Scotland. Health outcomes are recorded for around 1,000 GP surgeries that were operational during at least part of the study period. The target geography for small-area inference is the set of 6,976 Data Zones, which are the standard small-area units used in Scottish public health reporting. Data Zones typically contain around 500 to 1,000 residents. An intermediate geography of 14 Health Boards is also used for some descriptive summaries.

The principal data sources used in this thesis are the Scottish Air Quality database and the Pollution Climate Mapping model for measured and modelled pollutant concentrations, the Prescribing Information System for monthly GP-level prescription counts, the Scottish Primary Care Information Resource for patient registration and demographic data, the Scottish Index of Multiple Deprivation for area-level social and economic context, the HadUK-Grid dataset for meteorological covariates, and National Records of Scotland mid-year population estimates. Further details of these data are given in the relevant chapters.

The remainder of this thesis is organised as follows. Chapter 2 reviews the statistical methods that underpin the later analyses, including regression models, additive models, machine learning methods, Bayesian hierarchical modelling, and spatial and spatio-temporal models for areal data. Chapter 3 considers the spatio-temporal prediction of monthly  $\text{NO}_2$ ,  $\text{PM}_{10}$ , and  $\text{PM}_{2.5}$  concentrations across Scotland. Chapter 4 investigates the association between air pollution and respiratory ill health treated in primary care, using GP prescribing data and methods that account for the partially observed spatial distribution of GP patient populations. Chapter 5 develops a Bayesian disease mapping framework for realigning GP-level prescription counts to Data Zones. Finally, Chapter 6 summarises the main findings, discusses limitations, and outlines directions for future research.

# Chapter 2

## Statistical background

This chapter reviews the statistical methods that underpin the analyses presented in later chapters, with a particular focus on regression modelling, Bayesian hierarchical frameworks, and spatial and spatio-temporal smoothing models. The chapter is structured as follows: Section 2.1 introduces linear regression and generalised linear models, and describes the exponential family framework used for likelihood specification. Section 2.2 relaxes the linearity assumption in Section 2.1 by introducing regression splines, and shows how these are combined in generalised additive models. Section 2.3 introduces tree-based ensemble models, as non-parametric tools for the spatio-temporal prediction context. Section 2.4 presents the Bayesian framework, including Bayes' theorem, prior specification, and posterior inference via Markov chain Monte Carlo simulation and approximation methods. Section 2.5 reviews model selection criteria. Finally, Section 2.6 introduces key concepts and models for areal data, including disease mapping and a range of spatial and spatio-temporal random effects models.

### 2.1 Classic regression and GLMs

Regression modelling is an important tool in this thesis, both for predicting air pollution concentrations and for quantifying the health effects of air pollutants. This section briefly reviews classical linear regression and its extension to generalised linear models (GLMs), which provide the basic likelihood and mean-response structures used throughout the later chapters.

### 2.1.1 Linear regression

Linear regression is one of the most widely used statistical models. It assumes a linear relationship between a response variable and a set of predictors, and provides interpretable coefficients as well as a basis for more complex models. In Chapter 3, linear regression is used as a baseline model for air pollution prediction, with modelled climate outputs, meteorological factors, land cover, land use and road traffic conditions as covariates.

Assuming  $n$  observations of a response  $y_i$  and  $p$  covariates  $\{x_{i1}, \dots, x_{ip}\}$  for  $i = 1, \dots, n$ , a linear regression model can be expressed in matrix form as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathbf{N}(\mathbf{0}, \sigma^2 \mathbf{I}), \quad (2.1)$$

where  $\mathbf{y} = (y_1, \dots, y_n)^\top$  is an  $n \times 1$  vector of responses,  $\mathbf{X}$  is an  $n \times (p + 1)$  design matrix of predictors,  $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)^\top$  is a  $(p + 1) \times 1$  vector of the regression parameters, and  $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)^\top$  is an  $n \times 1$  vector of errors following an independent multivariate normal (white-noise) distribution. The  $i$ -th row of  $\mathbf{X}$  is denoted by  $\mathbf{x}_i^\top = (1, x_{i1}, \dots, x_{ip})$ , representing the predictors for the  $i$ -th observation, and the leading 1 corresponds to the intercept.

The ordinary least squares (OLS) estimator of  $\boldsymbol{\beta}$  minimises the residual sum of squares  $\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2$  and has closed form

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y},$$

provided that  $\mathbf{X}^\top \mathbf{X}$  is non-singular. The fitted values are

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y},$$

and an unbiased estimator of the error variance  $\sigma^2$  is

$$\hat{\sigma}^2 = \frac{\|\mathbf{y} - \hat{\mathbf{y}}\|^2}{n - p - 1}.$$

A commonly used measure of how well the linear model fits the data is the *coefficient of determination*,

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2},$$

where  $\bar{y} = n^{-1} \sum_{i=1}^n y_i$ . In practice,  $R^2$  represents the proportion of the total variability in the response that is explained by the fitted model, with values closer to 1 indicating a better fit.

For any covariate vector  $\mathbf{x}_0$ , the fitted mean  $\hat{y}_0 = \mathbf{x}_0^\top \hat{\boldsymbol{\beta}}$  has variance

$$\text{Var}(\hat{y}_0) = \hat{\sigma}^2 \mathbf{x}_0^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_0,$$

yielding a  $100(1 - \alpha)\%$  confidence interval

$$\hat{y}_0 \pm \hat{\sigma} \sqrt{\mathbf{x}_0^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_0} \times t_{n-p-1}(1 - \alpha/2),$$

where  $t_{n-p-1}(1 - \alpha/2)$  denotes the  $(1 - \alpha/2)$  quantile of the  $t$ -distribution with  $n - p - 1$  degrees of freedom.

A  $100(1 - \alpha)\%$  prediction interval for a new observation at covariate vector  $\mathbf{x}_{\text{new}}$  additionally incorporates the error variance, and is given by

$$\hat{y}_{\text{new}} \pm \hat{\sigma} \sqrt{1 + \mathbf{x}_{\text{new}}^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_{\text{new}}} \times t_{n-p-1}(1 - \alpha/2).$$

Here the factor “1” inside the square root reflects the additional uncertainty due to the random error in the new observation, beyond the uncertainty in the mean response.

### 2.1.2 Generalised linear models

In linear regression, the response is assumed to be continuous, with a conditional normal distribution given the predictors and an unbounded range of possible values. In many applications this is unrealistic, as responses may be counts, proportions, or binary indicators, and their variability can depend on the mean. These limitations motivate *generalised linear models* (GLMs; [Nelder and Wedderburn, 1972](#)), which extend linear regression to non-Gaussian responses while retaining a linear predictor on a transformed scale.

A GLM assumes that each response  $Y_i$  follows a distribution from the exponential family (introduced below), and links its conditional mean  $\mu_i = \mathbb{E}(Y_i | \mathbf{x}_i)$  to a set of covariates through a link function  $g(\cdot)$ :

$$Y_i \sim f(y_i | \theta_i), \quad g(\mu_i) = \eta_i = \mathbf{x}_i^\top \boldsymbol{\beta}, \quad i = 1, \dots, n,$$

where  $f(\cdot)$  is an exponential family density or mass function,  $g(\cdot)$  is the link function,

and  $\theta_i$  is a canonical parameter that determines the distribution of  $Y_i$ . The linear predictor  $\eta_i$  plays the same role as in classical linear regression, but acts on the transformed mean  $g(\mu_i)$  rather than directly on  $\mu_i$ .

### The exponential family of distributions

The GLM framework is built upon the *exponential family* of distributions. A probability density function (PDF) or probability mass function (PMF) in this family can be written as

$$f(y|\theta) = \exp \left\{ \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi) \right\}, \quad (2.2)$$

where  $a(\cdot)$ ,  $b(\cdot)$  and  $c(\cdot)$  are known functions,  $\theta$  is the *canonical parameter*, and  $\phi$  is a *scale* (or dispersion) parameter. Many standard distributions, including the Gaussian, Poisson, Binomial and Gamma distributions, can be written in this form.

The mean and variance of an exponential family distribution are given by

$$\mathbb{E}(Y) = \mu = b'(\theta) \quad \text{and} \quad \text{Var}(Y) = b''(\theta) a(\phi),$$

where  $b'(\theta)$  and  $b''(\theta)$  denote the first and second derivatives of  $b(\theta)$  with respect to  $\theta$ . In a GLM, the link between the regression parameters  $\beta$  and the distribution of  $Y_i$  arises through the relationship

$$g(\mu_i) = g\{b'(\theta_i)\} = \mathbf{x}_i^\top \beta,$$

so the choice of link function and exponential family distribution together determine the mean-variance relationship implied by the model. This framework is particularly important later in the thesis when specifying Poisson and negative binomial regression models for disease prescription counts, and when embedding these likelihoods within hierarchical spatial and spatio-temporal models. Maximum likelihood estimation for GLMs is typically carried out using the iterative re-weighted least squares (IRLS) algorithm, and further details can be found in [Dobson and Barnett \(2018\)](#).

### 2.1.3 Multicollinearity and the variance inflation factor

The regression models above assume that the columns of the design matrix  $\mathbf{X}$  are not linearly dependent, so that  $\mathbf{X}^\top \mathbf{X}$  is non-singular and  $\beta$  is identifiable. In practice the covariates are often strongly correlated, a situation known as *multicollinearity* ([Dobson and Barnett, 2018](#)). Near-linear dependence among the covariates does not

bias the ordinary least squares estimator, but it inflates the variance of the individual coefficients. The *variance inflation factor* (VIF) quantifies this inflation for each covariate (Fox and Monette, 1992).

Let  $R_i^2$  be the coefficient of determination from the auxiliary linear regression of  $\mathbf{x}_i$  on the remaining  $p - 1$  covariates (together with an intercept). The variance inflation factor for covariate  $i$  is defined as

$$\text{VIF}_i = \frac{1}{1 - R_i^2}.$$

Its role is made explicit by the variance of the corresponding OLS coefficient, which can be written as

$$\text{Var}(\hat{\beta}_i) = \frac{\sigma^2}{(n - 1) s_i^2} \text{VIF}_i, \quad (2.3)$$

where  $s_i^2$  is the sample variance of  $\mathbf{x}_i$ . The first factor in Equation (2.3) is the variance that would be obtained if  $\mathbf{x}_i$  were orthogonal to the other covariates, so  $\text{VIF}_i$  is precisely the factor by which the variance of  $\hat{\beta}_i$  is inflated by its linear association with the remaining covariates. When the covariates are mutually orthogonal,  $R_i^2 = 0$  and  $\text{VIF}_i = 1$ , indicating no inflation. As  $\mathbf{x}_i$  becomes more predictable from the other covariates,  $R_i^2 \rightarrow 1$  and  $\text{VIF}_i \rightarrow \infty$ . Commonly used rules of thumb flag a covariate as problematically collinear when its VIF exceeds 5 or 10 (James et al., 2013).

## 2.2 Smooth functions and additive models

Generalised linear models assume that the systematic component of the model, the linear predictor  $\eta_i$ , is a linear function of the covariates  $\mathbf{x}_i$ . In many applications, especially in environmental epidemiology, this assumption is restrictive, as relationships between outcomes and predictors such as temperature, air pollutant concentrations, or socio-economic indices are often non-linear. A flexible way to relax the linearity assumption while retaining much of the GLM framework is to replace  $\mathbf{X}\boldsymbol{\beta}$  by smooth functions of the covariates.

This section introduces regression splines as building bases for modelling non-linear effects, and shows how they can be combined into generalised additive models (GAMs). It then briefly discusses an extension, generalised additive models for location, scale and shape (GAMLSS), which allow smooth covariate effects on multiple parameters of the response distribution.

### 2.2.1 Regression splines

Regression splines represent a smooth function as a linear combination of basis functions that are piecewise polynomials joined at pre-specified locations called *knots*. Let  $x$  denote a univariate covariate and suppose we wish to model a smooth function  $f(x)$ . A general regression spline representation takes the form

$$f(x) = \sum_{k=1}^K \beta_k h_k(x),$$

where  $\{h_k(x)\}_{k=1}^K$  are chosen basis functions and  $\{\beta_k\}_{k=1}^K$  are coefficients to be estimated. The choice of basis and their number  $K$  controls the smoothness and numerical stability of the resulting fit. Below, several commonly used spline bases are reviewed.

#### Truncated power basis splines

A flexible way to model non-linear link functions is through *truncated power basis* splines. Let  $c_1, \dots, c_K$  denote  $K$  interior knots, which partition the predictor domain into non-overlapping intervals. An  $M$ th-order regression spline (i.e. polynomial pieces of degree  $M - 1$ ) can be written as

$$f(x) = \sum_{k=0}^{K+M-1} \beta_k h_k(x),$$

with basis functions

$$h_k(x) = \begin{cases} x^k & \text{if } k \in \{0, \dots, M-1\}, \\ (x - c_{k-M+1})_+^{M-1} & \text{if } k \in \{M, \dots, K+M-1\}, \end{cases}$$

where  $(x - c)_+^{M-1} = (x - c)^{M-1}$  if  $x > c$  and 0 otherwise. The set  $\{h_k(x)\}$  forms the truncated power basis of degree  $M - 1$ , and the coefficients  $\{\beta_k\}$  govern the contribution of each component. Two common special cases are:

**Linear splines** ( $M = 2$ ): continuous, but not differentiable at the knots;

**Cubic splines** ( $M = 4$ ): continuous with continuous first and second derivatives at knots, resulting in a much smoother curve.

Higher values of  $M$  yield smoother fits but may increase the risk of over-fitting if the number of knots is large and no additional regularisation is applied.



### B-splines

Truncated power bases are conceptually simple but can lead to numerically unstable design matrices, because the columns of the design matrix are often highly correlated. In practice, especially for least-squares or likelihood-based fitting, it is advantageous to use alternative bases with better numerical properties. The most common choice is the *B-spline* basis (De Boor, 1978; Dierckx, 1995), whose basis functions have local support and yield a well-conditioned design matrix even with many knots.

To construct an  $M$ th-order *B-spline* basis, one first chooses  $K$  interior knots  $c_1, \dots, c_K$  and boundary knots  $c_0 < c_1$  and  $c_{K+1} > c_K$ . An augmented knot sequence  $\kappa_1, \dots, \kappa_{K+2M}$  is then defined by

- $\kappa_1 \leq \kappa_2 \leq \dots \leq \kappa_M \leq c_0$ ,
- $\kappa_{j+M} = c_j$  for  $j = 1, \dots, K$ ,
- $c_{K+1} \leq \kappa_{K+M+1} \leq \dots \leq \kappa_{K+2M}$ .

Let  $B_{i,r}(x)$  denote the  $i$ -th *B-spline* basis function of order  $r$ . The Cox-de Boor recursion (De Boor, 1972; Cox, 1972) defines the order-1 basis functions as

$$B_{i,1}(x) = \begin{cases} 1 & \text{if } \kappa_i \leq x < \kappa_{i+1}, \\ 1 & \text{if } x = \kappa_{K+2M}, \\ 0 & \text{otherwise,} \end{cases} \quad i = 1, \dots, K + 2M - 1,$$

and, for  $1 < r \leq M$ , the higher-order basis functions are obtained by

$$B_{i,r}(x) = \frac{x - \kappa_i}{\kappa_{i+r-1} - \kappa_i} B_{i,r-1}(x) \mathbb{I}(\kappa_{i+r-1} > \kappa_i) + \frac{\kappa_{i+r} - x}{\kappa_{i+r} - \kappa_{i+1}} B_{i+1,r-1}(x) \mathbb{I}(\kappa_{i+r} > \kappa_{i+1}),$$

where  $\mathbb{I}(\cdot)$  is the indicator function.

An  $M$ -order regression spline represented by *B-spline* bases is thus

$$f(x) = \sum_{j=1}^{K+M} \beta_j B_{j,M}(x),$$

with the design matrix

$$\mathbf{B} = \begin{pmatrix} B_{1,M}(x_1) & \cdots & B_{K+M,M}(x_1) \\ \vdots & \ddots & \vdots \\ B_{1,M}(x_n) & \cdots & B_{K+M,M}(x_n) \end{pmatrix}.$$

Because each  $B$ -spline basis function is non-zero only over a small number of adjacent knot intervals, each row of  $\mathbf{B}$  contains relatively few non-zero entries. This local support reduces multicollinearity and improves numerical stability compared to truncated power basis splines.

### Smoothing splines

*Smoothing splines* avoid the explicit choice of a small set of knots. Instead, they typically use many (often all) distinct predictor values as knots, and control over-fitting via regularisation rather than by limiting the number of basis functions (Reinsch, 1967; Wang, 2011). Regularisation restricts the magnitude of the estimated coefficients which penalises the roughness of the fitted function to prevent it from being too complex.

Given observations  $(x_i, y_i)$ ,  $i = 1, \dots, n$ , the smoothing spline estimate of  $f(\cdot)$  minimises the *penalised residual sum of squares* (PRSS)

$$\text{PRSS}(f, \lambda) = \sum_{i=1}^n \{y_i - f(x_i)\}^2 + \lambda \int \{f''(t)\}^2 dt, \quad (2.4)$$

where the first term measures lack of fit and the second term penalises curvature through the integrated squared second derivative. The *smoothing parameter*  $\lambda > 0$  controls the trade-off between goodness-of-fit and smoothness, where small  $\lambda$  allows a wiggly curve that follows the data closely, while large  $\lambda$  enforces a nearly linear function.

In matrix notation, using a spline basis with design matrix  $\mathbf{B}$ , the PRSS can be written as

$$\text{PRSS}(\boldsymbol{\beta}, \lambda) = \|\mathbf{y} - \mathbf{B}\boldsymbol{\beta}\|^2 + \lambda \boldsymbol{\beta}^\top \mathbf{S} \boldsymbol{\beta},$$

where  $\mathbf{B}_{ij} = B_j(x_i)$ , and  $\mathbf{S}$  is a *penalty matrix* with entries  $\mathbf{S}_{ij} = \int B_i''(t) B_j''(t) dt$ . The minimiser is

$$\hat{\boldsymbol{\beta}} = (\mathbf{B}^\top \mathbf{B} + \lambda \mathbf{S})^{-1} \mathbf{B}^\top \mathbf{y} \quad \text{and} \quad \hat{\mathbf{y}} = \mathbf{B} \hat{\boldsymbol{\beta}} = \mathbf{S}_\lambda \mathbf{y},$$

where  $\mathbf{S}_\lambda = \mathbf{B} (\mathbf{B}^\top \mathbf{B} + \lambda \mathbf{S})^{-1} \mathbf{B}^\top$  is called the *smoother matrix*, and  $\text{df}_\lambda = \text{trace}(\mathbf{S}_\lambda)$  is the *effective degrees of freedom* providing an interpretable measure of model complexity. As  $\lambda \rightarrow 0$ ,  $\text{df}_\lambda \rightarrow n$  and the fit interpolates the data; as  $\lambda \rightarrow \infty$ ,  $\text{df}_\lambda \rightarrow 2$  and the fit approaches a straight line. In practice,  $\lambda$  is usually selected in a data-driven way using criteria such as generalised cross-validation (GCV), or restricted maximum likelihood (REML) in a mixed-model formulation, both of which are widely used in

practice and implemented in software such as `mgcv` in R. In this thesis, the former is used and the generalised cross-validation score for a given  $\lambda$  is

$$\text{GCV}(\lambda) = \frac{\|\mathbf{y} - \hat{\mathbf{y}}_\lambda\|^2/n}{(1 - \text{df}_\lambda/n)^2},$$

and  $\lambda$  is chosen to minimise  $\text{GCV}(\lambda)$ . This criterion balances lack of fit (the residual sum of squares) and model complexity (through  $\text{df}_\lambda$ ), and is widely used in spline-based regression and GAMs introduced in Section 2.2.2.

### ***P*-splines**

Penalised *B*-splines (*P*-splines) combine the numerical advantages of *B*-splines with a simple difference penalty on the coefficients (Eilers and Marx, 1996). The idea is to use a relatively rich *B*-spline basis (with many knots) but control smoothness through a penalty on neighbouring coefficients rather than through an integrated curvature penalty. Considering  $m$  basis functions, the *P*-spline estimate minimises

$$\text{PRSS}(\boldsymbol{\beta}, \lambda) = \sum_{i=1}^n \left\{ y_i - \sum_{j=1}^m \beta_j B_j(x_i) \right\}^2 + \lambda \sum_{j=k+1}^m (\Delta^k \beta_j)^2,$$

where  $\Delta^k$  denotes a  $k$ -order finite difference function (e.g.  $k = 2$  penalises deviations from local linearity in the coefficients). This discrete penalty replaces the numerical integration in Equation (2.4), which is a convenient approximation to curvature penalisation, and is more computationally efficient as it reduces the dimensionality of the problem from  $n$  to  $m$ .

### **2.2.2 Generalised additive models**

The *generalised additive model* (GAM) extends the GLM by allowing the linear predictor to include both linear and smooth functions of the covariates (Hastie and Tibshirani, 1986, 1990). Non-linear relationships are captured through smooth functions represented via spline bases such as those introduced above. For responses  $\{Y_i\}$  following an exponential family distribution with  $\mu_i = \mathbb{E}(Y_i|\mathbf{x}_i)$ , a GAM has the form

$$Y_i \sim f(\mu_i), \quad g(\mu_i) = \mathbf{x}_i^\top \boldsymbol{\beta} + \sum_{j=1}^J f_j(z_{ij}), \quad (2.5)$$

where  $\{f_j(\cdot)\}$  are semi-parametric smooths represented using the spline bases introduced in Section 2.2.1 (e.g. smoothing splines or  $P$ -splines). Each smooth in Equation (2.5) can be written in terms of spline basis function, for example

$$f_j(z) = \sum_{k=1}^{K_j} b_{jk}(z) \delta_{jk},$$

where  $\{b_{jk}(z)\}_k$  are basis functions with a sequence of  $K_j$  knots, and  $\{\delta_{jk}\}_k$  are the coefficients of a penalised spline. Because each smooth term includes its own intercept-like component, identifiability requires constraints to prevent confounding between smooths and the overall intercept. A common choice is to impose centring constraints  $\sum_{i=1}^n f_j(z_{ij}) = 0$  for each  $j$ . Under this constraint, the global intercept retains its interpretation as an overall mean effect, and  $f_j(\cdot)$  represents deviations from that mean associated with covariate  $\mathbf{z}_j$ .

Estimation in GAMs is typically carried out via penalised iteratively re-weighted least squares (PIRLS), which generalises the GLM fitting algorithm to include quadratic penalties on the spline coefficients (Wood, 2017). Each smooth  $\{f_j(\cdot)\}$  is associated with a smoothing parameter  $\lambda_j$  controlling the trade-off between model fit and smoothness, analogous to  $\lambda$  in Equation (2.4), which can be estimated using GCV or REML. In Chapter 3, GAMs are used to flexibly model non-linear exposure–response relationships and temporal trends in air pollution predictions.

### 2.2.3 Generalised additive models for location, scale and shape

Standard GAMs model only the mean of the response distribution via smooth covariate effects. In many applications, other aspects of the distribution, such as variability, skewness, or tail behaviour, may also depend on covariates. *Generalised additive models for location, scale and shape* (GAMLSS; Rigby and Stasinopoulos, 2005) extend GAMs by allowing multiple parameters of a (possibly non-exponential-family) response distribution to have their own linear and smooth predictors.

Suppose the response  $Y_i$  follows a parametric distribution with up to four parameters: location (mean)  $\mu_i$ , scale  $\sigma_i$ , skewness  $\nu_i$  and kurtosis  $\tau_i$ . A GAMLSS model has

the form

$$\begin{aligned}
 Y_i &\sim f(\mu_i, \sigma_i, \nu_i, \tau_i), \\
 \eta_{i,1} &= g_1(\mu_i) = \mathbf{x}_{i,1}^\top \boldsymbol{\beta}_1 + \sum_{k=1}^{K_1} f_k^{(1)}(z_{ik,1}), \\
 \eta_{i,2} &= g_2(\sigma_i) = \mathbf{x}_{i,2}^\top \boldsymbol{\beta}_2 + \sum_{k=1}^{K_2} f_k^{(2)}(z_{ik,2}), \\
 \eta_{i,3} &= g_3(\nu_i) = \mathbf{x}_{i,3}^\top \boldsymbol{\beta}_3 + \sum_{k=1}^{K_3} f_k^{(3)}(z_{ik,3}), \\
 \eta_{i,4} &= g_4(\tau_i) = \mathbf{x}_{i,4}^\top \boldsymbol{\beta}_4 + \sum_{k=1}^{K_4} f_k^{(4)}(z_{ik,4}).
 \end{aligned}$$

For the  $j$ -th distributional parameter in  $\{\mu, \sigma, \nu, \tau\}$ ,  $\eta_{i,j}$  is the linear predictor of the  $i$ -th observation linked to the corresponding parameter by a (possibly different) link function  $g_j(\cdot)$ . The vector  $\mathbf{x}_{i,j}$  contains the covariates that enter linearly in the  $i$ -th predictor vector for the  $j$ -th parameter, with associated regression coefficients  $\boldsymbol{\beta}_j$ . The functions  $f_k^{(j)}(\cdot)$  represent smooth functions acting on the  $k$ -th non-linear covariate for the  $j$ -th parameter  $\mathbf{z}_{k,j}$ , and there are  $K_j$  non-linear covariates for that parameter. In this way, each of the location, scale, skewness and kurtosis parameters can have its own combination of linear and smooth covariate effects, allowing different aspects of the response distribution to vary flexibly over the predictors. Estimation is typically performed via penalised maximum likelihood, with smoothness controlled by penalties analogous to those in GAMs, and further details are given in [Stasinopoulos et al. \(2017\)](#).

## 2.3 Tree-based machine learning methods

In this thesis tree-based machine learning methods are used as flexible predictive models for air pollution prediction in Chapter 3. In particular, I consider regression trees as base learners, together with their ensemble extensions via bootstrap aggregating (bagging) and random forests, and I utilise quantile regression forests to construct prediction intervals. This section introduces these methods in the regression setting relevant for air pollution prediction.

### 2.3.1 Regression trees

Regression trees (Breiman et al., 1984) partition the predictor space into regions within which the response is modelled by a simple constant. Let  $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$  denote training data, where  $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})$  is a vector of covariates and  $y_i$  is a continuous response (e.g. pollutant concentration). A regression tree predictor can be written as

$$\hat{f}(\mathbf{x}_i) = \sum_{m=1}^M c_m \mathbb{I}\{\mathbf{x}_i \in \mathbb{R}_m\},$$

where  $\{\mathbb{R}_m\}_{m=1}^M$  are disjoint regions (terminal nodes or leaves) that partition the covariate space, and  $c_m$  is typically the average of  $y_i$  over training points in region  $\mathbb{R}_m$ . The regions are constructed recursively by binary splits of the form

$$x_j \leq s \quad \text{vs.} \quad x_j > s,$$

where  $j \in \{1, \dots, p\}$  indexes a covariate and  $s$  is a candidate split point. Starting from the full dataset, the algorithm chooses the split  $(j, s)$  that maximises the reduction in residual sum of squares (RSS),

$$\Delta\text{RSS}(j, s) = \sum_{i=1}^n (y_i - \bar{y})^2 - \left\{ \sum_{\mathbf{x}_i \in \mathbb{R}_1(j, s)} (y_i - \bar{y}_{\mathbb{R}_1(j, s)})^2 + \sum_{\mathbf{x}_i \in \mathbb{R}_2(j, s)} (y_i - \bar{y}_{\mathbb{R}_2(j, s)})^2 \right\},$$

where  $\bar{y}$  is the overall mean and  $\bar{y}_{\mathbb{R}_m(j, s)}$  is the mean of  $y_i$  in child node  $\mathbb{R}_m(j, s)$  ( $m = 1, 2$ ) due to the binary split  $(j, s)$ . This greedy splitting continues within each resulting node until a stopping rule is met (e.g. minimum node size or maximum depth).

To predict at a new covariate vector  $\mathbf{x}_0$ , we send  $\mathbf{x}_0$  down the tree to the corresponding terminal node  $\mathbb{R}_m$  and set

$$\hat{f}(\mathbf{x}_0) = \frac{1}{|\mathbb{R}_m|} \sum_{\mathbf{x}_i \in \mathbb{R}_m} y_i$$

the mean value in that terminal node. Regression trees automatically capture non-linear effects and high-order interactions, but a single tree is typically unstable and exhibits high variance, as small changes in the training data can produce quite different trees. This motivates ensemble methods such as bagging and random forests that are introduced as follows.

### 2.3.2 Bagging of regression trees

Bootstrapping is a resampling procedure, which creates a number of new bootstrap samples by drawing samples with replacement of the original training data. This section illustrates how to use bootstrapping to create an ensemble of predictions. Bootstrap aggregating (also called *bagging*; Breiman, 1996) is designed to improve the stability and accuracy of the regression trees algorithm. By model averaging, bagging helps to reduce variance and minimise over-fitting. The basic idea is:

1. Draw  $B$  bootstrap samples from the training data of the same size by sampling  $n$  observations with replacement.
2. For each bootstrap sample  $b = 1, \dots, B$ , fit a regression tree  $\hat{f}^{(b)}(\cdot)$ .
3. For a new covariate vector  $\mathbf{x}_0$ , form the bagged predictor as the average over trees:

$$\hat{f}_{\text{Bag}}(\mathbf{x}_0) = \frac{1}{B} \sum_{b=1}^B \hat{f}^{(b)}(\mathbf{x}_0).$$

Because the bootstrap samples differ the individual trees will vary, and averaging them reduces the variance of the predictor while maintaining low bias. Bagging typically improves predictive accuracy compared to a single tree, at the cost of reduced interpretability of the overall model. In Chapter 3, bagging is implemented with regression trees as base learners to predict air pollution concentrations.

### 2.3.3 Random forests

Random forests (RF; Breiman, 2001) further improve bagged trees by deliberately de-correlating the individual trees. As in bagging, random forests grow many trees on bootstrap samples of the data, but introduce extra randomness at each split. Specifically, for each tree  $b = 1, \dots, B$ :

1. Draw a bootstrap sample from the training data.
2. Grow a regression tree  $\hat{f}^{(b)}(\cdot)$  at each node:
  - (a) Randomly select a subset of  $m_{\text{try}}$  candidates from the  $p$  covariates.
  - (b) Choose the best split using only these  $m_{\text{try}}$  covariates.

The RF predictor is again the average

$$\hat{f}_{\text{RF}}(\mathbf{x}_0) = \frac{1}{B} \sum_{b=1}^B \hat{f}^{(b)}(\mathbf{x}_0).$$

The RF algorithm reduces the correlation between trees by randomly restricting the set of candidate variables at each split, while increasing the variance reduction achieved by averaging. In regression problems, a common default is  $m_{\text{try}} \approx p/3$ , but this can also be treated as a tuning parameter. Key hyperparameters in RFs (also applicable in Bagging) include:

- the number of trees  $B$ ;
- the number of candidate variables to consider at each split  $m_{\text{try}}$ ;
- tree-complexity controls such as minimum node size or maximum depth.

Random forests typically yield strong predictive performance with relatively little tuning and handle complex high-dimensional relationships between predictors and the response. However, standard RFs provide only point predictions and are not able to directly quantify predictive uncertainty.

### Quantile regression forests

Quantile regression forests (QRFs; [Meinshausen and Ridgeway, 2006](#)) extend RFs to estimate the full conditional distribution of  $Y$  given  $\mathbf{X} = \mathbf{x}$ , rather than only its conditional mean. Instead of averaging leaf means across trees, QRFs keep track of the empirical distribution of training responses that “co-occur” with a new point  $\mathbf{x}_0$  in the forest. For each tree  $b$  in the forest, let  $\mathbb{R}_b(\mathbf{x}_0)$  denote the terminal node containing  $\mathbf{x}_0$ , and let

$$I_b(\mathbf{x}_0) = \{i : \mathbf{x}_i \in \mathbb{R}_b(\mathbf{x}_0)\}$$

be the indices of the training observations falling in the same leaf. QRF defines weights  $w_i(\mathbf{x}_0)$  for each training observation by averaging over trees:

$$w_i(\mathbf{x}_0) = \frac{1}{B} \sum_{b=1}^B \frac{\mathbb{I}\{i \in I_b(\mathbf{x}_0)\}}{|I_b(\mathbf{x}_0)|}, \quad i = 1, \dots, n,$$



so that  $w_i(\mathbf{x}_0) \geq 0$  and  $\sum_{i=1}^n w_i(\mathbf{x}_0) = 1$ . These weights define an estimate of the conditional distribution function

$$\hat{F}_{Y|\mathbf{X}}(y|\mathbf{x}_0) = \sum_{i=1}^n w_i(\mathbf{x}_0) \mathbb{I}\{y_i \leq y\}.$$

The estimated conditional  $\alpha$ -quantile is then obtained as

$$\hat{q}_\alpha(\mathbf{x}_0) = \inf\{y : \hat{F}_{Y|\mathbf{X}}(y|\mathbf{x}_0) \geq \alpha\},$$

and prediction intervals follow by taking, for example, the 2.5<sup>th</sup> and 97.5<sup>th</sup> percentiles,  $[\hat{q}_{0.025}(\mathbf{x}_0), \hat{q}_{0.975}(\mathbf{x}_0)]$  for a 95% prediction interval. Here “inf” denotes the *infimum*, meaning the smallest value  $y$  for which the estimated conditional distribution function  $\hat{F}_{Y|\mathbf{X}}(\cdot|\mathbf{x}_0)$  is at least  $\alpha$ . In this thesis, RFs and their QRF extension are used in Chapter 3 as flexible, non-parametric competitors to the parametric and semi-parametric models described earlier. RFs provide accurate point predictions of pollutant concentrations, while QRFs are used to quantify predictive uncertainty through data-driven prediction intervals.

## 2.4 Bayesian hierarchical modelling

### 2.4.1 Introduction

In frequentist statistics, model parameters are treated as fixed but unknown quantities and are typically estimated by maximising the likelihood, yielding the maximum likelihood estimator (MLE). In contrast, Bayesian statistics treats parameters as random variables and assigns them prior distributions that encode information (or uncertainty) about plausible parameter values before observing the data. Combining the prior with a likelihood model yields a posterior distribution, which forms the basis for inference.

Let  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_p)^\top$  denote a parameter vector and suppose  $\mathbf{y} = (y_1, \dots, y_n)^\top$  are  $n$  observations with likelihood  $f(\mathbf{y}|\boldsymbol{\theta})$ . Assigning a prior density  $\pi(\boldsymbol{\theta})$  gives the posterior density

$$\pi(\boldsymbol{\theta}|\mathbf{y}) = \frac{f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{f(\mathbf{y})}, \quad (2.6)$$

where  $f(\mathbf{y}) = \int f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}$  is a normalising constant. In many models  $f(\mathbf{y})$  does not need to be evaluated explicitly, as  $\pi(\boldsymbol{\theta}|\mathbf{y})$  is known up to proportionality.

Bayesian hierarchical models introduce additional layers into this framework. Model parameters are allowed to depend on hyperparameters, which may also have their own priors. This hierarchy allows borrowing strength across related groups or regions; explicit modelling of spatial and temporal dependencies; and incorporation of latent structures such as random effects. Such models are widely used in disease mapping and environmental epidemiology, where the goal is to estimate local risks while accounting for broader population patterns and spatial dependence.

### 2.4.2 Prior distributions

The prior distribution for each model parameter should reflect the researcher's knowledge before observing the data. When prior information is limited, one often uses non-informative or weakly informative priors. For example, a parameter  $\theta$  constrained to  $(0, 1)$  might be assigned a uniform prior, with density  $f(\theta) = 1$  for  $0 < \theta < 1$ , giving equal weight to all values in the interval. For a real-valued parameter, a weakly informative prior might be Gaussian, e.g.  $\theta \sim N(0, 100^2)$ , which expresses a belief that values near zero are somewhat more plausible but still allows a wide range of values due to the large variance.

A particularly useful class of priors are conjugate priors, in which a prior is conjugate for a likelihood if the posterior and prior belong to the same family of distributions. Many common likelihoods are members of the exponential family (see Equation (2.2)), and for each such likelihood there exists a natural conjugate prior (Gelman et al., 2013). Examples include a Beta prior for a Binomial likelihood; a Gamma prior for a Poisson likelihood; and an Inverse-Gamma prior for the variance of a Normal likelihood with known mean, etc. For instance, if  $Y_i \sim \text{Poisson}(\theta)$  for  $i = 1, \dots, n$  independently and  $\theta \sim \text{Gamma}(\alpha, \beta)$ , then

$$\pi(\theta|\mathbf{y}) \sim \text{Gamma}\left(\alpha + \sum_{i=1}^n y_i, \beta + n\right).$$

Conjugate priors can simplify analytical calculations and provide intuition, although modern computation often allows more flexible prior choices.

When one aims to minimise prior influence, non-informative priors such as Jeffreys priors can be used. Jeffreys prior (Jeffreys, 1946) is defined as  $f(\theta) \propto \sqrt{|I(\theta)|}$ , where

$I(\boldsymbol{\theta})$  is the Fisher information, defined as

$$I(\boldsymbol{\theta}) = -\mathbb{E} \left( \frac{\partial^2}{\partial \boldsymbol{\theta}^2} \ln f(\mathbf{y}|\boldsymbol{\theta}) | \boldsymbol{\theta} \right).$$

Such priors are invariant under reparameterisation but can sometimes be improper (do not integrate to one) and therefore must be used with care.

### 2.4.3 Bayesian inference

Inference in Bayesian models is based on the posterior distribution in Equation (2.6). In simple conjugate settings, the posterior can be derived and sampled from directly. In realistic hierarchical models, however, the posterior is rarely available in closed form and maybe high-dimensional and multimodal. Two main approaches are then used: simulation-based methods, primarily Markov chain Monte Carlo (MCMC, [Robert and Casella, 2005](#); [Gelman et al., 2013](#)) simulation; and deterministic approximations, such as integrated nested Laplace approximation (INLA, [Rue et al., 2009](#)) for latent Gaussian models.

#### Markov chain Monte Carlo methods

MCMC methods construct a Markov chain whose stationary distribution is the target posterior  $\pi(\boldsymbol{\theta}|\mathbf{y})$ . After an initial burn-in period, samples from the chain are treated as draws from the posterior and can be used to compute posterior means, medians, variances, and credible intervals. Let  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_p)^\top$ , possibly partitioned into  $d$  blocks  $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_d)$ . The chain is initialised at starting values  $\boldsymbol{\theta}^{(0)}$ , which can affect the number of iterations needed to reach stationarity. Running multiple chains from dispersed starting points is often recommended ([Brooks et al., 2011](#)). Here the values are updated iteratively, and at iteration  $i$  the new state  $\boldsymbol{\theta}^{(i)}$  depends only on the current state  $\boldsymbol{\theta}^{(i-1)}$  and not earlier values, which is known as the *Markov property*. Under mild conditions, the distribution of the sampled values converges to the true but unknown posterior density  $\pi(\boldsymbol{\theta}|\mathbf{y})$  as  $i \rightarrow \infty$ . Two widely used MCMC algorithms are the Gibbs sampler ([Geman and Geman, 1984](#)) and the Metropolis-Hastings (M-H) algorithm ([Metropolis et al., 1953](#); [Hastings, 1970](#)).

**Gibbs sampling** Gibbs sampling is used when the full conditional distributions of parameters are available in standard forms. At each iteration, the algorithm cycles through parameters (or parameter blocks), sampling each from its conditional

distribution given the current values of the others. The Gibbs sampler to draw  $S$  samples proceeds as:

1. Initialise  $\boldsymbol{\theta}^{(0)} = (\boldsymbol{\theta}_1^{(0)}, \dots, \boldsymbol{\theta}_d^{(0)})$ .
2. For iterations  $i = 1, \dots, S$  and for each block  $j = 1, \dots, d$ , draw

$$\boldsymbol{\theta}_j^{(i)} \sim \pi(\boldsymbol{\theta}_j | \boldsymbol{\theta}_{-j}^{(i)}, \mathbf{y}),$$

where  $\boldsymbol{\theta}_{-j}^{(i)}$  denotes all components of  $\boldsymbol{\theta}$  except  $\boldsymbol{\theta}_j$ , using the most recent updates:

$$\boldsymbol{\theta}_{-j}^{(i)} = (\boldsymbol{\theta}_1^{(i)}, \dots, \boldsymbol{\theta}_{j-1}^{(i)}, \boldsymbol{\theta}_{j+1}^{(i-1)}, \dots, \boldsymbol{\theta}_d^{(i-1)}).$$

This works particularly well in hierarchical models with conjugate priors, where full conditional distributions are known and easy to sample. When full conditionals are not analytically tractable, Gibbs sampling is not directly applicable and one can use more general methods such as Metropolis-Hastings.

**Metropolis-Hastings algorithm** The Metropolis-Hastings (M-H) algorithm is a more general MCMC method that can be used when full conditional distributions are only known up to proportionality (Metropolis et al., 1953; Hastings, 1970). For simplicity, consider updating a single block  $\boldsymbol{\theta}_j$  at iteration  $i$ :

1. Given the current value  $\boldsymbol{\theta}_j^{(i-1)}$ , propose a candidate  $\boldsymbol{\theta}_j^*$  from a proposal (or jumping) distribution  $g(\boldsymbol{\theta}_j^* | \boldsymbol{\theta}_j^{(i-1)})$ .
2. Compute the acceptance probability

$$p = \min \left\{ 1, \frac{\pi(\boldsymbol{\theta}_j^* | \mathbf{y}) g(\boldsymbol{\theta}_j^{(i-1)} | \boldsymbol{\theta}_j^*)}{\pi(\boldsymbol{\theta}_j^{(i-1)} | \mathbf{y}) g(\boldsymbol{\theta}_j^* | \boldsymbol{\theta}_j^{(i-1)})} \right\}.$$

3. Draw  $U \sim \text{Uniform}(0, 1)$ . If  $U \leq p$ , accept the proposal and set  $\boldsymbol{\theta}_j^{(i)} = \boldsymbol{\theta}_j^*$ ; otherwise, reject it and set  $\boldsymbol{\theta}_j^{(i)} = \boldsymbol{\theta}_j^{(i-1)}$ .

When the proposal is symmetric, i.e.  $g(\boldsymbol{\theta}_j^* | \boldsymbol{\theta}_j^{(i-1)}) = g(\boldsymbol{\theta}_j^{(i-1)} | \boldsymbol{\theta}_j^*)$  for all states, the acceptance probability simplifies to the Metropolis algorithm (Metropolis et al., 1953):

$$p = \min \left\{ 1, \frac{\pi(\boldsymbol{\theta}_j^* | \mathbf{y})}{\pi(\boldsymbol{\theta}_j^{(i-1)} | \mathbf{y})} \right\}.$$

A common choice for  $g(\cdot)$  is a Gaussian random-walk proposal with mean equal to the current value and a covariance matrix. In such a step if the proposal variance is too small, most proposals are accepted but steps are tiny and the chain explores the posterior slowly. But if the proposal variance is too large, proposals are often far from high-probability regions and are frequently rejected, leading to a very low acceptance rate. Empirical and theoretical work suggests targeting moderate acceptance (e.g. around 40–50% for scalar updates and around 20–30% for higher-dimensional blocks) to balance exploration and acceptance (see, for example, [Gelman et al., 1996](#); [Robert and Casella, 2005](#)). In practice, proposal variances are often tuned during pilot runs, or adaptively by software. In addition, parameters can be updated individually or in blocks. Blocking parameters with strong posterior dependence can improve mixing, but very large blocks may reduce acceptance and hinder exploration.

**Model convergence and diagnostics** MCMC-based inference is only valid once the Markov chain has converged to its stationary distribution, i.e. the posterior  $\pi(\theta|\mathbf{y})$ . Because early iterations may be influenced by arbitrary starting values, it is common to discard an initial burn-in period as these early iterations may not reflect the true posterior density. Since successive draws are correlated, one sometimes thins the chain by retaining every  $k$ -th sample to reduce storage. Note that thinning is optional and does not improve statistical efficiency if the effective sample size is already adequate. The required burn-in length and total number of iterations depend on model complexity and sampler efficiency. Rules of thumb are unreliable; instead, decisions should be guided by convergence diagnostics and experience with similar models.

Convergence is typically assessed by:

- Trace plots: plotting sampled values against iteration. A well-mixed chain should fluctuate around a stable mean without visible trends or long-term drift;
- Multiple chains: running several chains from dispersed starting values and visually checking that they all appear to mix and settle in the same region of parameter space;
- Formal diagnostics: the Gelman–Rubin statistic  $\hat{R}$  compares between-chain and within-chain variability. Suppose  $m$  parallel chains are run, each of length  $L$  after burn-in, and let  $\theta_{j\ell}$  denote the  $\ell$ -th draw from chain  $j$  for a given scalar

parameter. Define the chain means and the overall mean as

$$\bar{\theta}_j = \frac{1}{L} \sum_{\ell=1}^L \theta_{j\ell} \quad \text{and} \quad \bar{\theta} = \frac{1}{m} \sum_{j=1}^m \bar{\theta}_j,$$

the within-chain variance as

$$W = \frac{1}{m} \sum_{j=1}^m s_j^2 \quad \text{and} \quad s_j^2 = \frac{1}{L-1} \sum_{\ell=1}^L (\theta_{j\ell} - \bar{\theta}_j)^2,$$

and the between-chain variance as

$$B = \frac{L}{m-1} \sum_{j=1}^m (\bar{\theta}_j - \bar{\theta})^2.$$

An estimate of the Gelman-Rubin statistic is then

$$\hat{R} = \sqrt{\frac{\frac{L-1}{L}W + \frac{1}{L}B}{W}}.$$

Values of  $\hat{R}$  substantially larger than 1 indicate that between-chain variability exceeds within-chain variability, suggesting lack of convergence. If chains have converged, these should be similar and  $\hat{R}$  should be close to 1 (typically  $< 1.1$  is used as a rule of thumb; [Gelman and Rubin, 1992](#)).

Once convergence is established and a sufficient number of post-burn-in samples have been collected, posterior inference is straightforward. Point summaries such as posterior means and medians are obtained by averaging across samples, and uncertainty is naturally quantified via credible intervals based on posterior quantiles. For example, the 2.5<sup>th</sup> and 97.5<sup>th</sup> percentiles of the empirical posterior for a parameter provide a 95% credible interval. A key advantage of the Bayesian framework is that inference is based on the entire posterior distribution rather than solely on point estimates.

### Integrated nested Laplace approximation

Integrated nested Laplace approximation (INLA; [Rue et al., 2009](#)) is a deterministic alternative to MCMC for Bayesian inference, specifically developed for latent Gaussian models (LGMs), in which the unobserved components (latent effects) are assigned Gaussian prior distributions. LGMs include a wide range of commonly used

models, including Gaussian processes for time series, spatial and spatio-temporal smoothing models or grouped random effects models. While MCMC can, in principle, approximate posterior distributions arbitrarily well if run for sufficiently long, it can be prohibitively slow for complex hierarchical models and for large spatial or spatio-temporal datasets. INLA replaces simulation by a sequence of Laplace-type approximations to obtain accurate numerical approximations to the marginal posterior distributions of both latent effects and hyperparameters, and is therefore particularly well suited to the complicated spatio-temporal smoothing models considered in this thesis.

INLA is applicable when the model can be expressed in the latent Gaussian framework. Let  $\mathbf{y}$  denote the observed data,  $\mathbf{x}$  the latent field, and  $\boldsymbol{\theta}$  a vector of hyperparameters. A model is an LGM if (i) the observations are conditionally independent given the latent field and hyperparameters  $y_i \perp y_j | \mathbf{x}, \boldsymbol{\theta}$ , and (ii) the latent field has a Gaussian Markov random field (GMRF; [Rue and Held, 2005](#)) prior. A generic specification is

$$y_i \sim f(\eta_i, \boldsymbol{\theta}), \quad \boldsymbol{\eta} = \mathbf{A}\mathbf{x}, \quad \mathbf{x} | \boldsymbol{\theta} \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}^{-1}), \quad \boldsymbol{\theta} \sim g(\boldsymbol{\theta}).$$

Here  $f(\cdot)$  is a likelihood from a broad class (e.g. Gaussian, Poisson, Binomial, Gamma),  $\mathbf{x}$  contains all the fixed effects and random effects,  $\mathbf{A}$  is a typically sparse design matrix linking the latent field to the linear predictor, and  $\mathbf{Q}$  is a sparse precision matrix encoding the dependence structure of the GMRF (for example, a conditional autoregressive (CAR) or autoregressive (AR) structure). The vector  $\boldsymbol{\theta}$  contains hyperparameters from both the likelihood and the latent field (such as variances, range parameters, or autoregressive coefficients), with prior density  $g(\boldsymbol{\theta})$ . Note that the Gaussian assumption is imposed only on the prior for  $\mathbf{x}$ , not on the data model.

The INLA methodology can be summarised in two main stages. First, it approximates the posterior distribution of the hyperparameters,  $\pi(\boldsymbol{\theta} | \mathbf{y})$ . Starting from the joint posterior

$$\pi(\mathbf{x}, \boldsymbol{\theta} | \mathbf{y}) \propto \pi(\boldsymbol{\theta}) \pi(\mathbf{x} | \boldsymbol{\theta}) \prod_{i=1}^n f(y_i | \eta_i, \boldsymbol{\theta}).$$

INLA constructs a Gaussian (Laplace) approximation  $\pi_G(\mathbf{x} | \mathbf{y}, \boldsymbol{\theta})$  to the conditional posterior  $\pi(\mathbf{x} | \mathbf{y}, \boldsymbol{\theta})$  around its mode  $\boldsymbol{\mu}(\boldsymbol{\theta})$ . Substituting this approximation yields

$$\tilde{\pi}(\boldsymbol{\theta} | \mathbf{y}) = \frac{\pi(\mathbf{x}, \boldsymbol{\theta} | \mathbf{y})}{\pi_G(\mathbf{x} | \mathbf{y}, \boldsymbol{\theta})} \Big|_{\mathbf{x}=\boldsymbol{\mu}(\boldsymbol{\theta})},$$

which provides a numerically tractable approximation to  $\pi(\boldsymbol{\theta}|\mathbf{y})$ . Marginal posteriors for individual components  $\theta_j$  are then obtained by low-dimensional integration,

$$\tilde{\pi}(\theta_j|\mathbf{y}) = \int \tilde{\pi}(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta}_{-j},$$

where  $\boldsymbol{\theta}_{-j}$  denotes all elements of  $\boldsymbol{\theta}$  except  $\theta_j$ . In practice, this integration is carried out over a finite set of integration points (e.g. a central composite design) with associated numerical weights.

In the second stage, INLA approximates the conditional posterior of the latent field  $\mathbf{x}$  given  $\boldsymbol{\theta}$  and  $\mathbf{y}$  by a multivariate Gaussian distribution derived from the GMRF like

$$\pi(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y}) \approx \pi_G(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y}) = \mathcal{N}\left(\boldsymbol{\mu}(\boldsymbol{\theta}), \mathbf{Q}_x^{-1}(\boldsymbol{\theta})\right),$$

where  $\mathbf{Q}_x(\boldsymbol{\theta})$  is a sparse precision matrix. Marginal posteriors for each latent component  $x_j$  are obtained by integrating out the hyperparameters using the same set of integration points  $\boldsymbol{\theta}_k$  with area weights  $\Delta_k$ :

$$\tilde{\pi}(x_j|\mathbf{y}) = \int \pi_G(x_j|\boldsymbol{\theta}, \mathbf{y}) \tilde{\pi}(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta} \approx \sum_{k=1}^K \pi_G(x_j|\boldsymbol{\theta}_k, \mathbf{y}) \tilde{\pi}(\boldsymbol{\theta}_k|\mathbf{y}) \Delta_k.$$

Because the precision matrices are sparse, these computations remain efficient even when the latent field is high-dimensional, as is the case for the spatio-temporal random effects in this thesis.

The original formulation of INLA relied solely on Laplace approximations. In some non-Gaussian likelihood settings this can lead to biased posterior means for  $\mathbf{x}$  and  $\boldsymbol{\eta}$ , because the Gaussian assumption of the conditional posterior might be inaccurate (Van Niekerk et al., 2023). Recent developments, implemented as the default in the R-INLA package (Lindgren and Rue, 2015), combine INLA with a low-rank variational Bayes correction (Van Niekerk et al., 2023; Van Niekerk and Rue, 2024). This refinement adjusts the Gaussian approximation to  $\pi(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})$  by improving the posterior mean of the latent field while preserving the sparse precision structure. Empirical studies suggest that, for a wide class of LGMs, this yields accuracy comparable to MCMC, but with substantially reduced computational cost.



## 2.5 Model comparison

A number of methods have been proposed to evaluate and compare the fit of alternative statistical models applied to the same data set. The likelihood function measures how plausible the observed data are under a given model, and increasing the number of parameters typically increases the maximised likelihood. However, models with many parameters risk over-fitting, in which they may reproduce the observed data very well but perform poorly for new data. In practice, the modelling aim is to identify a model that provides an adequate fit to the data while remaining as parsimonious as possible. Model comparison criteria formalise this trade-off by combining a measure of fit (typically derived from the log-likelihood) with a penalty for model complexity.

Typically, these criteria take the form “goodness of fit + penalty”, where the penalty increases with model complexity. In this thesis, for non-Bayesian models I introduce the Akaike information criterion (AIC) and the Bayesian information criterion (BIC), which are based on maximum likelihood estimation and asymptotic arguments. For fully Bayesian hierarchical models, I focus on the deviance information criterion (DIC) and the Watanabe-Akaike information criterion (WAIC, also known as the widely applicable information criterion). For all criteria considered here, smaller values indicate models with better performance. In the following subsections, let  $\mathbf{y} = (y_1, \dots, y_n)^\top$  denote the observed data and  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_p)^\top$  denote the vector of model parameters.

### 2.5.1 Akaike information criterion

The Akaike information criterion (AIC; [Akaike, 1974](#)) is defined as

$$\text{AIC} = -2 \ln L(\hat{\boldsymbol{\theta}}_{\text{MLE}}) + 2p,$$

where  $L(\hat{\boldsymbol{\theta}}_{\text{MLE}})$  is the likelihood evaluated at the maximum likelihood estimate (MLE) of  $\boldsymbol{\theta}$ , and  $p$  is the number of estimated parameters in the model. The term  $-2 \ln L(\hat{\boldsymbol{\theta}}_{\text{MLE}})$  measures lack of fit (smaller is better), while the penalty  $2p$  adjusts for the optimistic bias in the in-sample fit when estimating  $p$  parameters. The model with the smallest AIC is preferred.

### 2.5.2 Bayesian information criterion

The Bayesian information criterion (BIC; [Schwarz, 1978](#)) is given by

$$\text{BIC} = -2 \ln L(\hat{\boldsymbol{\theta}}_{\text{MLE}}) + p \ln(n),$$

where  $n$  is the sample size. As with AIC, the first term measures lack of fit, while the second term penalises model complexity. Compared with AIC, BIC imposes a stronger penalty on models with many parameters, especially when  $n$  is large, because  $p \ln(n) > 2p$  for sufficiently large  $n$ . Consequently, BIC tends to favour simpler models than AIC, particularly in large samples.

### 2.5.3 Deviance information criterion

The deviance information criterion (DIC; [Spiegelhalter et al., 2002](#)) extends AIC idea to a Bayesian setting by replacing the MLE  $\hat{\boldsymbol{\theta}}_{\text{MLE}}$  with the posterior mean  $\hat{\boldsymbol{\theta}}_{\text{Bayes}} = \mathbb{E}(\boldsymbol{\theta}|\mathbf{y})$  and by defining an “effective number of parameters”. It is based on the deviance

$$D(\boldsymbol{\theta}) = -2 \ln L(\boldsymbol{\theta}) + C,$$

where  $C$  is a constant that does not depend on  $\boldsymbol{\theta}$  and cancels out in differences. Let  $\{\boldsymbol{\theta}^{(1)}, \dots, \boldsymbol{\theta}^{(S)}\}$  be posterior draws from an MCMC algorithm. Then the DIC can be computed as

$$\text{DIC} = \bar{D} + p_{\text{DIC}}, \quad \text{where } \bar{D} = \frac{1}{S} \sum_{s=1}^S D(\boldsymbol{\theta}^{(s)}).$$

Here  $\bar{D}$  represents the average fit over the posterior distribution, while  $p_{\text{DIC}}$  is the effective number of parameters defined as

$$p_{\text{DIC}} = \bar{D} - D(\hat{\boldsymbol{\theta}}_{\text{Bayes}}) = 2 \left( \ln L(\hat{\boldsymbol{\theta}}_{\text{Bayes}}) - \frac{1}{S} \sum_{s=1}^S \ln L(\boldsymbol{\theta}^{(s)}) \right).$$

This measures the extent to which the posterior mean of the parameters over-fits the data relative to the average posterior fit. Thus the criterion can be finally given as

$$\text{DIC} = -2 \ln L(\hat{\boldsymbol{\theta}}_{\text{Bayes}}) + 2p_{\text{DIC}},$$

which is relatively easy to compute from MCMC output or from approximate Bayesian methods such as INLA, and has been widely used in hierarchical modelling.

### 2.5.4 Watanabe-Akaike information criterion

The Watanabe-Akaike criterion (WAIC; [Watanabe and Opper, 2010](#)) is a fully Bayesian generalisation of AIC that operates based on the pointwise predictive distribution rather than a single summary such as the posterior mean in DIC. Under fairly general conditions, WAIC is asymptotically equivalent to leave-one-out cross-validation. It estimates the expected log pointwise predictive density (elpd) for new data, by computing the log pointwise posterior predictive density (lppd) and then adding a correction for the effective number of parameters  $p_{\text{WAIC}}$  ([Gelman et al., 2014](#)). The WAIC is defined as

$$\text{WAIC} = -2(\text{lppd} - p_{\text{WAIC}}).$$

Let  $p(y_i|\boldsymbol{\theta})$  denote the likelihood of observation  $y_i$ , and let  $\boldsymbol{\theta}^{(1)}, \dots, \boldsymbol{\theta}^{(S)}$  be posterior draws. The log pointwise predictive density is approximated by

$$\text{lppd} = \sum_{i=1}^n \ln \left( \int p(y_i|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta} \right) \approx \sum_{i=1}^n \ln \left( \frac{1}{S} \sum_{s=1}^S p(y_i|\boldsymbol{\theta}^{(s)}) \right),$$

and the effective number of parameters is estimated as

$$p_{\text{WAIC}} = \sum_{i=1}^n \text{Var}_{\text{post}} (\ln p(y_i|\boldsymbol{\theta})) \approx \sum_{i=1}^n \text{Var}_{s=1}^S (\ln p(y_i|\boldsymbol{\theta}^{(s)})).$$

## 2.6 Spatial and spatio-temporal modelling

Spatial statistics involves the analysis and modelling of data associated with specific geographic locations. A fundamental feature distinguishing spatial data, and analogously time series data, from standard statistical data is that observations that are close in space or time tend to be more similar than those further apart. This phenomenon follows Tobler's *First Law of Geography*, which states that “Everything is related to everything else, but near things are more related than distant things” ([Tobler, 1970](#)). Consequently, statistical models must explicitly account for this autocorrelation to ensure valid inference.

Spatial data are typically categorised into three distinct types: point process data, geostatistical data, and areal unit (or lattice) data ([Cressie, 2015](#)). Point process data arise when the spatial locations of events are the random variables of interest (e.g., the specific locations of trees in a forest). Geostatistical data (also called point-referenced data) consist of observations measured at specific locations. Finally, areal unit data

are aggregated over a set of non-overlapping geographical regions, such as census tracts or electoral wards. In this thesis, geostatistical data arise in the environmental context (e.g., predicted air pollution surfaces), while areal unit data arise in the disease context (e.g., health outcomes aggregated to Data Zones or GP catchment areas).

The remainder of this section reviews concepts and models that are central to the spatio-temporal analysis and disease mapping undertaken in later chapters. Section 2.6.1 discusses how to quantify spatial proximity via neighbourhood matrices. Section 2.6.2 reviews Moran's I as a measure of spatial autocorrelation. Section 2.6.3 presents the spatio-temporal Gaussian process models used for air pollution prediction. Section 2.6.4 introduces basic disease mapping concepts like standardised ratios. Section 2.6.5 presents commonly used spatial models for areal unit data, focusing on conditional autoregressive (CAR) priors, while Section 2.6.6 reviews several spatio-temporal extensions.

### 2.6.1 Defining spatial closeness

For geostatistical (point-referenced) data, spatial closeness is characterised through distances between locations and selected covariance functions. Letting  $\mathbf{s}_i$  and  $\mathbf{s}_j$  denote two locations, their Euclidean distance can be expressed as  $d_{ij} = \|\mathbf{s}_i - \mathbf{s}_j\|$ . Correlation typically decreases with increasing distance, and a general and widely used class of distance-based dependence functions is given by the Matérn family, which is the standard choice for covariance or correlation functions in geostatistics (Cressie, 2015). For an isotropic process with unit marginal variance, the Matérn correlation between two locations  $\mathbf{s}_i$  and  $\mathbf{s}_j$  can be written as

$$\text{Corr}(d_{ij}; \nu, \rho) = \frac{(\rho d_{ij})^\nu K_\nu(\rho d_{ij})}{2^{\nu-1} \Gamma(\nu)},$$

where  $\nu > 0$  is a smoothness parameter,  $\rho > 0$  is a range (or scale) parameter,  $\Gamma(\cdot)$  is the Gamma function, and  $K_\nu(\cdot)$  is the modified Bessel function of order  $\nu$ . The parameter  $\nu$  controls the smoothness of the underlying spatial field (larger  $\nu$  corresponds to smoother realisations), while  $\rho$  determines how quickly correlation decays with distance. Common models such as the exponential and Gaussian correlations arise as special or limiting cases of the Matérn family. For example, the exponential correlation

function corresponds to  $\nu = 1/2$  and has the form

$$\rho_{\text{exp}}(d_{ij}; \rho) = \exp(-\rho d_{ij}),$$

while the Gaussian correlation function can be written as

$$\rho_{\text{Gauss}}(d_{ij}; \rho) = \exp\{-(\rho d_{ij})^2\}.$$

For areal unit data, spatial models require a formal representation of which areas are “close” to each other. This spatial structure is encoded through a *neighbourhood matrix* (or adjacency matrix)  $\mathbf{W}$ , which is an  $n \times n$  non-negative matrix. Here the element  $w_{ij}$  quantifies the spatial closeness between areas  $\mathcal{A}_i$  and  $\mathcal{A}_j$ . In this thesis I adopt the following conventions: (i)  $w_{ii} = 0$  for all  $i$ , since an area is not considered a neighbour of itself; (ii)  $w_{ij} = 0$  implies that  $\mathcal{A}_i$  and  $\mathcal{A}_j$  are not treated as neighbours; and  $w_{ij} > 0$  implies that  $\mathcal{A}_i$  and  $\mathcal{A}_j$  are neighbours, with larger values indicating stronger spatial closeness. Typically the matrix is also taken to be symmetric ( $w_{ij} = w_{ji}$ ), which ensures that the neighbourhood relationship is mutual.

The most common specification in disease mapping is a binary neighbourhood matrix, where  $w_{ij} = 1$  if  $\mathcal{A}_i$  and  $\mathcal{A}_j$  are neighbours and 0 otherwise. Two main definitions of “neighbour” are widely used: (i)  $\mathcal{A}_i$  and  $\mathcal{A}_j$  are neighbours if they share a common boundary (first-order contiguity); or (ii) the centroid of  $\mathcal{A}_j$  (defined as  $\mathbf{s}_j$ ) is among the  $k$  nearest centroids to  $\mathcal{A}_i$  (typically according to Euclidean distance). For the latter construction, the resulting matrix is generally directed (area  $\mathcal{A}_j$  can be among the  $k$  nearest neighbours of  $\mathcal{A}_i$  without the converse holding), and is usually made symmetric by setting  $w_{ij} = w_{ji} = 1$  whenever either  $\mathcal{A}_j$  is a neighbour of  $\mathcal{A}_i$  or  $\mathcal{A}_i$  is a neighbour of  $\mathcal{A}_j$ , and 0 otherwise. In both cases,  $\mathbf{W}$  is typically sparse, meaning that most entries are zero.

An alternative is to define  $\mathbf{W}$  using continuous measures of distance between centroids ( $\mathbf{s}_i, \mathbf{s}_j$ ). In principle, one can construct a continuous neighbourhood matrix by setting, for example,  $w_{ij} = \rho(d_{ij}; \nu, \rho)$ , possibly scaled or transformed, so that the weights directly reflect a Matérn-type spatial correlation structure. However, a pure Matérn specification yields a dense matrix, because every pair of areas has a non-zero correlation. This can substantially increase computational burden and is often unnecessary if long-range interactions are weak. As with simpler distance-based weights, it is therefore common to threshold the weights in practice, for example by setting  $w_{ij} = 0$  whenever the distance exceeds a fixed cut-off or whenever the weight falls below a small tolerance. This yields a sparse matrix that still reflects graded

spatial closeness rather than simple adjacency.

### 2.6.2 Spatial autocorrelation

Spatial autocorrelation refers to the tendency for observations from nearby areas to be more similar (or dissimilar) than would be expected under spatial independence (Cressie, 2015). Global spatial autocorrelation can be summarised by a single statistic that measures the overall strength of spatial clustering or spatial dispersion. A widely used measure is Moran's I (Moran, 1950), which has a similar form to Pearson's correlation coefficient and the lag 1 autocorrelation in time series. For data  $\mathbf{y} = (y_1, \dots, y_n)^\top$  with mean  $\bar{y} = \sum_{i=1}^n y_i/n$  and a given neighbourhood matrix  $\mathbf{W} = \{w_{ij}\}$ , Moran's I is defined as

$$I = \frac{n \sum_{i=1}^n \sum_{j=1}^n w_{ij} (y_i - \bar{y})(y_j - \bar{y})}{\left( \sum_{i=1}^n \sum_{j=1}^n w_{ij} \right) \sum_{i=1}^n (y_i - \bar{y})^2}.$$

The numerator is a spatially weighted covariance, and the denominator is a scaled total variance. It measures the strength of the linear relationship in the spatial data, suitably weighted for their spatial locations. Here  $I \approx 0$  means no spatial correlation,  $I > 0$  means positive spatial autocorrelation (similar values cluster together), while  $I < 0$  indicates negative spatial correlation (neighbouring areas tend to be dissimilar).

To formally assess whether observed spatial autocorrelation is statistically different to what would be expected under a randomisation of no spatial structure, a permutation (Monte Carlo) test is commonly used. The basic idea is to compare the Moran's I computed from the original data to a reference distribution obtained by repeatedly permuting the observed values randomly over the set of areas while keeping the neighbourhood structure fixed. For each permuted dataset, Moran's I is recalculated, producing a large collection of values that represent what would be expected under spatial randomness. The position of the observed Moran's I within this empirical distribution then indicates how unusual it is. If only a small proportion of permuted datasets yield a value as extreme as, or more extreme than, the observed Moran's I statistic, there is evidence of statistically significant spatial autocorrelation.

### 2.6.3 Spatio-temporal Gaussian process models

Spatio-temporal Gaussian process (GP) models are designed for geostatistical (point-referenced) data, which are widely used for environmental applications such as air

pollution (e.g., [Wikle, 2003](#); [Banerjee et al., 2014](#); [Sahu, 2022](#)). They provide a flexible way to model smoothly varying dependence in space and time, while separating measurement error from the underlying latent process of interest. In this subsection, they are used for modelling and predicting ambient air pollution concentrations in Chapter 3, and follow the hierarchical Bayesian framework introduced in Section 2.4.

Let  $Y_{it}$  denote the observed response at spatial location  $\mathbf{s}_i \in \mathcal{D} \subset \mathbb{R}^2$  for  $i = 1, \dots, n$  and discrete time  $t = 1, \dots, T$ . A typical spatio-temporal GP model can be written as

$$\begin{aligned} Y_{it}|O_{it} &\sim \mathcal{N}(O_{it}, \sigma^2), \\ O_{it} &= \mathbf{x}_{it}^\top \boldsymbol{\beta} + \omega_{it}, \end{aligned} \tag{2.7}$$

where  $O_{it}$  is the latent “true” spatio-temporal process of interest,  $\mathbf{x}_{it}$  is a vector of covariates with regression coefficients  $\boldsymbol{\beta}$ , and  $\omega_{it}$  is a zero-mean spatio-temporal random effect. The term  $\epsilon_{it} = Y_{it} - O_{it}$  is often interpreted as a nugget effect, capturing measurement error and unresolved small-scale variability. The key modelling choice is the specification of the joint spatio-temporal structure of  $\omega_{it}$ . Define a covariance function

$$\text{Cov}\{\omega_{it}, \omega_{it'}\} = \sigma_\omega^2 C((\mathbf{s}, t), (\mathbf{s}', t'); \boldsymbol{\psi}).$$

where  $\sigma_\omega^2$  is a marginal variance parameter and  $\boldsymbol{\psi}$  collects the correlation parameters. A common simplification is to assume a *separable* covariance structure,

$$C((\mathbf{s}, t), (\mathbf{s}', t'); \boldsymbol{\psi}) = C_s(\mathbf{s}, \mathbf{s}'; \boldsymbol{\psi}_s) C_t(t, t'; \boldsymbol{\psi}_t),$$

where  $C_s(\cdot, \cdot)$  is a spatial correlation function (often taken from the exponential or Matérn families introduced in Section 2.6.1), and  $C_t(\cdot, \cdot)$  is a purely temporal correlation function.

**Temporally independent spatial Gaussian process model.** A simple but widely-used special case assumes that the spatial process is independent across time, so that temporal dependence is ignored. In this case,

$$C_t(t, t'; \boldsymbol{\psi}_t) = \begin{cases} 1, & t = t', \\ 0, & t \neq t', \end{cases}$$

so that  $\boldsymbol{\omega}_t = (\omega_{1t}, \dots, \omega_{nt})$  is a spatial GP at each time  $t$  and is independent over  $t$ . That is,  $\boldsymbol{\omega}_t \sim N(\mathbf{0}, \boldsymbol{\Sigma}_w = \sigma_w^2 \mathbf{S}_w)$ , where  $\sigma_w^2$  is the marginal spatial variance and  $\mathbf{S}_w$  is an  $n \times n$  spatial correlation matrix with entries

$$(\mathbf{S}_w)_{ij} = C_s(\mathbf{s}_i, \mathbf{s}_j; \boldsymbol{\psi}_s), \quad \text{for } i, j = 1, \dots, n.$$

This model is appropriate when spatial dependence is strong but residual temporal dependence (after accounting for covariates) is weak. It is also computationally convenient because the likelihood factorises over time.

**Spatio-temporal autoregressive Gaussian process model.** Another common approach is to impose a first-order autoregressive (AR(1)) evolution in time for the latent process  $\{O_{it}\}$ , which explicitly accommodates temporal dependence by replacing the linear predictor in Equation (2.7) with

$$O_{it} = \alpha O_{i,t-1} + \mathbf{x}_{it}^\top \boldsymbol{\beta} + \omega_{it},$$

where  $|\alpha| < 1$  is a temporal autoregressive parameter, such that when  $\alpha = 0$  this AR model reduces to the temporally independent GP model. In this thesis, spatio-temporal GP models of the independent and AR(1) temporal forms described above are used as part of the model comparison for air pollution prediction in Chapter 3, where they are implemented using the `spTimer` package (Bakar and Sahu, 2015) within a fully Bayesian framework.

## 2.6.4 Disease mapping

A central goal in epidemiology and public health is to understand how disease risk varies over space and time, and how such variation relates to potential risk factors such as environmental exposures, socio-economic deprivation, or demographic structure. *Disease maps* are commonly used to visualise spatial patterns in disease incidence, prevalence or mortality, by displaying relative risk estimates for each areal unit (Lawson, 2018; MacNab, 2022). Early work by Clayton and Kaldor (1987) formalised the concept of disease mapping, emphasising the importance of borrowing strength across neighbouring areas via spatial smoothing.

In practice, health data are almost always available in aggregated form for pre-defined geographical units, rather than for individuals, due to confidentiality and reporting constraints. For high-prevalence diseases, relatively small areas can be used



to represent nearly homogeneous communities, whereas for rare diseases larger areas may be needed to obtain stable counts. In Scotland, commonly used spatial units include Data Zones (6,976 areas), Intermediate Zones (1,279 areas), Council Areas (32 areas) and Health Boards (14 areas). The chosen spatial resolution constrains both the statistical model and the inferences that can be drawn. In particular, ecological analyses that relate area-level disease measures to area-level exposures are subject to ecological bias, because within-area variability in exposures and individual-level confounding are ignored. This bias tends to increase as the areas become larger and more heterogeneous.

To introduce notation, consider a study region  $\mathcal{A} = \{\mathcal{A}_1, \dots, \mathcal{A}_n\}$  partitioned into  $n$  non-overlapping areal units, and let  $y_i$  denote the observed number of disease events in area  $\mathcal{A}_i$ . Directly comparing the raw counts  $\{y_i\}$  across areas is misleading, because both population size and demographic composition vary between areas. For example, chronic obstructive pulmonary disease (COPD) is substantially more common in older individuals (Chapman et al., 2006; Raheison and Girodet, 2009), so areas with more elderly populations will naturally exhibit higher counts even if their underlying risk is not elevated. A standard approach to adjust for such confounding is *indirect standardisation*. Suppose the population in area  $\mathcal{A}_i$  is stratified into  $J$  age-sex groups, and let  $n_{ij}$  denote the population size in area  $\mathcal{A}_i$  and stratum  $j$  ( $j = 1, \dots, J$ ). Let  $r_j$  denote the disease rate for stratum  $j$  in a suitable reference population (for example, the national population). The expected number of cases in area  $\mathcal{A}_i$  under the reference rate is

$$e_i = \sum_{j=1}^J n_{ij} r_j,$$

which represents how many cases would be expected if  $\mathcal{A}_i$  had the same age–sex-specific rates as the reference population.

A commonly used summary of disease risk is the *standardised morbidity/mortality ratio* (SMR), sometimes also referred to as the *standardised disease ratio*, defined as

$$\text{SMR}_i = \frac{y_i}{e_i}.$$

An SMR greater than one indicates that more cases are observed than expected under the reference rates (suggesting elevated risk), while an SMR less than one indicates fewer cases than expected (suggesting reduced risk). For example,  $\text{SMR} = 1.2$  corresponds to a 20% higher risk than the reference level, while  $\text{SMR} = 0.8$  corresponds to a 20% lower risk. Although SMRs are intuitive and easy to compute,

they can be highly variable when the expected counts  $e_i$  are small, as is the case for sparsely populated areas or rare diseases. In such settings, extreme SMR values may arise purely from random variation. To stabilise these health risk estimates and quantify the uncertainty, modern disease mapping studies typically model the counts directly using statistical models.

For example, a standard likelihood assumption is

$$Y_i \sim \text{Poisson}(e_i\theta_i), \quad \text{for } i = 1, \dots, n,$$

where  $\theta_i$  is the latent relative risk in area  $\mathcal{A}_i$ . The SMR can be viewed as the maximum likelihood estimate of  $\theta_i$  under the Poisson model. Strictly speaking, one could model the counts using a binomial distribution  $Y_i \sim \text{Binomial}(n_i, \theta_i)$ , where  $n_i$  is the population size in  $\mathcal{A}_i$  and  $\theta_i$  is a latent incidence rate or prevalence. However, for rare diseases with typically  $y_i \ll n_i$ , the Poisson model is generally a good approximation and is more convenient for modelling purposes. In addition, to overcome over-dispersion ( $\text{Var}(Y_i) > \mathbb{E}(Y_i)$ ) that may exist in such models, the quasi-Poisson and negative binomial models are commonly used extensions.

In general when jointly modelling  $n$  data points at area units  $\{\mathcal{A}_1, \dots, \mathcal{A}_n\}$ , the log-relative risks are modelled using a generalised linear (mixed) model of the form

$$\begin{aligned} Y_i &\sim f(\mu_i = e_i\theta_i), \\ \ln(\theta_i) &= \mathbf{x}_i^\top \boldsymbol{\beta} + \phi_i, \end{aligned}$$

where  $f(\cdot)$  is an exponential family distribution that can be used to model count data such as a Poisson or negative binomial distribution,  $\mathbf{x}_i$  denotes a vector of area-level covariates,  $\boldsymbol{\beta}$  is a vector of regression coefficients, and  $\boldsymbol{\phi} = (\phi_1, \dots, \phi_n)^\top$  is a vector of random effects that captures residual spatial variability. If important unmeasured risk factors vary smoothly in space, then  $\boldsymbol{\phi}$  is typically assumed to be spatially correlated, for example via a conditional autoregressive prior. These spatial models are discussed in detail in Section 2.6.5.

### 2.6.5 Spatial models for areal unit data

In this thesis, one of the main responses of interest is areal unit data in an environmental epidemiology context. As in Section 2.6.4, let  $e_i$  denote the expected count in area  $\mathcal{A}_i$ , obtained by indirect standardisation with respect to age-sex strata. The quantity of primary interest is the latent relative risk  $\theta_i$ , which measures how much

the disease risk in  $\mathcal{A}_i$  deviates from the reference (national) risk level. The data model can be written in a general form as

$$Y_i|\theta_i, \boldsymbol{\xi} \sim f(y_i|\mu_i = e_i\theta_i, \boldsymbol{\xi}), \quad i = 1, \dots, n,$$

$$\ln(\theta_i) = \mathbf{x}_i^\top \boldsymbol{\beta} + \phi_i,$$

where  $\mu_i = \mathbb{E}(Y_i) = e_i\theta_i$ ,  $f(\cdot)$  is a suitable likelihood for count data (e.g., a Poisson or negative binomial distribution), and  $\boldsymbol{\xi}$  collects additional likelihood-specific hyperparameters (such as a dispersion or size parameter). On the log scale the relative risks are modelled using a generalised linear (mixed) form, where  $\mathbf{x}_i$  is a vector of observed covariates with regression coefficients  $\boldsymbol{\beta}$ , and  $\phi_i$  is a spatial random effect that captures residual spatial structure not explained by the covariates. The latent field  $\boldsymbol{\phi} = (\phi_1, \dots, \phi_n)^\top$  is modelled as a Gaussian random field with spatial dependence governed by the neighbourhood matrix  $\mathbf{W}$  introduced in Section 2.6.1.

In Bayesian disease mapping, spatial dependence among the random effects  $\boldsymbol{\phi}$  can be modelled using conditional autoregressive (CAR) priors, which specify the joint distribution of  $\boldsymbol{\phi}$  through a set of full conditional distributions  $f(\phi_i|\boldsymbol{\phi}_{-i})$ , where  $\boldsymbol{\phi}_{-i}$  denotes all random effects except  $\phi_i$ . Below I will briefly review several commonly used CAR priors.

### Intrinsic CAR model

The intrinsic CAR model proposed by (Besag et al., 1991) is one of the simplest and most widely used spatial priors for areal data. The joint prior is specified as

$$\boldsymbol{\phi} \sim N(\mathbf{0}, [\tau \mathbf{Q}(\mathbf{W})]^-), \quad \mathbf{Q}(\mathbf{W}) = \text{diag}(\mathbf{W}\mathbf{1}) - \mathbf{W}, \quad (2.8)$$

where  $\tau > 0$  is a precision (inverse variance) parameter,  $\mathbf{1}$  is an  $n$ -vector of ones, and  $(\cdot)^-$  denotes a generalised (Moore-Penrose) inverse because the intrinsic CAR precision matrix  $\mathbf{Q}(\mathbf{W})$  is singular. Equivalently, the ICAR prior can be written via its full conditional distributions as

$$\phi_i|\boldsymbol{\phi}_{-i} \sim N\left(\frac{\sum_{j=1}^n w_{ij}\phi_j}{\sum_{j=1}^n w_{ij}}, \frac{\tau^{-1}}{\sum_{j=1}^n w_{ij}}\right).$$

Thus, the conditional mean of  $\phi_i$  is the average of the random effects in neighbouring areas, while the conditional variance is inversely proportional to the number of neighbours if a binary neighbourhood matrix applies. This expresses the intuitive idea

that each area's random effect is smoothed towards the average of its neighbours, and that more neighbours provide more information and hence a smaller conditional variance. Because  $\mathbf{Q}(\mathbf{W})$  is singular, the joint distribution is improper and an identifiability constraint such as  $\sum_{i=1}^n \phi_i = 0$  is required to anchor the overall level of the random effects. Note that the model cannot handle areas with no neighbours (i.e.  $\sum_{j=1}^n w_{ij} = 0$ ), since in that case the conditional mean and variance for  $\phi_i$  are not defined.

### **BYM CAR model**

[Besag et al. \(1991\)](#) extended the ICAR prior by adding an unstructured random effect to capture non-spatial extra-variability. The resulting Besag-York-Mollié (BYM) model specifies

$$\ln(\theta_i) = \mathbf{x}_i^\top \boldsymbol{\beta} + \phi_i + \psi_i,$$

where  $\boldsymbol{\phi}$  follows an ICAR prior as in Equation (2.8), and the unstructured effects  $\boldsymbol{\psi} = (\psi_1, \dots, \psi_n)^\top$  are independent and identically distributed as  $\psi_i \sim N(0, \tau_\psi^{-1})$ , with  $\tau_\psi > 0$  an additional precision parameter.

The structured component  $\boldsymbol{\phi}$  captures smoothly varying spatial patterns, while the unstructured component  $\boldsymbol{\psi}$  accounts for local heterogeneity that is not spatially correlated. The relative magnitudes of  $\tau$  and  $\tau_\psi$  control the balance between spatially structured and unstructured variability, in which large values of  $\tau_\psi/\tau$  correspond to relatively strong spatial smoothing, whereas small values place more weight on unstructured variation. A limitation of the original BYM formulation is that only the sum  $\phi_i + \psi_i$  can be identified, so that the decomposition into structured and unstructured components is not uniquely determined, making interpretation of the individual component difficult.

### **Proper CAR model**

The intrinsic CAR prior induces strong spatial smoothing and yields an improper joint distribution for  $\boldsymbol{\phi}$ . Proper CAR models modify this construction to obtain a proper (non-singular) Gaussian prior by introducing an explicit spatial autocorrelation parameter  $\rho$  that allows the overall strength of spatial dependence to be estimated from the data. A common proper CAR specification proposed by [Stern and Cressie](#)

(2000) assumes a joint Gaussian prior of the form

$$\begin{aligned}\boldsymbol{\phi} &\sim N(\mathbf{0}, [\tau \mathbf{Q}(\rho, \mathbf{W})]^{-1}), \\ \mathbf{Q}(\rho, \mathbf{W}) &= \text{diag}(\mathbf{W}\mathbf{1}) - \rho \mathbf{W},\end{aligned}\tag{2.9}$$

with  $0 \leq \rho \leq 1$ . Here  $\rho$  controls the global level of spatial smoothing, where values close to 1 induce strong positive spatial correlation, whereas  $\rho = 0$  corresponds to spatial independence. The corresponding full conditional distribution is

$$\phi_i | \boldsymbol{\phi}_{-i} \sim N\left(\frac{\rho \sum_{j=1}^n w_{ij} \phi_j}{\sum_{j=1}^n w_{ij}}, \frac{\tau^{-1}}{\sum_{j=1}^n w_{ij}}\right),$$

which reduces to the ICAR prior when  $\rho = 1$ . A drawback of this specification is that when  $\rho = 0$  the random effects are no longer spatially correlated, but the conditional variance still depends on the number of neighbours. This weakens the interpretation of  $\rho$  as a pure spatial correlation parameter and motivates alternative parameterisations, such as the Leroux CAR model in Section 2.6.5, that give more interpretable boundary cases.

### Leroux CAR model

The Leroux model (Leroux et al., 2000) provides a flexible and widely used proper CAR prior that smoothly interpolates between spatial independence and the ICAR model. The joint prior for  $\boldsymbol{\phi}$  is specified as

$$\begin{aligned}\boldsymbol{\phi} &\sim N\left(\mathbf{0}, [\tau \mathbf{Q}(\rho, \mathbf{W})]^{-1}\right), \\ \mathbf{Q}(\rho, \mathbf{W}) &= \rho(\text{diag}(\mathbf{W}\mathbf{1}) - \mathbf{W}) + (1 - \rho)\mathbf{I},\end{aligned}\tag{2.10}$$

where  $0 \leq \rho \leq 1$ ,  $\mathbf{1}$  is an  $n$ -vector of ones and  $\mathbf{I}$  is the  $n \times n$  identity matrix. The corresponding full conditional distribution for  $\phi_i$  is

$$\phi_i | \boldsymbol{\phi}_{-i} \sim N\left(\frac{\rho \sum_{j=1}^n w_{ij} \phi_j}{\rho \sum_{j=1}^n w_{ij} + 1 - \rho}, \frac{\tau^{-1}}{\rho \sum_{j=1}^n w_{ij} + 1 - \rho}\right).$$

When  $\rho = 0$ , the model reduces to an exchangeable Gaussian prior with no spatial structure. As  $\rho \rightarrow 1$ , the prior behaves increasingly like the ICAR prior with strong spatial smoothing. The Leroux CAR prior is proper provided each area has at least one neighbour (i.e. each row of  $\mathbf{W}$  contains at least one positive entry).

### 2.6.6 Spatio-temporal models for areal unit data

The spatial models in Section 2.6.5 assume that data are observed at a single time point across  $n$  areal units. In many applications, data are collected repeatedly over time, and it is necessary to account for both spatial and temporal dependence. This section reviews several commonly used spatio-temporal extensions of spatial disease mapping models. Suppose for areas  $\mathcal{A} = \{\mathcal{A}_1, \dots, \mathcal{A}_n\}$  and time period  $t = 1, \dots, T$ , there is a data likelihood model as

$$Y_{it} | \theta_{it}, \boldsymbol{\xi} \sim f(y_{it} | \mu_{it} = e_{it} \theta_{it}, \boldsymbol{\xi}),$$

where  $\ln(\theta_{it}) = \eta_{it}$  is a linear predictor including covariate effects and spatio-temporal random effects. The models below differ in how they decompose  $\eta_{it}$  into spatial, temporal, and interaction components, and in how they specify the dependence structure of the random effects.

#### Bernardinelli model

One of the earliest spatio-temporal smoothing models was proposed by [Bernardinelli et al. \(1995\)](#), which allows area-specific linear temporal trends, with the linear slope itself varying across space. The linear predictor is specified as

$$\ln(\theta_{it}) = \mathbf{x}_{it}^\top \boldsymbol{\beta} + \phi_i + (\omega + \delta_i)t, \quad (2.11)$$

where  $\phi_i$  is a spatial random effect capturing baseline spatial variation (area-specific intercepts),  $\omega$  is a global linear time trend common to all areas, and  $\delta_i$  is an area-specific deviation from the global slope, so that  $\omega + \delta_i$  is the linear time trend for area  $\mathcal{A}_i$ . Typically,  $\boldsymbol{\phi} = (\phi_1, \dots, \phi_n)^\top$  and  $\boldsymbol{\delta} = (\delta_1, \dots, \delta_n)^\top$  are each assigned either independent Gaussian priors or spatially structured CAR priors. The latter allows both the baseline risk and the area-specific time trends to be spatially smooth, reflecting the idea that neighbouring areas may share similar levels and trends in disease risk. For identifiability,  $\sum_{i=1}^n \phi_i = 0$  and  $\sum_{i=1}^n \delta_i = 0$  are commonly imposed. However, the assumption that both the global trend and the area-specific deviations are restricted to be linear in  $t$  may be unrealistic for long time series or when the temporal pattern exhibits marked nonlinearity.

### Knorr-Held model

Knorr-Held (2000) proposed a more flexible spatio-temporal model by extending the BYM framework to include both spatial and temporal main effects, along with a space-time interaction term, which is given by

$$\ln(\theta_{it}) = \mathbf{x}_{it}^\top \boldsymbol{\beta} + \phi_i + \psi_i + \alpha_t + \gamma_t + \delta_{it}, \quad (2.12)$$

where  $\phi_i$  and  $\psi_i$  are structured and unstructured spatial random effects (as in the BYM model),  $\alpha_t$  and  $\gamma_t$  are structured and unstructured temporal random effects, and  $\delta_{it}$  is a spatio-temporal interaction term capturing additional variation that cannot be explained by the additive spatial and temporal components alone. Typically,  $\boldsymbol{\phi}$  follows a CAR prior to model spatial correlation,  $\boldsymbol{\alpha}$  follows a structured temporal prior such as first-order random walk (e.g.  $\alpha_1 \sim N(0, \tau_\alpha^{-1})$  and  $\alpha_t | \alpha_{t-1} \sim N(\alpha_{t-1}, \tau_\alpha^{-1})$  for  $t \geq 2$ ), and  $\boldsymbol{\psi}$  and  $\boldsymbol{\gamma}$  follow independent and identical normal distributions to capture spatial heterogeneity and unstructured temporal noise respectively.

Four types of prior structures for the interaction term  $\delta_{it}$  corresponding to different combinations of spatial and temporal structure are proposed and given by

**Type I** (unstructured  $\times$  unstructured):  $\delta_{it} \sim N(0, \tau_\delta^{-1})$ ; no spatial or temporal structure in the interaction.

**Type II** (structured space  $\times$  unstructured time): for each time  $t$ ,  $\boldsymbol{\delta}_{\cdot t} = (\delta_{1t}, \dots, \delta_{nt})^\top$  independently follows a CAR prior, where interactions are spatially smooth.

**Type III** (unstructured space  $\times$  structured time): for each area  $\mathcal{A}_i$ ,  $\boldsymbol{\delta}_{i\cdot} = (\delta_{i1}, \dots, \delta_{iT})^\top$  independently follows a temporal random walk.

**Type IV** (structured space  $\times$  structured time):  $\boldsymbol{\delta} = (\delta_{11}, \dots, \delta_{nT})^\top$  is assigned a multivariate Gaussian prior with precision matrix  $\mathbf{Q}_\delta = \mathbf{Q}_{\text{time}} \otimes \mathbf{Q}_{\text{space}}$ , where  $\mathbf{Q}_{\text{time}}$  and  $\mathbf{Q}_{\text{space}}$  are the structure matrices of the chosen temporal and spatial priors, and  $\otimes$  denotes the Kronecker product. This yields interaction effects that are correlated in both space and time.

This framework is very flexible and can represent a wide range of spatio-temporal dependence structures by appropriate choices of main effects and interaction type. However, the model is also highly parameterised, as spatial patterns are captured by  $\boldsymbol{\phi}$ ,  $\boldsymbol{\psi}$  and part of  $\boldsymbol{\delta}$ , while temporal trends are captured by  $\boldsymbol{\alpha}$ ,  $\boldsymbol{\gamma}$  and  $\boldsymbol{\delta}$ . This may lead to identifiability issues and make it difficult to interpret which component is responsible for which aspect of the spatio-temporal pattern.

### Ugarte model

Ugarte et al. (2012) proposed a simpler spatio-temporal structure in which the spatial and temporal main effects follow structured priors and the interaction term is defined using a separable space-time structure. The linear predictor is

$$\ln(\theta_{it}) = \mathbf{x}_{it}^\top \boldsymbol{\beta} + \phi_i + \alpha_t + \delta_{it}, \quad (2.13)$$

where  $\phi_i$  is a spatial random effect, typically following a proper CAR or Leroux prior;  $\alpha_t$  is a temporal random effect, typically following a first-order random walk or AR(1) process; and  $\delta_{it}$  is a space-time interaction term similar to Type IV form as defined in Section 2.6.6. Compared to the Knorr-Held model, the Ugarte model uses fewer random effects. There are no unstructured spatial or temporal components, and the form of the interaction is fixed to be structured in both space and time. This reduces the number of parameters and can make estimation more stable.

### Rushworth model

Rushworth et al. (2014) proposed a spatio-temporal model with a non-separable dependence structure that directly couples spatial and temporal correlation through a multivariate autoregressive formulation. They consider a single spatio-temporal random-effect surface  $\{\phi_{it}\}$ , and write

$$\begin{aligned} \ln(\theta_{it}) &= \mathbf{x}_{it}^\top \boldsymbol{\beta} + \phi_{it}, \quad t = 1, \dots, T, \\ \boldsymbol{\phi}_1 &\sim N(\mathbf{0}, [\tau \mathbf{Q}(\rho, \mathbf{W})]^{-1}), \\ \boldsymbol{\phi}_t | \boldsymbol{\phi}_{t-1} &\sim N(\alpha \boldsymbol{\phi}_{t-1}, [\tau \mathbf{Q}(\rho, \mathbf{W})]^{-1}), \quad t = 2, \dots, T, \end{aligned} \quad (2.14)$$

where  $\boldsymbol{\phi}_t = (\phi_{1t}, \dots, \phi_{nt})^\top$  denotes the vector of random effects for time period  $t$  and the precision matrix  $\mathbf{Q}(\rho, \mathbf{W})$  is that of the Leroux CAR prior (see Section 2.6.5). Here,  $\tau > 0$  is a precision parameter controlling the overall variability of the random effects,  $\rho \in [0, 1]$  controls the strength of spatial autocorrelation, and  $\alpha \in (-1, 1)$  controls the temporal autocorrelation. Here  $\alpha = 0$  indicates temporal independence, while  $\alpha = 1$  indicates strong temporal autocorrelation and turns this conditional part of the decomposition into a first-order random walk. Such specification is attractive because it has a relatively parsimonious parameterisation, with only three hyperparameters  $(\tau, \rho, \alpha)$  controlling the strength of spatial and temporal smoothing, and yields a non-separable spatio-temporal structure that often provides a realistic representation of how disease risk evolves over both space and time.



## Chapter 3

# A comparison of statistical and machine learning models for spatio-temporal prediction of air pollutant concentrations in Scotland

### 3.1 Introduction

Global concern about air pollution has intensified in recent decades, driven primarily by its detrimental impacts on human health and ecological systems (Brauer et al., 2016; Cohen et al., 2017). Air pollution typically comprises a mixture of particles and gaseous chemicals in the ambient atmosphere, including sulphur dioxide (SO<sub>2</sub>), nitrogen oxides (NO<sub>x</sub>), carbon monoxide (CO), ozone (O<sub>3</sub>) and particulate matter (PM) (Wu and Lin, 2019). Exposure to ambient air pollution has been linked to a wide range of adverse health outcomes, including chronic respiratory conditions, cardiovascular disease, mental health impacts, and premature mortality (Cohen et al., 2017). The World Health Organisation (WHO) has estimated that almost the entire global population (99%) is exposed to air pollution levels exceeding WHO guideline limits (World Health Organization, 2021). This burden is disproportionately borne by low- and middle-income countries, where populations are more likely to experience persistently high pollution levels (<https://www.who.int/health-topics/air-pollution>). Beyond public health, air pollution can also cause substantial ecological damage and exacerbate global climate change (Manisalidis et al., 2020); for example, it can drive soil and water acidification, harm forests and crops, and

threaten biodiversity.

In the United Kingdom, including Scotland, air pollution management is supported by a robust legislative framework, including key provisions such as the *Clean Air Act 1993* (UK Parliament, 1993) and the *Environment Act 2021* (UK Parliament, 2021). These regulations establish air quality standards, define enforcement mechanisms, and aim to reduce emissions to protect environmental quality and public health. In parallel, advisory authorities such as the Committee on the Medical Effects of Air Pollutants (COMEAP, <https://www.gov.uk/government/groups/committee-on-the-medical-effects-of-air-pollutants-comeap>) play an important role in quantifying the health burden attributable to air pollution and translating scientific evidence into policy-relevant guidance. Together, legal frameworks and scientific assessment underpin the UK's efforts to mitigate air pollution and evaluate the impacts of interventions.

Robust mapping of air pollutant concentrations is essential for effective air quality management and for safeguarding public health. Ideally, pollution maps would be derived from dense monitoring networks that provide point-level observations across space. In practice, however, monitoring data are often spatially sparse and temporally incomplete, reflecting the impermanence of sensors and the asynchronous installation and retirement of monitoring sites. To address limited monitoring coverage, numerical air quality models can provide spatially complete predictions, such as the national Pollution Climate Mapping (PCM) model (<https://uk-air.defra.gov.uk/data/pcm-data>) produced under the auspices of the Department for Environment, Food & Rural Affairs (DEFRA). These models simulate pollutant transport and atmospheric chemical reactions using emissions inventories, meteorology, and computational representations of dispersion and transformation processes (Byun and Schere, 2006).

Nevertheless, numerical model outputs and emissions based inputs may be affected by systematic biases and may be available only at relatively coarse temporal or spatial resolutions. In Scotland, the widely used modelled product, the PCM, is provided only as annual averages. Although some products provide finer temporal resolution, e.g. the European-scale CAMS reanalysis, these are produced on coarser spatial grids than the 1 km<sup>2</sup> PCM surface. This limits the study of seasonal variation in exposure, and restricts investigation of interactions between air pollution and other risk factors that vary at sub-annual scales. A natural strategy to mitigate these limitations is to integrate numerical model outputs with monitoring data, with the goal of producing more accurate and higher-resolution concentration surfaces. The real-world motivation for this chapter is therefore to develop monthly air pollution

maps for Scotland by effectively calibrating and temporally “downscaling” the numerical predictions using observations from the monitoring network. Calibration is needed because numerical model outputs are constructed from emissions inventories and dispersion modelling, and so carry systematic biases relative to the measured concentrations that the monitoring data can correct. In addition, quantifying predictive uncertainty is equally important. The resulting exposure surface feeds the epidemiological analyses in Chapter 4, where ignoring prediction error would understate exposure measurement error and bias the estimated health effects, and the uncertainty maps additionally indicate where the network is least informative and new monitoring would be most valuable.

A range of approaches has been used for spatio-temporal prediction of air pollution, including land use regression, generalised additive models, hierarchical Bayesian models, and machine learning methods. In Scotland, exposure information for population level studies has typically been drawn directly from the sparse monitoring network or from the annual PCM model outputs, and flexible spatio-temporal prediction methods have rarely been systematically applied or compared in this setting. However, there remains no clear consensus as to which class of methods is most appropriate, either in general or specifically in nations characterised by sparse and inhomogeneous monitoring networks such as Scotland, where monitoring sites are concentrated in central regions (Chen et al., 2019; Berrocal et al., 2020; Ren et al., 2020). The primary aim of this chapter is therefore to compare different statistical and machine learning methods for spatio-temporal air pollution prediction in the context of a sparse and unevenly distributed monitoring network. Using numerical model outputs, meteorological variables and land cover covariates as predictors, I predict air pollutant concentrations at unmonitored locations across Scotland. Specifically, I investigate monthly average pollutant concentrations and associated predictors from January 2016 to December 2020.

I evaluate an extensive set of candidate models for predicting concentrations of three key pollutants, namely nitrogen dioxide ( $\text{NO}_2$ ), coarse particulate matter ( $\text{PM}_{10}$ ) and fine particulate matter ( $\text{PM}_{2.5}$ ), and select a final model for each pollutant based on out-of-sample predictive performance. The selected models are then used to generate monthly concentration maps for Scotland at a  $1 \text{ km}^2$  spatial resolution. These predictions enable identification of areas experiencing long-term improvements or deteriorations in air quality and support subsequent analyses of health impacts (the focus of Chapter 4), with the potential to inform future public health policy.

In presenting these final air pollution prediction maps, I focus on the following

important features of air pollution in Scotland:

- (i) The overall spatial pattern in concentrations over the entire 5-year period, to identify where concentrations are highest.
- (ii) The temporal trend in concentrations over the 60-month study period, to assess long-term and seasonal patterns in air pollution and identify the locations where concentrations have decreased most.
- (iii) The level of uncertainty in the predictions as measured by the average widths of the 95% prediction intervals, to potentially inform the placement of new monitoring sites for the greatest improvement in prediction precision.

This chapter is organised as follows. Section 3.2 reviews existing approaches to spatio-temporal air pollution prediction, focusing on linear regression, generalised additive models, hierarchical spatio-temporal smoothing models, and machine learning methods. Section 3.3 introduces the study region and datasets, summarising the monitoring observations and the covariates used for prediction. In Section 3.4, I present the modelling framework, including key assumptions, mathematical specifications, and estimation procedures for each candidate model. Section 3.5 describes the model validation strategy and performance metrics, including hyperparameter tuning for the machine learning methods. In Section 3.6, I compare predictive performance across models and pollutants, aiming to identify a best performing model for each pollutant. Section 3.7 then presents the resulting monthly prediction maps, highlighting spatial and temporal patterns as well as prediction uncertainty quantification. Finally, in Section 3.8, I discuss limitations of the work, and outline how these results motivate and support the analyses in the subsequent chapter.

## 3.2 Literature review of modelling techniques in air pollution prediction

This section reviews modelling approaches that are commonly used for spatio-temporal prediction of ambient air pollutant concentrations. In practice, prediction is typically informed by multiple data sources (e.g., monitoring observations, numerical model outputs, meteorology and land use covariates), and methods must balance predictive accuracy, interpretability and the ability to quantify uncertainty. The review is organised around four broad classes of approaches: linear regression

models (including land use regression), generalised additive models, hierarchical spatio-temporal models, and machine learning methods.

### 3.2.1 Linear models

Linear regression models have a long history in air pollution prediction and forecasting, and remain widely used because they are interpretable, computationally efficient, and straightforward to implement and validate. In this setting, pollutant concentrations are related to predictors such as modelled concentrations, meteorological variables, spatial and temporal indicators, land use data, and traffic-related covariates. Land use regression (LUR) models, in particular, represent a prominent and practically influential application of the linear modelling framework in environmental and ecological studies, aiming to explain fine-scale spatial variation in pollution using geographically referenced predictors.

A substantial literature demonstrates the utility of LUR for predicting air pollutant concentrations and supporting exposure assessment in epidemiological studies. For example, [Gilbert et al. \(2005\)](#) developed an LUR model linking log NO<sub>2</sub> concentrations to land use variables, reporting a coefficient of determination ( $R^2$ ) of 0.54 and highlighting the role of traffic-related predictors. [Ross et al. \(2006\)](#) fitted multiple regression models to NO<sub>2</sub> at 39 monitoring sites in San Diego, USA, using predictors such as road conditions, traffic flow, land use, population and housing, and reported that four of the covariates explained 79% of the variation in NO<sub>2</sub>. National-scale applications have also been developed. [Novotny et al. \(2011\)](#) produced LUR-based estimates of ground-level NO<sub>2</sub> across the contiguous USA for 2006 ( $R^2 = 0.78$ ), demonstrating the ability of the model to capture within-urban and near-road variation in NO<sub>2</sub> concentrations, while [Larkin et al. \(2017\)](#) developed global LUR models using data from over 5,200 monitoring sites across 58 countries, explaining 54% of variation in NO<sub>2</sub> with a mean absolute error of 3.7 ppb. Linear regression has also been used to calibrate particulate matter estimates derived from satellite-based products and chemical transport model outputs against ground observations ([GBD 2013 Risk Factors Collaborators, 2015](#); [Brauer et al., 2016](#)), enabling improved global exposure assessment and long-term trend analysis.

A key limitation of standard linear regression in this context is that the relationship between measurements and predictors may vary spatially due to regional differences in emissions, meteorology, topography and monitoring-site characteristics. Methods that allow coefficients to vary across space have therefore been developed. Geograph-

ically weighted regression (GWR) estimates location-specific regression parameters and is designed to capture spatial heterogeneity in covariate effects (Brunsdon et al., 1998). For example, Van Donkelaar et al. (2016) used GWR to allow calibration relationships between measurements and modelled estimates to vary spatially, incorporating additional land use and compositional information. Related approaches include linear mixed models and hierarchical linear models, which can represent spatially varying relationships more flexibly through random effects and structured correlation assumptions (Liu et al., 2009; Wang et al., 2016; Sahu, 2022). Despite their wide use, linear models typically impose a restrictive functional form, which can limit their ability to represent non-linear relationships and interactions that are common in atmospheric processes. These limitations motivate more flexible regression frameworks, such as additive models.

### 3.2.2 Generalised additive models

Generalised additive models (GAMs) extend linear regression by allowing non-linear relationships between the response and predictors through smooth functions (Hastie and Tibshirani, 1986, 1990). This flexibility is attractive in air pollution applications because relationships with meteorology, seasonality and land use often exhibit non-linearity and thresholding. GAMs have been used to predict  $\text{NO}_2$  and PM using predictors such as traffic indicators, satellite products and meteorological variables, often including smooth spatial terms based on monitoring-site coordinates to represent broad-scale spatial structure (Liu et al., 2009; Hart et al., 2009).

Several studies further improve GAM-based prediction by explicitly modelling residual spatial dependence. For example, Li et al. (2012) combined GAMs with cokriging of the residuals to estimate  $\text{NO}_2$  and  $\text{NO}_x$  surfaces for winter and summer in Southern California, reporting strong cross-validation performance ( $R^2$  between 0.88 and 0.92). In that approach, GAMs model the mean structure driven by covariates and large-scale trends, while cokriging captures residual spatial correlation under a second-order stationarity assumption. GAMs have also been used for interpretive analyses of pollutant–meteorology relationships. In a study of Beijing, Hou and Xu (2022) used GAM-based analyses to quantify marginal effects after controlling for long-term trends and seasonality, identifying key meteorological predictors and providing a better understanding into how changes in meteorology may influence air quality. Overall, GAMs offer a useful compromise between interpretability and flexibility. However, their performance can be sensitive to the choice of smoothness

penalties and may still be limited when spatial structure is highly localised, non-smooth, or driven by unobserved factors, motivating hierarchical spatio-temporal models that treat spatial and temporal dependence explicitly.

### 3.2.3 Hierarchical spatio-temporal models

Hierarchical spatio-temporal models provide a robust framework for analysing point-referenced air pollution data while explicitly accounting for both spatial and temporal dependency structures (Wikle et al., 1998; Wikle, 2003; Banerjee et al., 2014; Cressie, 2015). Typically, these models decompose variation into a structured latent process (capturing spatial and/or spatio-temporal dependence) and an unstructured component capturing measurement error or micro-scale variability (often referred to as the *nugget effect*) (Sahu, 2022). A common formulation uses spatio-temporal Gaussian processes (GPs), where dependence is encoded through a spatial or spatio-temporal covariance function. Banerjee et al. (2014) provide many general suggestions for modelling such spatio-temporal covariance structure. Because fully non-separable spatio-temporal covariance models can be computationally demanding, separable correlation structures in space and time are often adopted to reduce complexity (Mardia and Goodall, 1993). Temporal dependence is frequently incorporated through autoregressive (AR) structures, which can capture persistent month-to-month behaviour while retaining interpretable parameters and tractable computation (Cocchi et al., 2007; Cameletti et al., 2011; Sahu, 2012).

A major practical barrier for GP-based models is computation, as dense covariance matrices lead to expensive matrix operations, particularly under Markov chain Monte Carlo (MCMC) inference. To address this, sparse approximations and Markov representations have become central to modern spatial modelling. The stochastic partial differential equation (SPDE) approach of Cameletti et al. (2013), building on the link between Matérn Gaussian fields and Gaussian Markov random fields (GMRFs), enables representation of spatial processes through sparse precision matrices (Rue and Held, 2005; Lindgren et al., 2011). This sparsity supports scalable inference using efficient numerical linear algebra and facilitates approximate Bayesian inference via the integrated nested Laplace approximation (INLA; Rue et al., 2009). Hierarchical spatio-temporal models are therefore attractive because they provide coherent uncertainty quantification and explicitly account for dependence. However, they typically require stronger modelling assumptions and greater computational effort than regression-based approaches, and their predictive performance can be sensi-



tive to misspecification of covariance structure or to sparse and uneven monitoring networks.

### 3.2.4 Machine learning

Machine learning methods have become increasingly popular in air pollution modelling due to their flexibility, scalability, and ability to capture complex non-linear relationships and high-order interactions (Reid et al., 2015; Hu et al., 2017b; Liu et al., 2017). They also exhibit excellent computational efficiency and suitability for handling large datasets, which is common in environmental and epidemiological studies. Commonly used methods include random forests (Hu et al., 2017b; Meng et al., 2018; Zhan et al., 2018), gradient boosting algorithms (Reid et al., 2015; Zhan et al., 2017), and support vector regression (Hu et al., 2017a). Comparative studies report mixed conclusions that depend on pollutant, spatial domain, and validation design. For example, Reid et al. (2015) compared multiple approaches for daily PM<sub>2.5</sub> prediction in California and found machine learning methods outperformed linear regression, whereas Van den Bossche et al. (2018) reported little improvement from support vector regression relative to linear models for black carbon in a mobile monitoring campaign. In Europe, Chen et al. (2019) found broadly similar predictive performance between a set of machine learning algorithms and linear models for PM<sub>2.5</sub> and NO<sub>2</sub>. Reviews have also highlighted settings in which explicitly spatial statistical methods perform strongly, e.g., Berrocal et al. (2020) reported that downscaler models and universal kriging outperformed several machine learning algorithms for PM<sub>2.5</sub> exposure assessment in the contiguous United States, plausibly because spatial dependence is modelled directly. In contrast, Ren et al. (2020) found that machine learning approaches provided practical and competitive estimates in a large-scale US comparison to LUR models.

A growing literature seeks to combine the predictive strength of machine learning with explicit modelling of spatial dependence. For example, Liu et al. (2018) proposed random forest regression kriging (RFRK), where a random forest models the mean structure and kriging is applied to residuals to account for remaining spatial correlation (Hengl et al., 2015). Other developments explicitly target spatial heterogeneity; for instance, geographical random forests train local models using nearby observations to capture spatially varying covariate effects (Georganos et al., 2021; Luo et al., 2021). Despite their promise, machine learning methods can be sensitive to the choice of tuning parameters, validation scheme, and covariate availability,



and they often provide limited inferential interpretation and less direct uncertainty quantification than hierarchical Bayesian approaches. Consequently, careful model selection, spatially appropriate validation, and uncertainty assessment remain essential for reliable exposure prediction (Meyer et al., 2018).

## 3.3 Data and exploratory analysis

### 3.3.1 Study region and data collection

This study focuses on Scotland, a nation in the north of the United Kingdom. Scotland has a diverse physical environment, including mountainous regions, rolling hills, coastal plains and numerous lochs (McKernan, 2004). Land cover is similarly varied, spanning extensive peatlands, woodland, farmland and urbanised areas (Alshammari et al., 2018). Scotland's meteorology is highly variable in both space and time, which strongly influences pollutant dispersion, chemical transformation and deposition, and therefore plays an important role in determining ambient air quality.

The objective of this chapter is to predict monthly average concentrations (in  $\mu\text{g}/\text{m}^3$ ) of nitrogen dioxide ( $\text{NO}_2$ ), particulate matter with aerodynamic diameter  $\leq 10 \mu\text{m}$  ( $\text{PM}_{10}$ ) and  $\leq 2.5 \mu\text{m}$  ( $\text{PM}_{2.5}$ ) on a  $1 \text{ km}^2$  grid across Scotland, over the five-year period from January 2016 to December 2020. The choice of a monthly temporal resolution is motivated by two considerations. First, the predicted concentrations are intended for linkage to monthly aggregated health outcomes in the subsequent epidemiological analyses, where the focus is on longer-term associations rather than short-term forecasting. Second, daily modelling is challenging in this setting due to substantial missingness in the monitoring data, both across sites and across time.

The choice of a  $1 \text{ km}^2$  spatial resolution is driven by the following factors. The intended downstream analyses are conducted at small-area population units whose spatial scales are typically larger than  $1 \text{ km}^2$ , so finer spatial resolution is not essential for the exposure estimates. In addition, many key predictors used in this chapter (including numerical model outputs and gridded meteorological variables) are naturally available at, or can be spatially realigned to, a  $1 \text{ km}^2$  grid. Finally, moving to a finer spatial resolution would substantially increase the number of prediction locations and hence the computational burden.

### 3.3.2 Monitoring pollution concentrations

Daily mean monitoring concentrations (in  $\mu\text{g}/\text{m}^3$ ) for  $\text{NO}_2$ ,  $\text{PM}_{10}$  and  $\text{PM}_{2.5}$  were obtained from the Scottish Air Quality website (<https://www.scottishairquality.scot/>). These measurements span January 1, 2016 to December 31, 2020 (1,827 days) and were recorded at the sites shown in Figure 3.1. These sites are part of the Automatic Urban and Rural Network (AURN, <https://uk-air.defra.gov.uk/networks/network-info?view=aurn>), which is established and maintained by the Department for Environment, Food & Rural Affairs (DEFRA) to monitor ambient air quality in the UK.

Note that not all monitoring sites have the ability to measure all pollutants. Over the study period,  $\text{NO}_2$  was measured at 94 sites,  $\text{PM}_{10}$  at 93 sites, and  $\text{PM}_{2.5}$  at 79 sites. These monitoring site locations are displayed in Figure 3.1, and are colour-coded by the type of local environment in which they reside, which comprises kerbside (7 monitors, 6.6% of sites), roadside (78, 73.6%), rural (5, 4.7%), suburban (3, 2.8%), urban background (10, 9.4%) and urban industrial (3, 2.8%). The definitions of these types are given below, and are taken from <https://www.scottishairquality.scot/air-quality/monitoring>.

**Kerbside:** A site sampling within 1 m of the kerbside of a busy road.

**Roadside:** A site sampling between 1 m of the kerbside of a busy road and the back of the pavement. Typically this will be within 5 m of the road, but could be up to 15 m.

**Rural:** An open countryside location, in an area of low population density distanced as far as possible from roads, populated and industrial areas.

**Suburban:** A location type situated in a residential area on the outskirts of a town or city.

**Urban background:** An urban location distanced from sources and therefore broadly representative of city-wide background conditions, e.g. urban residential areas.

**Urban industrial:** An area where industrial sources make an important contribution to the total pollution burden, located approximately 20–30 m from the kerb of a busy road.

The majority of the sites are next to main roads (Kerbside and Roadside), which will naturally have higher concentrations than nearby background (Rural, Suburban,

and Urban) locations, indicating potential monitoring bias towards traffic-influenced locations (Beckwith et al., 2019). Moreover, the sites are very unevenly spread across the country, with the vast majority in the central belt containing Glasgow in the west and Edinburgh in the east, while Northern and Southern Scotland contain hardly any monitoring sites. Because monitor siting in Scotland is driven by regulatory priorities, with sites concentrated near busy roads and in the population dense central belt, the locations of the network are not independent of the pollution surface being measured. This is a form of preferential sampling (Diggle et al., 2010), which can bias estimates of the underlying concentration field and tends to make out-of-sample predictive performance most optimistic where the network is densest.

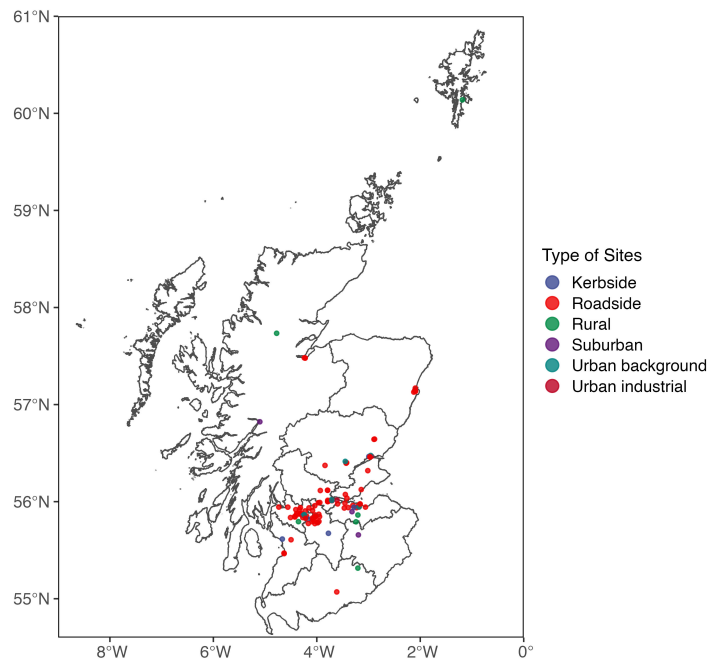


Figure 3.1: Map showing the air pollution monitoring sites in Scotland, which are colour-coded according to the type of local environment in which they reside.

A summary of the daily mean concentrations measured by the monitoring sites is presented in the left panel of Table 3.1 for the three pollutants. The table shows that monitoring site data for all three pollutants include some negative concentrations, which contradicts the physical constraint that air pollutant concentrations should be non-negative. This discrepancy suggests potential recording errors in the data or instrument malfunction, and as this only affects 53 (0.0274%) measurements they are removed and considered as missing (NA) in the modelling.

The amount of missingness in the daily mean concentrations is complex to summarise because: (i) some monitors only measured a subset of the three pollutants;

Table 3.1: Summary statistics of the monitoring site (measured) data and the modelled concentration data for  $\text{NO}_2$ ,  $\text{PM}_{10}$  and  $\text{PM}_{2.5}$ . The measured data (left) are daily averages, while the modelled data (right) are yearly averages for those  $1 \text{ km}^2$  grid squares that contain at least one monitoring site.

|         | Monitoring concentrations |                  |                   | Modelled concentrations |                  |                   |
|---------|---------------------------|------------------|-------------------|-------------------------|------------------|-------------------|
|         | $\text{NO}_2$             | $\text{PM}_{10}$ | $\text{PM}_{2.5}$ | $\text{NO}_2$           | $\text{PM}_{10}$ | $\text{PM}_{2.5}$ |
| Min.    | -0.10                     | -1.50            | -2.00             | 1.79                    | 4.72             | 2.29              |
| 1st Qu. | 11.80                     | 7.20             | 3.30              | 9.28                    | 8.66             | 5.13              |
| Median  | 20.80                     | 10.10            | 4.70              | 12.15                   | 9.74             | 5.99              |
| Mean    | 24.34                     | 11.90            | 5.97              | 12.97                   | 9.56             | 5.75              |
| 3rd Qu. | 33.30                     | 14.50            | 7.00              | 15.21                   | 10.48            | 6.42              |
| Max.    | 167.90                    | 359.20           | 58.20             | 31.71                   | 12.62            | 7.67              |
| NA's    | 57355                     | 69680            | 117007            |                         |                  |                   |

and (ii) not all monitoring sites were present over the entire five year study duration. These two factors lead to a non-rectangular data array over pollutants, monitoring sites and calendar days. Additionally, some of the monitors had periods when they were non-operational, which induced further missing values into the data set. A summary of these patterns of “missingness” is presented in Figures 3.2 to 3.4, where grey denotes the observation is not present (missing) while black denotes it is present. Examining the distribution of grey cells can provide an indicative narrative for missing records at a given site. For example, long consecutive grey blocks may indicate that a site was not able to measure the pollutant or was not operational during that period, while intermittent grey blocks could suggest sensor malfunction or invalid data storage. For  $\text{NO}_2$  and  $\text{PM}_{10}$ , missing records appear commonly due to both reasons, whereas for  $\text{PM}_{2.5}$ , the missing records exist mainly because most monitoring sites did not measure this pollutant during the entire five-year study period.

These daily averages are then aggregated to a monthly temporal resolution by averaging (mean), as long as there is one or more observations recorded within the month. Additionally, I record the numbers of days that measurements are available within each month, which will be used to allow for heteroscedasticity in the additive models for location, scale, and shape (see Section 3.4.2). Figure 3.5 shows box plots of the temporal trends in these monthly mean concentrations for all three pollutants, with lines denoting the median values across the set of available monitoring sites. A clear seasonality can be observed in  $\text{NO}_2$  concentrations, which peak in winter and decline in summer. This trend could be attributed to increased vehicular usage and heating requirements during the colder months, leading to higher

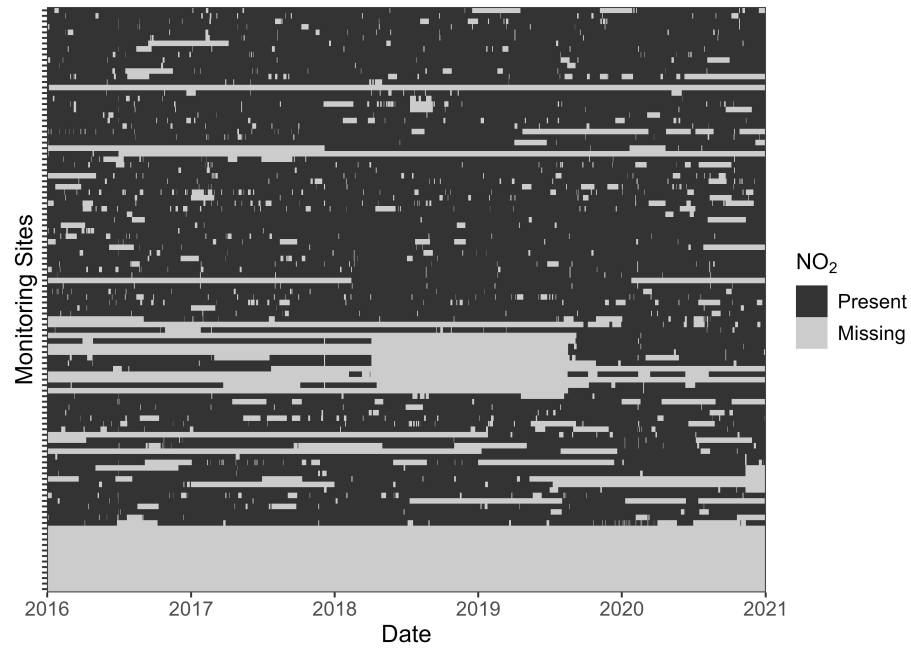


Figure 3.2: Heat map summarising the presence or absence of NO<sub>2</sub> concentration data by day and site.

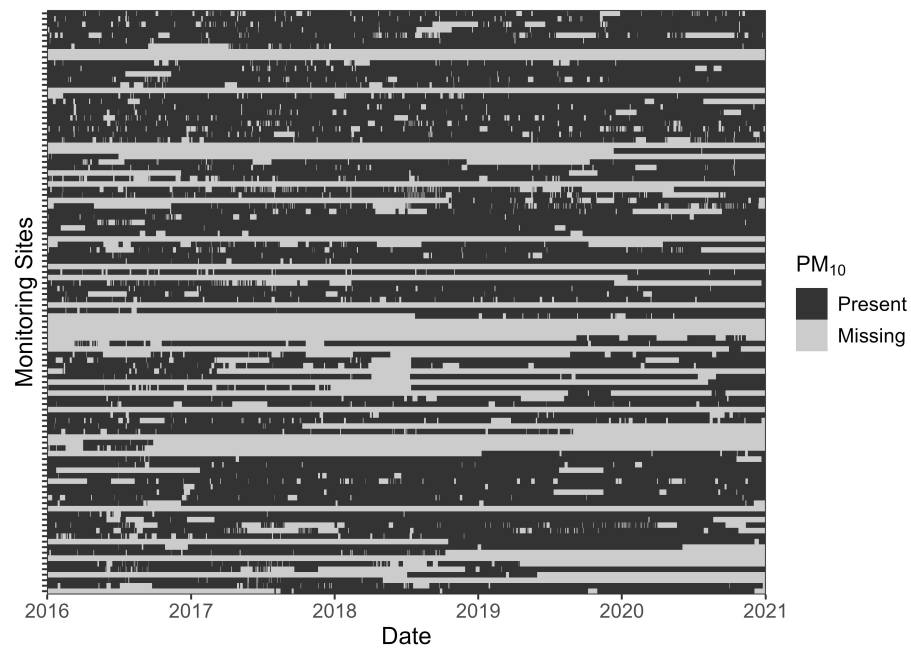


Figure 3.3: Heat map summarising the presence or absence of PM<sub>10</sub> concentration data by day and site.

nitrogen oxide emissions. Conversely, in summer, NO<sub>2</sub> is more easily converted to O<sub>3</sub> due to enhanced photochemical reactions under stronger sunlight and higher

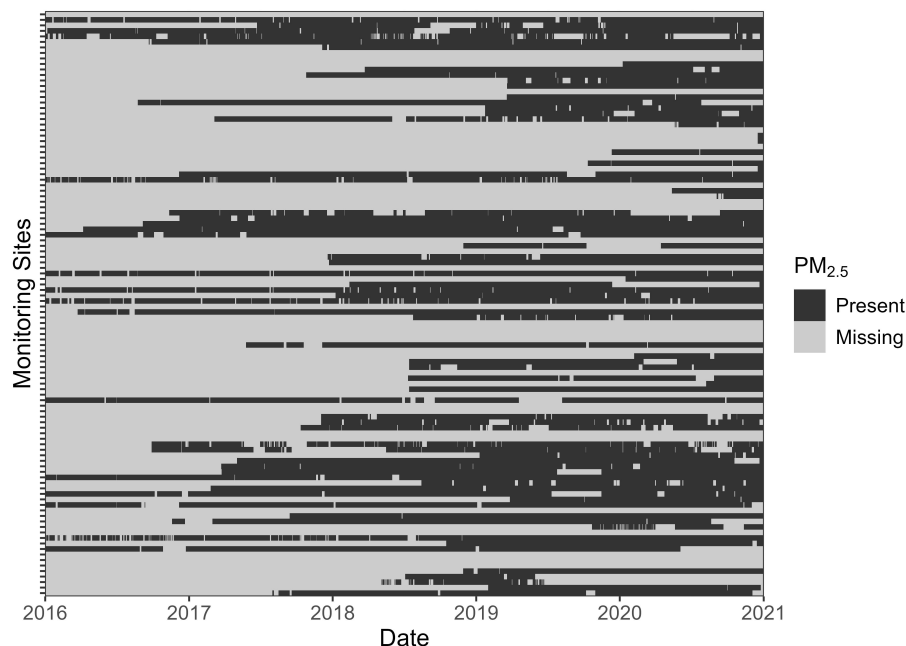


Figure 3.4: Heat map summarising the presence or absence of PM<sub>2.5</sub> concentration data by day and site.

temperatures, thereby reducing atmospheric NO<sub>2</sub> concentrations (Air Quality Expert Group, 2004). In addition, PM<sub>10</sub> and PM<sub>2.5</sub> exhibit weaker and less regular seasonality than NO<sub>2</sub>, though some seasonal variation is present. Two apparent spikes in particulate matter levels are also observed in February and April 2019, which can be traced back to two concerning secondary pollution events from continental Europe (Department for Environment, Food and Rural Affairs, 2020). The specific meteorological conditions during these periods facilitated the transport of foreign particulate matter to the UK. Subsequently, a phase of warm, sunny weather with low wind speeds impeded the dispersion of these particles, leading to their accumulation in the atmosphere (Department for Environment, Food and Rural Affairs, 2020).

Figure 3.6 provides a year-by-year view of seasonality in monthly mean concentrations. NO<sub>2</sub> concentrations were broadly stable between 2016 and 2019, followed by an evident reduction in 2020. In particular, a pronounced decline is visible from March 2020 onwards, coinciding with the first nationwide lockdown in response to the Covid-19 pandemic. For PM<sub>10</sub> and PM<sub>2.5</sub>, concentrations in early 2020 were generally lower than in previous years, consistent with a combination of reduced activity and meteorological conditions that favoured dispersion and wet deposition (DEFRA, <https://www.gov.uk/government/statistics/air-quality-statistics/concentrations-of-particulate-matter-pm10-and-pm25>).

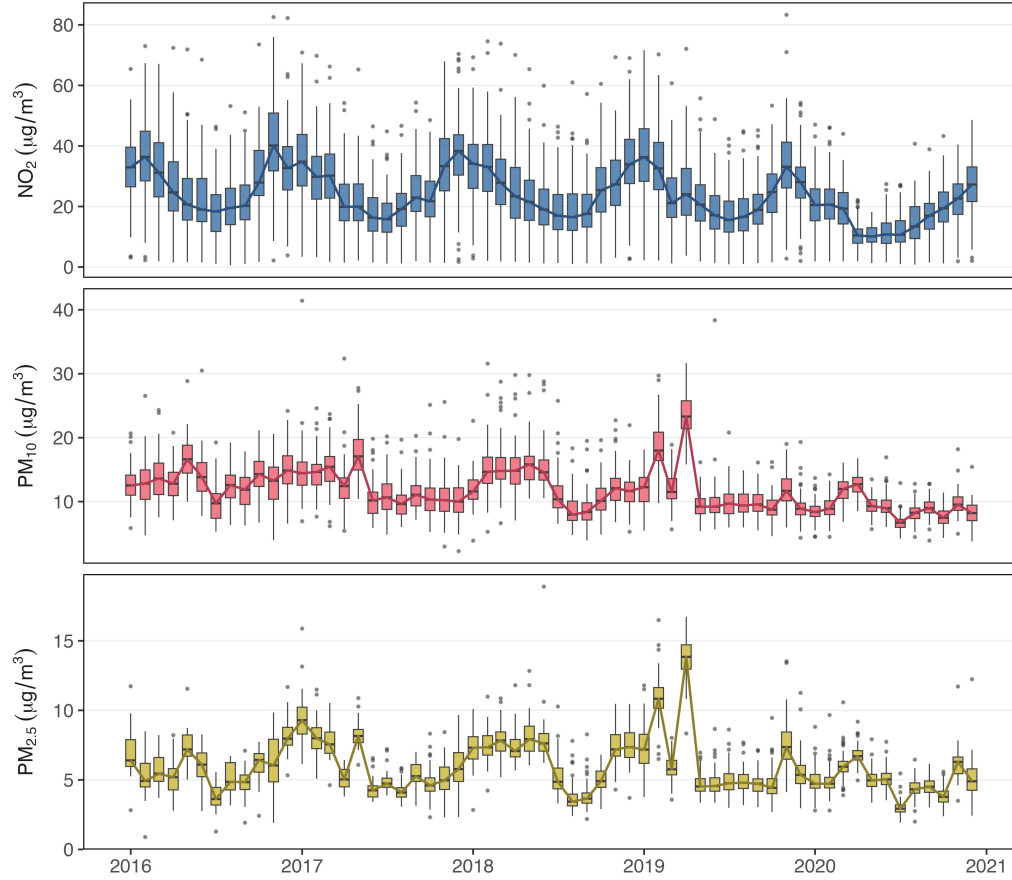


Figure 3.5: Monthly temporal trends in the monitored concentrations of  $\text{NO}_2$ ,  $\text{PM}_{10}$  and  $\text{PM}_{2.5}$  via box plots and medians (across sites).

### 3.3.3 Modelled pollution concentrations

As the monitoring sites are spatially sparse and unevenly distributed, I also utilise annual average modelled concentrations on a  $1 \text{ km}^2$  grid from the Pollution Climate Mapping (PCM; <https://uk-air.defra.gov.uk/data/pcm-data>) model under DEFRA's Modelling of Ambient Air Quality (MAAQ) contract. These PCM concentrations will be used as an annual predictor in the regression models for the monthly monitoring site concentrations, i.e., all 12 months of the same calendar year will have the same value. Thus, the statistical/machine learning models presented in this chapter essentially downscale these PCM data to a monthly resolution. These numerical modelled concentrations are generated by integrating data sources, such as road traffic data, meteorological data and air dispersion model outputs (Pugsley et al., 2022). In the PCM gridded dataset, cells without values are treated as missing, typically corresponding to coastal grid squares where the cell centroid falls offshore. To obtain a spatially complete predictor surface, missing PCM values are imputed

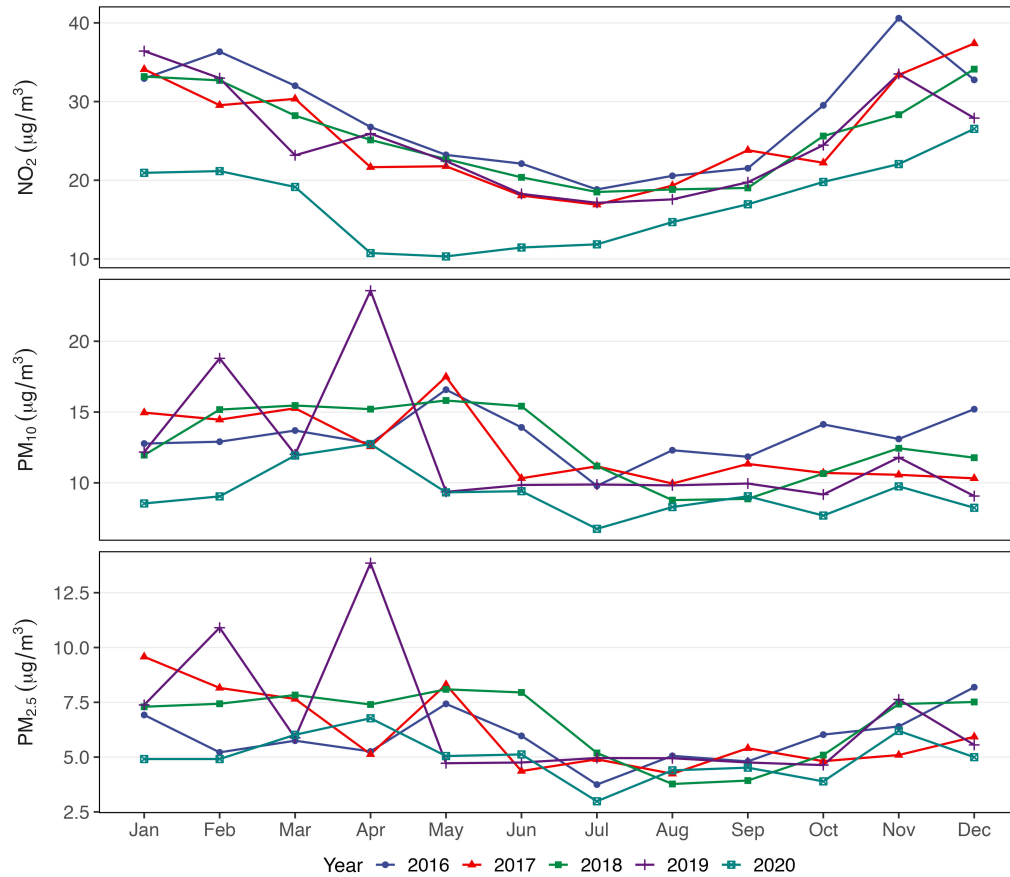


Figure 3.6: Line charts showing the seasonality of monthly mean concentrations of  $\text{NO}_2$ ,  $\text{PM}_{10}$  and  $\text{PM}_{2.5}$  by year.

using the value from the nearest non-missing grid cell.

The right panel of Table 3.1 presents summary statistics for the annual mean concentrations obtained from the PCM model for  $1 \text{ km}^2$  grid cells that contain at least one monitoring site. These modelled concentrations exhibit narrower ranges and lower peak values compared to the measurements, which is likely due to the following two factors:

- **Spatio-temporal resolution:** The modelled data represent annual mean concentrations over a  $1 \text{ km}^2$  grid cell, whereas the monitoring data are daily averages at a single point location. Thus the spatio-temporal aggregation of the modelled data smooths out local variations in air pollution levels, leading to narrower ranges and lower peak concentrations.
- **Model assumption:** The PCM model outputs incorporate some assumptions and simplifications about the processes that drive air pollution concentrations,



which may not capture all the real-world complexities, particularly at small spatial scales. For example, the model might not fully account for local pollution sources, such as road traffic, or for local conditions relating to geography or the built environment that affects pollutant dispersion.

Figure 3.7 summarises the annual PCM concentrations for those 1 km<sup>2</sup> grid cells that contain at least one monitoring site. For NO<sub>2</sub>, concentrations are highest in 2016, followed by a modest decline and relative stability between 2017 and 2019. A further reduction is evident in 2020, with the median falling below 10 µg/m<sup>3</sup>, consistent with reduced traffic volumes during the Covid-19 lockdown period. Modelled PM<sub>10</sub> exhibits a broadly similar pattern, with elevated levels in 2016, a small decrease in 2017, little change thereafter until 2019, and a clearer decline in 2020. In contrast, PM<sub>2.5</sub> peaks in 2017 and then decreases gradually, followed by a sharper drop in 2020, when the median reaches approximately 4.5 µg/m<sup>3</sup>.

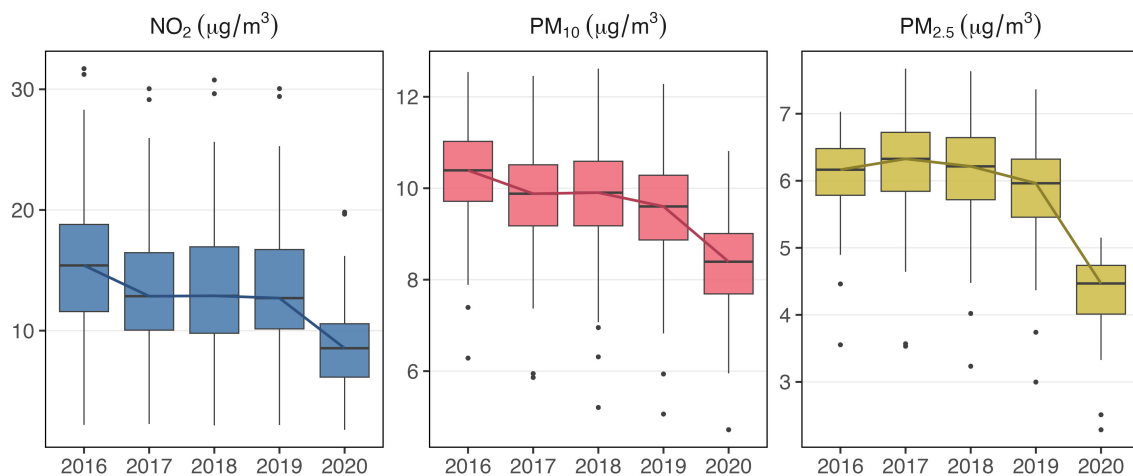


Figure 3.7: Box plots of annual PCM modelled concentrations (NO<sub>2</sub>, PM<sub>10</sub> and PM<sub>2.5</sub>) for grid cells that contain at least one monitoring site (2016–2020).

Figure 3.8 presents the five-year mean concentration maps for the three pollutants between 2016 and 2020. For NO<sub>2</sub> a discernible urban-rural difference is visible, with clearly elevated concentrations in the four largest cities of Aberdeen, Dundee, Edinburgh and Glasgow, while the northern Highlands and southern Scotland demonstrate lower concentrations with means below 10 µg/m<sup>3</sup>. This pattern is primarily due to vehicular emissions and heating combustion, which are more prevalent in urban areas due to their increased population density. In contrast, the spatial distribution of PM<sub>10</sub> concentrations reveals elevated levels in both urban areas and eastern regions, with the former again being driven by dense populations while the latter is partially

driven by the influence of long-range trans-boundary pollution from continental Europe (Department for Environment, Food and Rural Affairs, 2023). Finally,  $PM_{2.5}$  concentrations exhibit a similar pattern to  $PM_{10}$ , which is unsurprising given that the former is a subset of the latter. Glasgow and Edinburgh, being the two largest cities in Scotland, have experienced the highest  $PM_{2.5}$  pollution levels.

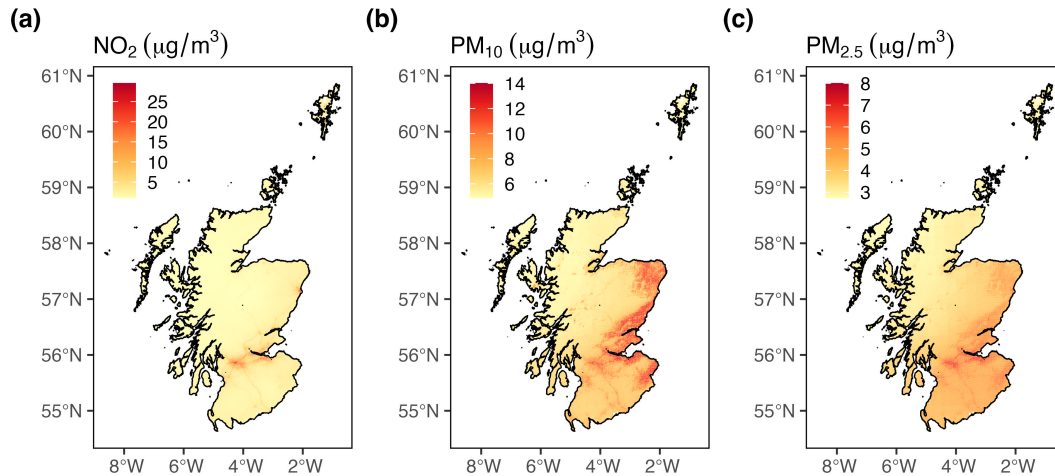


Figure 3.8: Spatial distributions of the temporal mean (between 2016 and 2020) of the 1 km<sup>2</sup> estimates of NO<sub>2</sub>, PM<sub>10</sub> and PM<sub>2.5</sub> from the Pollution Climate (PCM) Mapping model.

### 3.3.4 Meteorological predictors

Meteorological conditions play an important role in governing the formation, dispersion, transport, and deposition of ambient air pollutants, and are therefore routinely included as predictors in air pollution models (Guo and Feng, 2019; Liu et al., 2022). In this study, meteorological variables were obtained from the HadUK-Grid products produced by the UK Met Office (<https://www.metoffice.gov.uk>). HadUK-Grid provides gridded climate summaries derived from the UK land surface observation network by interpolating station observations to a regular 1 km<sup>2</sup> grid (Met Office et al., 2022). In this study I use six meteorological predictors, summarised in Table 3.2, with each variable aggregated to the monthly scale to match the exposure data.

Figure 3.9 presents box plots of monthly meteorological conditions at the monitoring locations. Relative humidity shows a pronounced seasonal cycle, with higher values in late autumn and winter (median close to 95%) and lower values in late spring and summer. In contrast, mean sea level pressure and 10 m wind speed exhibit

Table 3.2: Meteorological predictors derived from HadUK-Grid and used in the modelling.

| Label    | Name                       | Unit | Aggregation |
|----------|----------------------------|------|-------------|
| Hurs     | Relative humidity          | %    | mean        |
| Psl      | Pressure at mean sea level | hPa  | mean        |
| SfcWind  | Wind speed at 10 m         | m/s  | mean        |
| Sun      | Sunshine hours             | h    | sum         |
| Rainfall | Total precipitation amount | mm   | sum         |
| Tas      | Air temperature            | °C   | mean        |

weaker seasonal patterns over the study period. The relatively narrow interquartile range for wind speed suggests limited spatial variability at the monitoring locations. Sunshine duration and air temperature display strong seasonality, as expected for Scotland, whereas total precipitation shows substantial month-to-month variability with a less distinct seasonal signal. Notably, in February 2020, Scotland experienced its heaviest rainfall in half a decade, followed by a sharp drop in rainfall, reaching its lowest point in April of that year.

Figure 3.10 maps the five-year mean of each meteorological predictor across Scotland. Relative humidity tends to be higher in southern Scotland and along northern coastal regions, with lower values across parts of the central belt. Mean sea level pressure exhibits a south–north gradient. Wind speed shows relatively weak spatial structure at this scale. Sunshine duration is generally higher in coastal and southern areas and lower across the north-western Highlands. Spatial variation in air temperature is modest, although cooler conditions are evident across some central and higher-elevation areas. Finally, precipitation displays a clear west–east contrast, with the western coast receiving substantially higher rainfall than the east.

### 3.3.5 NDVI and land use data

To characterise vegetation cover and seasonal greenness, this study uses the Normalised Difference Vegetation Index (NDVI) data derived from the Terra Moderate Resolution Imaging Spectroradiometer (MODIS; <https://lpdaac.usgs.gov/products/mod13a3v006/>). The MODIS NDVI product provides monthly gridded values at a 1 km<sup>2</sup> spatial resolution. MODIS is a widely used remote sensing technology that provides an Earth observation platform for land surface property measurements, including vegetation and land cover, with a 16-day revisit time. MODIS-based vegetation indices provide a consistent way to quantify changes in vegetation canopy

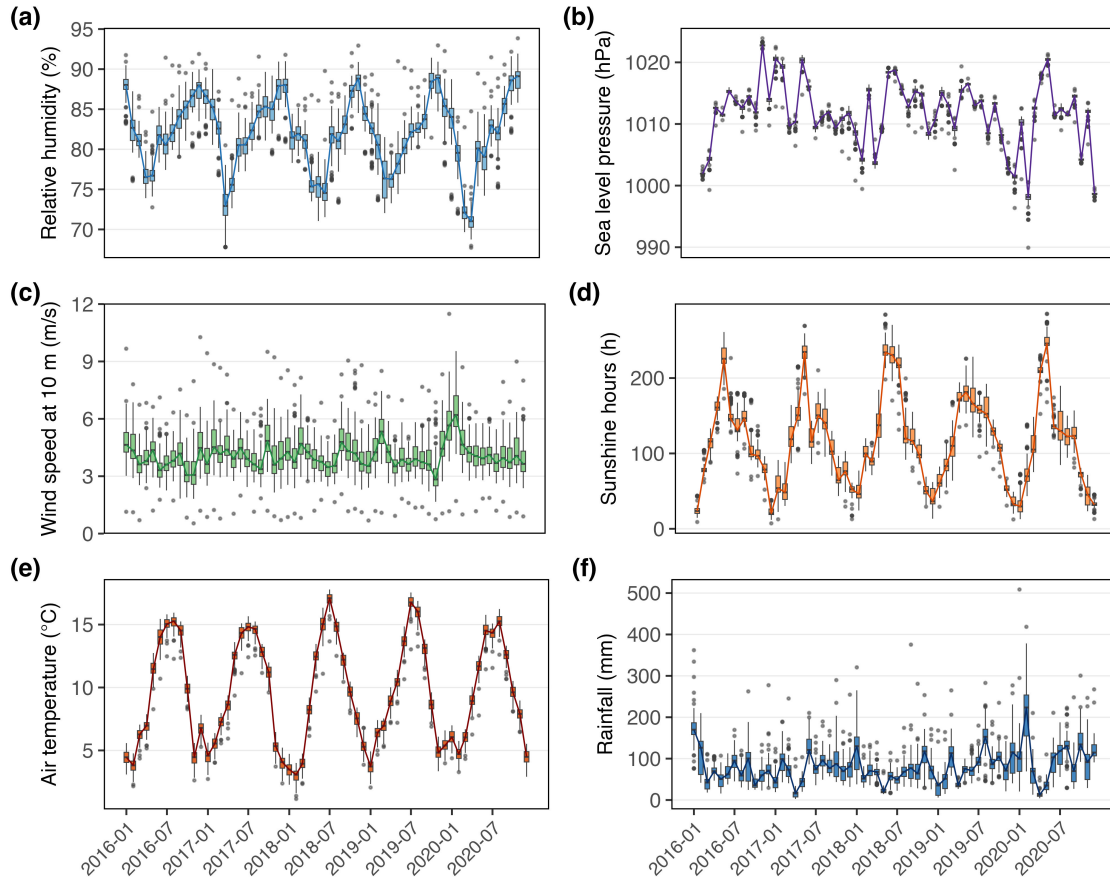


Figure 3.9: Box plots of monthly meteorological predictors (relative humidity, sea-level pressure, 10 m wind speed, sunshine hours, total precipitation and air temperature) at the monitoring-site locations over 2016–2020. Source: HadUK-Grid.

greenness, which is a combined attribute of leaf area, chlorophyll, and canopy structure. NDVI summarises the contrast between the visible and near-infrared bands of the electromagnetic spectrum, and provides a standardised proxy for vegetation canopy greenness. It takes values in  $[-1, 1]$ , with negative values typically corresponding to water, values close to zero indicating bare soil or sparse vegetation, and higher positive values indicating denser vegetation. NDVI is defined as

$$\text{NDVI} = \frac{\rho_{\text{NIR}} - \rho_{\text{Red}}}{\rho_{\text{NIR}} + \rho_{\text{Red}}},$$

where  $\rho_{\text{NIR}}$  and  $\rho_{\text{Red}}$  denote reflectance in the near-infrared and red spectral bands, respectively. The five-year mean NDVI across Scotland is mapped in Figure 3.11 (b), which shows higher greenness across the rural uplands and lower values in the urban central belt.

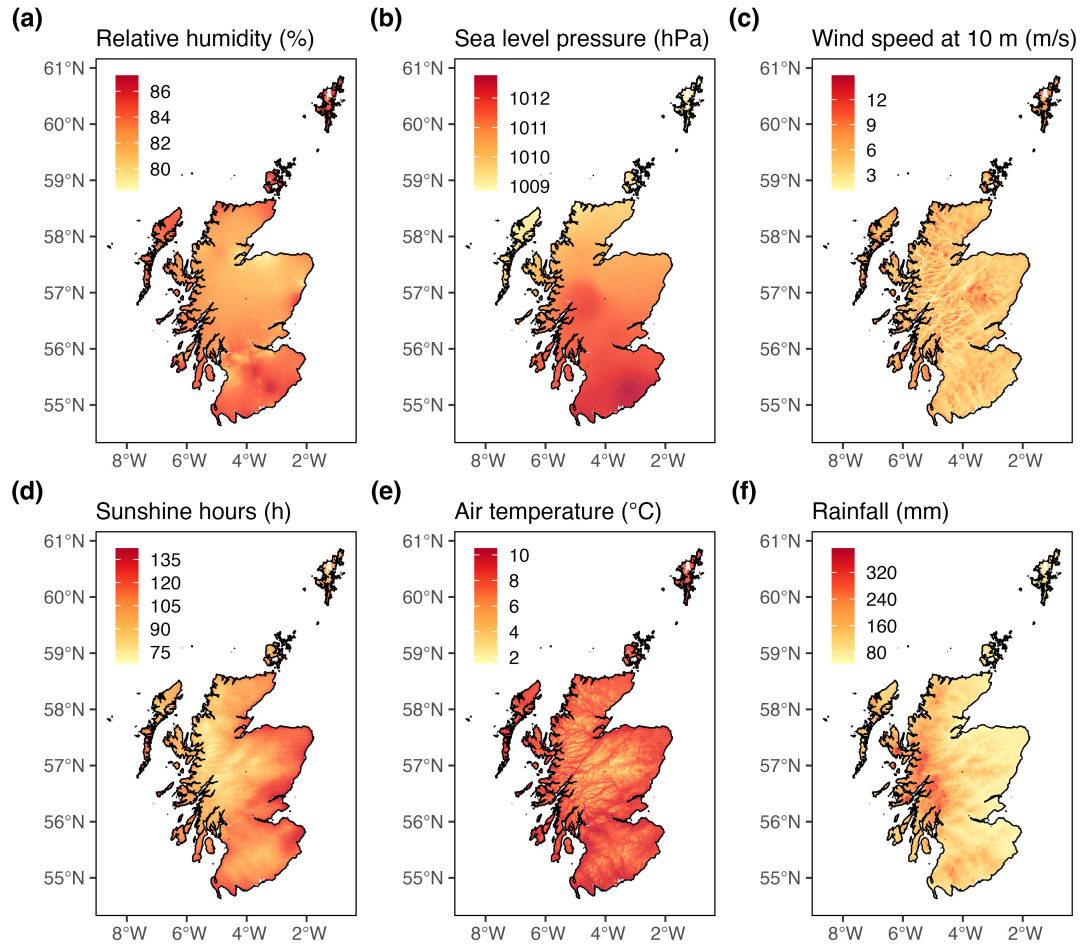


Figure 3.10: Spatial distribution of the five-year mean (2016–2020) of each monthly meteorological predictor (relative humidity, sea-level pressure, 10 m wind speed, sunshine hours, total precipitation and air temperature) across Scotland on the 1 km<sup>2</sup> grid. Source: HadUK-Grid.

This study also uses land-use information from the Scottish Government Urban Rural Classification (2020 version; [Rural and Environment Science and Analytical Services Division, 2022](#)). This classification provides a consistent framework for defining urban and rural areas across Scotland using settlement size and measures of accessibility. The classification is based on population data from the National Records of Scotland and accessibility analysis, which distinguishes between accessible and remote areas in Scotland. While the original system is available in both 6-fold and 8-fold forms, I aggregated categories into three groups to support stable estimation given that most monitoring sites are located in urban areas. The resulting three-class scheme is summarised in Table 3.3. Figure 3.11 (a) maps the classification and highlights that, outside the four major cities and their surrounding areas, most of

Scotland is classified as rural.

Table 3.3: Aggregated (three-class) Scottish Urban Rural Classification used in this study.

| Class | Class name        | Description                      |
|-------|-------------------|----------------------------------|
| 1     | Large urban areas | Populations of 125,000 or more.  |
| 2     | Other urban areas | Population of 10,000 to 124,999. |
| 3     | Rural areas       | Population less than 10,000.     |

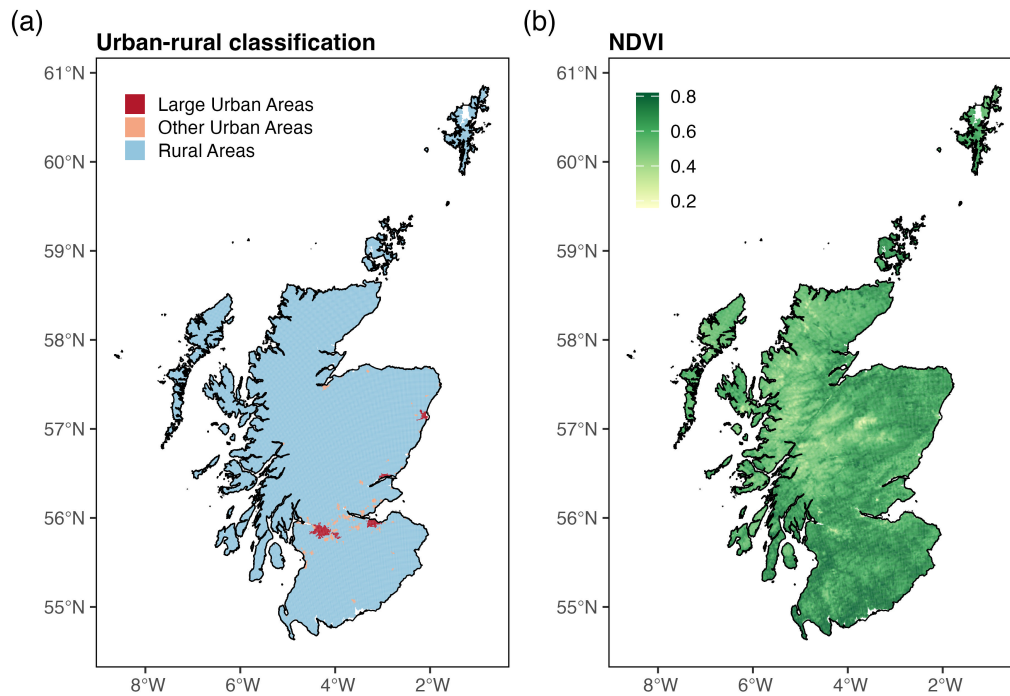


Figure 3.11: (a) The aggregated three-class Scottish Urban Rural Classification and (b) the five-year mean (2016–2020) NDVI, both mapped on the 1 km<sup>2</sup> grid.

### 3.3.6 Population size and traffic conditions

In addition to the meteorological and land cover predictors described above, both population density and proximity to the road network are well-established drivers of spatial variation in ambient air pollution concentrations and are therefore included as predictors in this study. Population density reflects the intensity of domestic heating, vehicular activity and other anthropogenic emission sources, and has been shown to explain a substantial proportion of urban–rural differences in pollutant



levels (Novotny et al., 2011; Larkin et al., 2017). Proximity to major roads is particularly relevant for traffic-related pollutants such as  $\text{NO}_2$ , for which vehicular exhaust constitutes a primary emission source, but also contributes meaningfully to particulate matter concentrations through both direct tailpipe emissions and road dust resuspension.

Population density data were obtained from the mid-year population estimates published by the National Records of Scotland. These estimates are provided at the Data Zone (DZ) level, a small-area administrative geography comprising 6,976 spatial units across Scotland. The DZ-level population counts were spatially realigned to the  $1 \text{ km}^2$  grid by area weighted averaging, yielding a gridded population density surface consistent with the resolution of the other predictors. The spatial distribution of population density across Scotland is presented in Figure 3.12 (a). As expected, the map reveals that the population density is higher in the central belt, with the two hotspots corresponding to Glasgow in the west and Edinburgh in the east. Beyond these two cities, the remainder of Scotland, including the northern and western Highlands, the island groups, and large parts of southern Scotland, exhibits population densities close to zero, reflecting the predominantly rural and sparsely settled character of these regions.

To characterise proximity to road infrastructure, the straight-line distance from the centroid of each monitoring site and each  $1 \text{ km}^2$  grid square to the nearest major road was computed using road network data obtained from OpenStreetMap (<https://download.geofabrik.de/europe/united-kingdom/scotland.html>). Major roads were defined as those classified as motorways, trunk roads, primary roads, secondary roads, and tertiary roads within the OpenStreetMap hierarchy. The resulting distance covariate (in metres) provides a continuous, time-invariant measure of road proximity for each location across Scotland. The resulting gridded distance-to-nearest-road surface is shown in Figure 3.12 (b). The underlying road network is most dense in the central belt, particularly in the Glasgow–Edinburgh corridor. Major arterial routes extend northward through the A9 corridor toward Inverness and Aberdeen, and southward toward the border with England, while the western Highlands are served by comparatively sparse road network. Consequently, grid squares in the remote north and west of Scotland are characterised by larger distances to the nearest major road than those in the urban central belt.

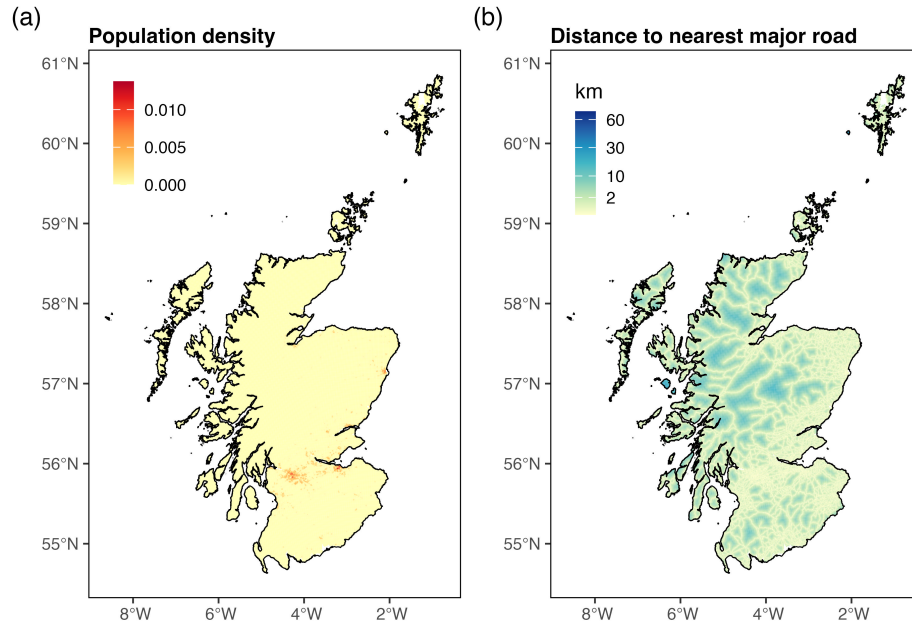


Figure 3.12: Spatial distribution of population density (normalised) across Scotland, derived from National Records of Scotland mid-year population estimates and harmonised to the  $1 \text{ km}^2$  prediction grid (left panel), together with the gridded distance from each  $1 \text{ km}^2$  grid cell to the nearest major road (in km), derived from OpenStreetMap road data (right panel).

### 3.4 Methods for predicting ambient air pollutant concentrations

The data described in Section 3.3 present both opportunities and challenges for spatio-temporal prediction. The monitoring network provides direct observations of pollutant concentrations but is spatially sparse and unevenly distributed, with coverage concentrated in the central belt while large parts of Scotland, particularly the northern Highlands, contain few or no monitoring sites. The PCM model outputs offer spatially complete annual concentration surfaces that can serve as a powerful predictor, but require temporal downscaling to the monthly resolution needed for downstream epidemiological analyses. The remaining predictors – meteorological variables, NDVI, urban–rural classification, population density, and distance to major roads – are available at the  $1 \text{ km}^2$  monthly grid resolution, providing a rich covariate set to explain both spatial and temporal variation in concentrations. Together, these data characteristics motivate a modelling framework that can calibrate PCM outputs against monitoring observations, exploit complex and potentially non-linear



relationships between air pollutant concentrations and predictors, and generate spatially complete monthly predictions with uncertainty quantification. To identify the approach best suited to this setting, this chapter compares a range of candidate models spanning four broad classes: normal linear models, additive models, hierarchical spatio-temporal models, and random forests. Each is described in turn below, and all are evaluated under a common spatial cross-validation framework detailed in Section 3.5.

Throughout this section, let  $Y(\mathbf{s}_i, t)$  denote the monthly mean concentration of a given pollutant at monitoring location  $\mathbf{s}_i$ ,  $i = 1, \dots, n$ , in month  $t = 1, \dots, T$ , where  $n$  is the number of monitoring sites for that pollutant and  $T = 60$ . Let  $\mathbf{x}(\mathbf{s}_i, t)$  be the corresponding predictor vector from Section 3.3, comprising the PCM modelled concentration, the meteorological variables, NDVI, urban–rural class, population density and distance to the nearest major road. Each pollutant is modelled separately, so pollutant subscripts are suppressed. For every model, a 95% prediction interval is produced for each held-out observation and grid cell, and back-transformed to the original concentration scale. For linear and additive models (LM, AM, AMLSS), intervals are obtained as Gaussian predictive intervals on the log scale using the predictive standard error. For hierarchical spatio-temporal models (GP, AR), intervals are the 2.5th and 97.5th percentiles of the posterior predictive samples. For random forests (RF), intervals are the 2.5th and 97.5th conditional quantiles from quantile regression forests.

### 3.4.1 Normal linear models

The simplest class of models considered here assumes that log-transformed monthly mean concentrations are linearly related to the full set of predictors described in Section 3.3. Formally, the normal linear model (LM) takes the form

$$Y(\mathbf{s}_i, t) = \alpha + \mathbf{x}(\mathbf{s}_i, t)^\top \boldsymbol{\beta} + \epsilon(\mathbf{s}_i, t), \quad \epsilon(\mathbf{s}_i, t) \stackrel{\text{iid}}{\sim} N(0, \sigma_\epsilon^2), \quad (3.1)$$

for monitoring locations  $\{\mathbf{s}_i\}_{i=1}^n$  and months  $t = 1, \dots, T$ , where  $\alpha$  is the intercept,  $\boldsymbol{\beta}$  is the coefficient vector, and  $\epsilon(\mathbf{s}_i, t)$  is an independent Gaussian error with constant variance  $\sigma_\epsilon^2$ . Further details of estimation and inference are given in Chapter 2.

Despite its simplicity, the linear model serves as an important baseline. The covariate set is rich, and in particular the PCM modelled concentrations already encode much of the spatial structure, so a well-specified linear model may perform competitively even against more complex approaches. Two variants are considered,

differing on only in how temporal evolution is parameterised.  $LM_{cs}$  assumes a *constant seasonal* pattern across all five years by including separate factor variables for *Year* and *Month*, implying that the monthly seasonal effect is invariant over time.  $LM_{vs}$  allows for *varying seasonality* by replacing these with a single variable with one level per calendar month across the 60-month study period, thereby permitting the seasonal pattern to differ from year to year. The inclusion of spatial coordinates (longitude and latitude) as additional linear terms was also assessed but did not improve predictive performance in either variant, and these terms were therefore excluded.

### 3.4.2 Additive models

Additive models (AMs) extend linear models by allowing continuous predictors to enter through smooth, data-driven functions rather than fixed linear terms (Hastie and Tibshirani, 1986), which is attractive in the air pollution prediction context where relationships with meteorological variables and temporal trends are often non-linear. The general form is

$$Y(\mathbf{s}_i, t) = \alpha + \mathbf{x}^{(1)}(\mathbf{s}_i, t)^\top \boldsymbol{\beta} + \sum_{j=p_1+1}^p f_j(x_j(\mathbf{s}_i, t)) + \epsilon(\mathbf{s}_i, t), \quad (3.2)$$

$$\epsilon(\mathbf{s}_i, t) \stackrel{\text{iid}}{\sim} N(0, \sigma_\epsilon^2),$$

where  $\mathbf{x}^{(1)}(\mathbf{s}_i, t)$  is a set of  $p_1$  covariates containing the categorical predictors entering linearly, and  $\{x_j(\mathbf{s}_i, t)\}$  are the continuous predictors whose effects  $\{f_j(\cdot)\}$  are estimated from the data via penalised regression splines. Specifically, univariate  $P$ -splines are used for continuous covariates, a Gaussian process smooth for the overall temporal trend, and a cyclic  $P$ -spline for seasonality to ensure continuity across year boundaries. This formulation directly extends  $LM_{vs}$  by allowing non-linear predictor effects and a flexible year-varying seasonal structure. Smoothing parameters are selected by the restricted maximum likelihood (REML) method, which is generally preferred over the generalised cross-validation (GCV) criterion as it tends to be more stable and less prone to under-smoothing (Wood, 2017). Further details of parameter estimation and inference are given in Chapter 2 and Wood (2017).

Two variants of the additive model are compared: (i) AM includes all predictors as described above without any explicit spatial trend, and (ii)  $AM_{sp}$  additionally incorporates longitude and latitude as a bivariate smooth term to capture any residual large-scale spatial structure not already explained by the covariates.

A potential limitation of the additive model in this setting is its assumption of homoscedastic errors. Because monthly average concentrations are aggregated from varying numbers of daily observations due to equipment malfunction, site downtime, and irregular presence of monitors across the study period (see Section 3.3), the within-month variance is likely to differ across sites and months. To address this, I also consider additive models for location, scale and shape (AMLSS; [Rigby and Stasinopoulos, 2005](#)), which extend the additive models by allowing the error variance depend on predictors via

$$\begin{aligned} \log(\sigma_\epsilon(\mathbf{s}_i, t) - b) &= \alpha_\sigma + \mathbf{x}_\sigma(\mathbf{s}_i, t)^\top \boldsymbol{\beta}_\sigma + \sum_{k=q_1+1}^q f_k(x_{\sigma_k}(\mathbf{s}_i, t)) + \nu(\mathbf{s}_i, t), \\ \nu(\mathbf{s}_i, t) &\stackrel{\text{iid}}{\sim} N(0, \sigma_\nu^2). \end{aligned} \quad (3.3)$$

Here the natural log of the error variance  $\log(\sigma_\epsilon(\mathbf{s}_i, t) - b)$  can depend linearly and/or non-linearly on predictors  $\{x_{\sigma_k}(\mathbf{s}_i, t)\}$ , and thus modelled are in an analogous way to the mean function in Equation (3.2). Specifically,  $x_{\sigma_k}(\mathbf{s}_i, t)$  is the number of missing daily records for site  $\mathbf{s}_i$  within month  $t$  and  $f_k(\cdot)$  is a  $P$ -spline smooth. In this specification  $\nu(\cdot)$  is an independent white-noise process, while  $b > 0$  pre-specifies a minimum value for  $\{\sigma(\mathbf{s}_i, t)\}$  to avoid singularities in the model likelihood ([Wood et al., 2016](#)). Two variants are again considered: AMLSS without a spatial trend, and AMLSS<sub>sp</sub> with a bivariate spatial smooth. Details of parameter estimation and prediction can again be found in [Wood \(2017\)](#).

### 3.4.3 Hierarchical spatio-temporal models

Unlike the regression-based approaches described above, which model pollutant concentrations purely through covariate effects, hierarchical spatio-temporal models additionally incorporate a latent process to represent structured dependence in space and time. This provides a principled framework for capturing short-range spatial autocorrelation and temporal evolution in pollutant concentrations that may remain after accounting for the predictors, while simultaneously delivering uncertainty quantification through the Bayesian framework. Following [Bakar and Sahu \(2015\)](#), the model assumes that the observed concentration  $Y(\mathbf{s}_i, t)$  is an error-prone measurement of the true underlying pollutant surface  $O(\mathbf{s}_i, t)$  as

$$Y(\mathbf{s}_i, t) \mid O(\mathbf{s}_i, t) \sim N\left(O(\mathbf{s}_i, t), \sigma_\epsilon^2\right), \quad (3.4)$$

where  $\sigma_\epsilon^2$  captures measurement error and micro-scale variability not explained by the latent process. Two specifications of  $O(\mathbf{s}_i, t)$  are considered here, chosen for the availability of efficient software implementation and their established use in the air pollution literature (Mukhopadhyay and Sahu, 2017).

The first is the spatio-temporal Gaussian process (GP) model, which assumes temporal independence in the latent surface while allowing for spatial autocorrelation as

$$O(\mathbf{s}_i, t) = \mathbf{x}(\mathbf{s}_i, t)^\top \boldsymbol{\beta} + w(\mathbf{s}_i, t), \quad (3.5)$$

where  $\mathbf{w}_t = (w(\mathbf{s}_1, t), \dots, w(\mathbf{s}_n, t))^\top \sim N(\mathbf{0}, \sigma_w^2 \mathbf{S}_w)$  independently across months. Here  $\sigma_w^2$  governs the magnitude of spatially structured variation and  $\mathbf{S}_w$  is a spatial correlation matrix specified via the exponential correlation function (Bakar and Sahu, 2015).

The second is the spatio-temporal autoregressive (AR) model, which extends the GP model by introducing temporal dependence through a first-order autoregressive structure as

$$O(\mathbf{s}_i, t) = \alpha O(\mathbf{s}_i, t-1) + \mathbf{x}(\mathbf{s}_i, t)^\top \boldsymbol{\beta} + w(\mathbf{s}_i, t), \quad (3.6)$$

where  $\alpha \in (-1, 1)$  is the temporal autocorrelation parameter. When  $\alpha = 0$  the AR model reduces to the GP model. The inclusion of  $\alpha$  allows the model to capture persistent month-to-month behaviour in pollution concentrations while retaining an interpretable and tractable parameterisation.

Both models are fitted within a Bayesian framework using Markov chain Monte Carlo (MCMC) simulation, implemented via the `spTimer` package in R (Bakar and Sahu, 2015). Four parallel chains are run, each discarding the first 2,000 iterations as burn-in before retaining a further 8,000 samples, yielding 32,000 posterior samples in total. Convergence is assessed using the Gelman–Rubin diagnostic (Gelman and Rubin, 1992). Point predictions are taken as the posterior median of the predictive samples, and 95% prediction intervals are obtained as the 2.5th and 97.5th percentiles. Further details of prior specifications and posterior computation are given in Chapter 2.

### 3.4.4 Random forests

Random forests (RF; Breiman, 2001) are one of the most widely used prediction algorithms, and have been applied in air pollution modelling contexts (Hu et al., 2017b; Liu et al., 2018; Zhan et al., 2018). The aim of this chapter is to contrast broad modelling paradigms rather than to benchmark every machine learning algorithm, so

a single well-understood ensemble method is preferred over an exhaustive comparison. The method builds on bootstrap aggregation, whereby multiple copies of the training data are created by sampling with replacement, and a decision tree is fitted to each bootstrapped sample through recursive binary partitioning of the predictor space. Each tree grows until either the number of observations in a terminal node falls below a minimum threshold `min.size`, or the depth of the tree exceeds a maximum `max.depth`. To reduce the correlation between individual trees, only a randomly selected subset of `mtry` predictors is considered as candidates at each split of each tree. The final prediction for a new observation is obtained by averaging the predictions from all  $n_{\text{tree}}$  trees, with the variance reduction achieved through this averaging improving predictive stability relative to any single tree.

Since random forests do not make distributional assumptions, they can in principle be applied directly to the original concentration scale. To ensure a fair comparison with the other models in this study, which are all fitted on the log scale, two variants are compared:  $\text{RF}_{\text{oc}}$ , fitted to the original concentrations, and  $\text{RF}_{\text{lc}}$ , fitted to the log-transformed concentrations with predictions subsequently back-transformed. In addition, because random forests are in essence non-spatial and make predictions based solely on covariate information, two further variants are considered that augment the predictor set with longitude and latitude as additional features, denoted  $\text{RF}_{\text{oc\_sp}}$  and  $\text{RF}_{\text{lc\_sp}}$  respectively. Including spatial coordinates in this way allows the random forest to capture non-linear spatial trends and spatial heterogeneity in covariate effects that would otherwise be ignored.

Unlike the statistical models described in earlier subsections, random forests do not naturally produce predictive intervals. To address this, quantile regression forests (QRF; [Meinshausen and Ridgeway, 2006](#)) are used to obtain 95% prediction intervals, by estimating the lower and upper quantiles of the empirical conditional distribution of the response given the predictors across all observations in the terminal leaves. Implementation uses the `ranger` package in R ([Wright and Ziegler, 2017](#)), and the optimal combination of tuning parameters (`num.trees`, `mtry`, `min.size`, and `max.depth`) is selected via 10-fold cross-validation as described in Section 3.5.

## 3.5 Study design

### 3.5.1 Spatial cross-validation

The primary goal of this study is prediction at unmonitored locations across Scotland, so model performance must be evaluated through an out-of-sample spatial validation experiment rather than a purely temporal one. For each pollutant, the set of monitoring sites is randomly partitioned into a training set comprising 80% of sites and a test set comprising the remaining 20%, with all observations from a given site assigned entirely to one partition. Predictions are then made for the held-out test sites and compared against the corresponding observations using the performance metrics described in Section 3.5.2. To ensure that results are not overly sensitive to any single partition, this process is repeated ten times using different random splits, and the reported performance metrics are averaged across ten replicates.

It is worth noting why a spatial rather than temporal validation design is adopted here. A temporal split, in which earlier months are used for training and later months for testing, evaluates the ability of a model to forecast future concentrations at already monitored locations, which is however not the primary aim here. Instead, the key inferential goal is to predict concentrations at the large number of unmonitored  $1 \text{ km}^2$  grid squares across Scotland, which requires assessing how well each model generalises to new spatial locations. The spatial cross-validation design directly targets this objective, and is therefore the more appropriate choice for the present study.

### 3.5.2 Performance metrics

To compare predictive performance across models and pollutants, five metrics are computed on the test set observations. Let  $y(\mathbf{s}_i, t)$  and  $\hat{y}(\mathbf{s}_i, t)$  denote the observed and predicted concentrations on the original scale at location  $\mathbf{s}_i$  in month  $t$ , and let  $(L(\mathbf{s}_i, t), U(\mathbf{s}_i, t))$  denote the corresponding 95% prediction interval.

**RMSE:** The root mean squared error,

$$\text{RMSE} = \sqrt{\frac{1}{nT} \sum_{i=1}^n \sum_{t=1}^T (y(\mathbf{s}_i, t) - \hat{y}(\mathbf{s}_i, t))^2},$$

measures overall prediction accuracy and penalises large errors more heavily than small ones.

**MAE:** The median absolute error,

$$\text{MAE} = \text{median}\{|y(\mathbf{s}_i, t) - \hat{y}(\mathbf{s}_i, t)|\},$$

provides a complementary measure of accuracy that is more robust to extreme prediction errors than the RMSE.

**Bias:** Systematic over- and under-prediction is quantified by the bias,

$$\text{Bias} = \frac{1}{nT} \sum_{i=1}^n \sum_{t=1}^T (\hat{y}(\mathbf{s}_i, t) - y(\mathbf{s}_i, t)),$$

where values close to zero indicate that predictions are free of systematic error.

**CP:** Coverage probability,

$$\text{CP} = \frac{100}{nT} \sum_{i=1}^n \sum_{t=1}^T \mathbb{I}(L(\mathbf{s}_i, t) \leq y(\mathbf{s}_i, t) \leq U(\mathbf{s}_i, t)),$$

measures the percentage of test observations falling within their 95% prediction interval, with values close to 95% indicating well-calibrated uncertainty. A coverage probability substantially above 95% suggests the intervals are unnecessarily wide and the model is under-confident, while a coverage substantially below 95% indicates intervals that are too narrow and overconfident.

**MIW:** For the two models with similar CP, the mean interval width,

$$\text{MIW} = \frac{1}{nT} \sum_{i=1}^n \sum_{t=1}^T (U(\mathbf{s}_i, t) - L(\mathbf{s}_i, t)),$$

favours the model that achieves the same nominal coverage with more precise intervals.

Model selection is therefore primarily driven by RMSE and MAE, with CP and MIW used to assess whether the accompanying uncertainty quantification is appropriate.

### 3.5.3 Implementation details

Prior to modelling, both the monitoring observations and the PCM modelled concentrations are log-transformed. Air pollutant concentrations are strictly non-negative

and typically exhibit right skewed distributions with a long tail of elevated values, and the log transformation stabilises within-site variance, renders the marginal distributions approximately Gaussian, and ensures compatibility with the Gaussian error assumptions underlying the linear, additive and hierarchical models (Cameletti et al., 2011). All predictions, uncertainty intervals, and performance metrics are subsequently computed on the original concentration scale for interpretability, requiring back-transformation of model outputs as described below.

The complete set of candidate models compared in this study is summarised in Table 3.4, together with their key characteristics. In total, 12 models are evaluated for each pollutant. It should be noted that an initial investigation also included bagging models as additional candidates, but these were ultimately excluded from the final comparison. Since random forests include the core aggregation principle of bagging while additionally reducing between-tree correlation through predictor subsampling, bagging offered no meaningful advantage or empirical predictive improvement, and is omitted here for conciseness.

Table 3.4: Summary of the candidate models compared in this study.

| Model               | Description                                            | Scale    |
|---------------------|--------------------------------------------------------|----------|
| LM <sub>cs</sub>    | Linear model, constant seasonality                     | Log      |
| LM <sub>vs</sub>    | Linear model, varying seasonality                      | Log      |
| AM                  | Additive model                                         | Log      |
| AM <sub>sp</sub>    | Additive model with spatial smooth                     | Log      |
| AMLSS               | Additive model, heteroscedastic errors                 | Log      |
| AMLSS <sub>sp</sub> | AMLSS with spatial smooth                              | Log      |
| SP <sub>gp</sub>    | Hierarchical Gaussian process model                    | Log      |
| SP <sub>ar</sub>    | Hierarchical autoregressive model                      | Log      |
| RF <sub>oc</sub>    | Random forest, original scale                          | Original |
| RF <sub>lc</sub>    | Random forest, log scale                               | Log      |
| RF <sub>oc_sp</sub> | Random forest with spatial coordinates, original scale | Original |
| RF <sub>lc_sp</sub> | Random forest with spatial coordinates, log scale      | Log      |

### 3.5.4 Construction of predictive uncertainty

For models building on the log scale, a naive back-transformation via exponentiation yields an estimate of the median rather than the mean of the original-scale distribution. Under a lognormal assumption, the mean is  $\exp\left(\mu(\mathbf{s}_i, t) + \frac{1}{2}\sigma_\epsilon^2\right)$ , so point predictions are corrected accordingly by multiplying  $\exp(\hat{\mu}(\mathbf{s}_i, t))$  by  $\exp\left(\frac{1}{2}\hat{\sigma}_\epsilon^2\right)$ . For AMLSS



models, where the variance is location- and time-varying, the correction becomes  $\exp\left(\hat{\mu}(\mathbf{s}_i, t) + \frac{1}{2}\hat{\sigma}_\epsilon^2(\mathbf{s}_i, t)\right)$ . For the hierarchical spatio-temporal models, predictions are taken as the posterior median of the MCMC samples on the log scale and then exponentiated directly, with 95% prediction intervals obtained from the 2.5th and 97.5th percentiles of the posterior predictive samples. For the random forest models applied to the log-transformed concentrations, predictions and interval bounds are similarly exponentiated without further correction, as quantile forests directly target the conditional quantiles of the response distribution rather than its mean.

### 3.5.5 Hyperparameter tuning for random forests

Random forests require the selection of four tuning parameters: the number of trees `num.trees`, the number of predictor candidates considered at each split `mtry`, the minimum number of observations in a terminal node `min.size`, and the maximum tree depth `max.depth`. These parameters collectively govern the complexity and diversity of the ensemble, and their optimal values are data dependent. They are therefore selected empirically using 10-fold cross-validation applied to the training set, separately for each pollutant and each of the ten training-test splits described in Section 3.5.1.

Specifically, the training set is randomly divided into ten folds at the monitoring site level. For each candidate combination of tuning parameters, the model is fitted to nine folds and used to predict the held-out tenth, and this process is repeated across all ten folds. The combination minimising the average RMSE across the ten folds is then selected, and the random forest is re-fitted to the entire training set using these optimal values before making predictions for the test set. This ensures that tuning parameter selection relies solely on training data, so that all test set predictions remain fully out of sample.

The grid of candidate values explored is `num.trees`  $\in \{100, 200, 500, 1000\}$ , `mtry`  $\in \{4, 6, 8, 10\}$ , `min.size`  $\in \{1, 3, 5, 10, 15\}$ , and `max.depth`  $\in \{1, 5, 10, 15, 20, 30\}$ , yielding 480 combinations in total. The optimal tuning parameters selected for the best performing random forest models are presented in Table 3.5, which shows that 500 trees were sufficient to stabilise predictive performance for all three pollutants, while the optimal values of the other three varied across pollutants.

Table 3.5: Optimal hyperparameters selected via 10-fold cross-validation for the random forest models.

|                   | num.trees             | mtry          | min.size          | max.depth              |
|-------------------|-----------------------|---------------|-------------------|------------------------|
| Range             | {100, 200, 500, 1000} | {4, 6, 8, 10} | {1, 3, 5, 10, 15} | {1, 5, 10, 15, 20, 30} |
| NO <sub>2</sub>   | 500                   | 4             | 3                 | 10                     |
| PM <sub>10</sub>  | 500                   | 4             | 3                 | 15                     |
| PM <sub>2.5</sub> | 500                   | 8             | 3                 | 20                     |

### 3.6 Model comparison for the three air pollutants

This section presents the results of the spatial cross-validation experiment described in Section 3.5, comparing the predictive performance of the 12 candidate models across the three pollutants. Table 3.6 summarises the five performance metrics averaged across the ten training-test splits. Scatter plots of observed versus predicted concentrations for all models and pollutants are presented in Figures 3.13 to 3.15.

For NO<sub>2</sub>, the random forest model applied to the original concentration scale (RF<sub>oc</sub>) achieves the lowest RMSE (7.70) and a coverage probability close to the nominal 95% level (94.97%), making it the best overall performer for this pollutant. The strong performance of random forests for NO<sub>2</sub> is consistent with the highly localised and traffic-driven nature of this pollutant, whose spatial variation is largely captured through the covariate set rather than residual spatial autocorrelation. The linear model with varying seasonality (LM<sub>vs</sub>) also performs competitively, achieving an RMSE of 7.82 and appropriate coverage, which highlights the value of the rich covariate set even under a simple linear structure. In contrast, models incorporating explicit spatial smoothing components (AM<sub>sp</sub>, AMLSS<sub>sp</sub>, SP<sub>gp</sub> and SP<sub>ar</sub>) perform markedly worse, with higher RMSE and MAE values and poor coverage probabilities. This suggests that spatial smoothing is not appropriate for NO<sub>2</sub> in this setting, likely because the highly localised nature of traffic-related emissions produces spatial variation that is better explained by covariates than by smooth spatial processes. The scatter plots in Figure 3.13 confirm that the random forest models provide the closest agreement between observed and predicted concentrations, though they tend to under-predict the highest observed NO<sub>2</sub> values.

For PM<sub>10</sub>, the linear model LM<sub>vs</sub> achieves the lowest RMSE (2.73) and MAE (1.44) with near-nominal coverage probability (93.81%), and is selected as the best performing model for this pollutant based on both predictive performance and parsimony. The random forest variants perform similarly, with RMSEs between 2.75

and 2.78, and also provide good uncertainty quantification. As with  $\text{NO}_2$ , models with spatial smoothing terms perform worse, with the additive spatial variants ( $\text{AM}_{\text{sp}}$  and  $\text{AMLSS}_{\text{sp}}$ ) exhibiting elevated RMSE values and poor coverage. The hierarchical spatio-temporal models ( $\text{SP}_{\text{gp}}$  and  $\text{SP}_{\text{ar}}$ ) achieve competitive RMSE values but produce severely over-covering prediction intervals, with coverage probabilities approaching 100%, indicating that their posterior predictive variances are large and the models are under-confident. For  $\text{PM}_{2.5}$ , the four random forest variants collectively achieve the lowest RMSE values (all equal to 1.15) and the most appropriate coverage probabilities, ranging from 94.58% to 94.88%. Among these,  $\text{RF}_{\text{lc}}$  is selected as the preferred model as it achieves a slightly lower MAE (0.54) than the other variants.

Across all three pollutants, the results consistently favour models that rely on the covariate set to explain spatial variation rather than those that impose explicit spatial smoothing structures. Random forests emerge as the best or near-best performers in terms of both point prediction accuracy and uncertainty quantification, while simple linear models with flexible temporal parameterisation also perform surprisingly well. Based on these results,  $\text{RF}_{\text{oc}}$ ,  $\text{LM}_{\text{vs}}$  and  $\text{RF}_{\text{lc}}$  are selected as the final models for generating monthly concentration maps for  $\text{NO}_2$ ,  $\text{PM}_{10}$  and  $\text{PM}_{2.5}$  respectively, and their predictions are examined in detail in Section 3.7.

## 3.7 Results from the best performing prediction models

This section presents predictions from the three best performing models identified in Section 3.6:  $\text{RF}_{\text{oc}}$  for  $\text{NO}_2$ ,  $\text{LM}_{\text{vs}}$  for  $\text{PM}_{10}$ , and  $\text{RF}_{\text{lc}}$  for  $\text{PM}_{2.5}$ . Each model is re-fitted to the entire dataset before generating monthly predictions and 95% prediction intervals on a regular  $1 \text{ km}^2$  grid across Scotland for all 60 months between January 2016 and December 2020. The analysis focuses on three aspects: (i) the spatio-temporal patterns in predicted concentrations, (ii) the long-term changes in concentrations over the study period, and (iii) the spatial distribution of predictive uncertainty.

### 3.7.1 Spatio-temporal trends in predicted concentrations

Figure 3.16 presents line plots of the spatially averaged monthly predicted concentrations across all  $1 \text{ km}^2$  grid squares in Scotland for each pollutant between 2016

Table 3.6: Predictive model comparison for NO<sub>2</sub>, PM<sub>10</sub> and PM<sub>2.5</sub>, averaged across ten spatial cross-validation splits.

| Pollutant         | Model                  | RMSE  | MAE  | Bias  | CP     | MIW    |
|-------------------|------------------------|-------|------|-------|--------|--------|
| NO <sub>2</sub>   | LM <sub>cs</sub>       | 7.97  | 4.23 | 0.78  | 92.85% | 29.51  |
|                   | LM <sub>vs</sub>       | 7.82  | 4.16 | 0.75  | 92.28% | 28.42  |
|                   | AM                     | 9.71  | 4.94 | 0.25  | 84.02% | 25.37  |
|                   | AM <sub>sp</sub>       | 12.29 | 5.92 | 1.51  | 68.64% | 23.48  |
|                   | AMLSS                  | 9.55  | 4.79 | -0.44 | 82.87% | 24.80  |
|                   | AMLSS <sub>sp</sub>    | 12.31 | 5.84 | 1.06  | 67.26% | 22.80  |
|                   | SP <sub>gp</sub>       | 12.14 | 6.01 | 1.45  | 99.46% | 98.23  |
|                   | SP <sub>ar</sub>       | 12.15 | 5.99 | 1.37  | 97.98% | 134.49 |
|                   | <b>RF<sub>oc</sub></b> | 7.70  | 4.36 | 0.32  | 94.97% | 32.26  |
|                   | RF <sub>lc</sub>       | 7.98  | 4.19 | -0.82 | 95.02% | 32.67  |
|                   | RF <sub>oc_sp</sub>    | 7.91  | 4.49 | 0.38  | 95.27% | 33.52  |
|                   | RF <sub>lc_sp</sub>    | 8.15  | 4.30 | -0.85 | 95.20% | 34.03  |
| PM <sub>10</sub>  | LM <sub>cs</sub>       | 3.09  | 1.71 | 0.08  | 94.04% | 10.55  |
|                   | <b>LM<sub>vs</sub></b> | 2.73  | 1.44 | 0.15  | 93.81% | 9.32   |
|                   | AM                     | 3.01  | 1.57 | 0.19  | 89.44% | 9.02   |
|                   | AM <sub>sp</sub>       | 3.81  | 1.74 | 0.06  | 83.21% | 8.53   |
|                   | AMLSS                  | 2.99  | 1.56 | -0.03 | 88.37% | 8.74   |
|                   | AMLSS <sub>sp</sub>    | 4.01  | 1.79 | -0.12 | 80.94% | 8.22   |
|                   | SP <sub>gp</sub>       | 2.75  | 1.47 | -0.11 | 99.86% | 25.30  |
|                   | SP <sub>ar</sub>       | 2.80  | 1.47 | -0.17 | 99.93% | 35.29  |
|                   | RF <sub>oc</sub>       | 2.78  | 1.52 | 0.19  | 93.89% | 10.99  |
|                   | RF <sub>lc</sub>       | 2.76  | 1.49 | 0.04  | 94.11% | 11.00  |
|                   | RF <sub>oc_sp</sub>    | 2.77  | 1.51 | 0.22  | 93.99% | 11.17  |
|                   | RF <sub>lc_sp</sub>    | 2.75  | 1.47 | 0.04  | 94.07% | 11.16  |
| PM <sub>2.5</sub> | LM <sub>cs</sub>       | 1.56  | 0.81 | -0.00 | 94.19% | 5.30   |
|                   | LM <sub>vs</sub>       | 1.18  | 0.57 | 0.02  | 94.11% | 4.12   |
|                   | AM                     | 1.30  | 0.66 | 0.08  | 90.79% | 3.95   |
|                   | AM <sub>sp</sub>       | 1.27  | 0.69 | 0.12  | 89.66% | 3.70   |
|                   | AMLSS                  | 1.31  | 0.62 | -0.02 | 88.24% | 3.69   |
|                   | AMLSS <sub>sp</sub>    | 1.30  | 0.67 | 0.04  | 85.24% | 3.41   |
|                   | SP <sub>gp</sub>       | 1.15  | 0.54 | -0.12 | 99.98% | 24.43  |
|                   | SP <sub>ar</sub>       | 1.20  | 0.61 | 0.02  | 99.49% | 11.54  |
|                   | RF <sub>oc</sub>       | 1.15  | 0.56 | 0.08  | 94.67% | 4.69   |
|                   | <b>RF<sub>lc</sub></b> | 1.15  | 0.54 | 0.03  | 94.58% | 4.75   |
|                   | RF <sub>oc_sp</sub>    | 1.15  | 0.57 | 0.09  | 94.81% | 4.74   |
|                   | RF <sub>lc_sp</sub>    | 1.15  | 0.55 | 0.03  | 94.88% | 4.85   |

and 2020, with each year represented by a separate line. Figure 3.17 presents the

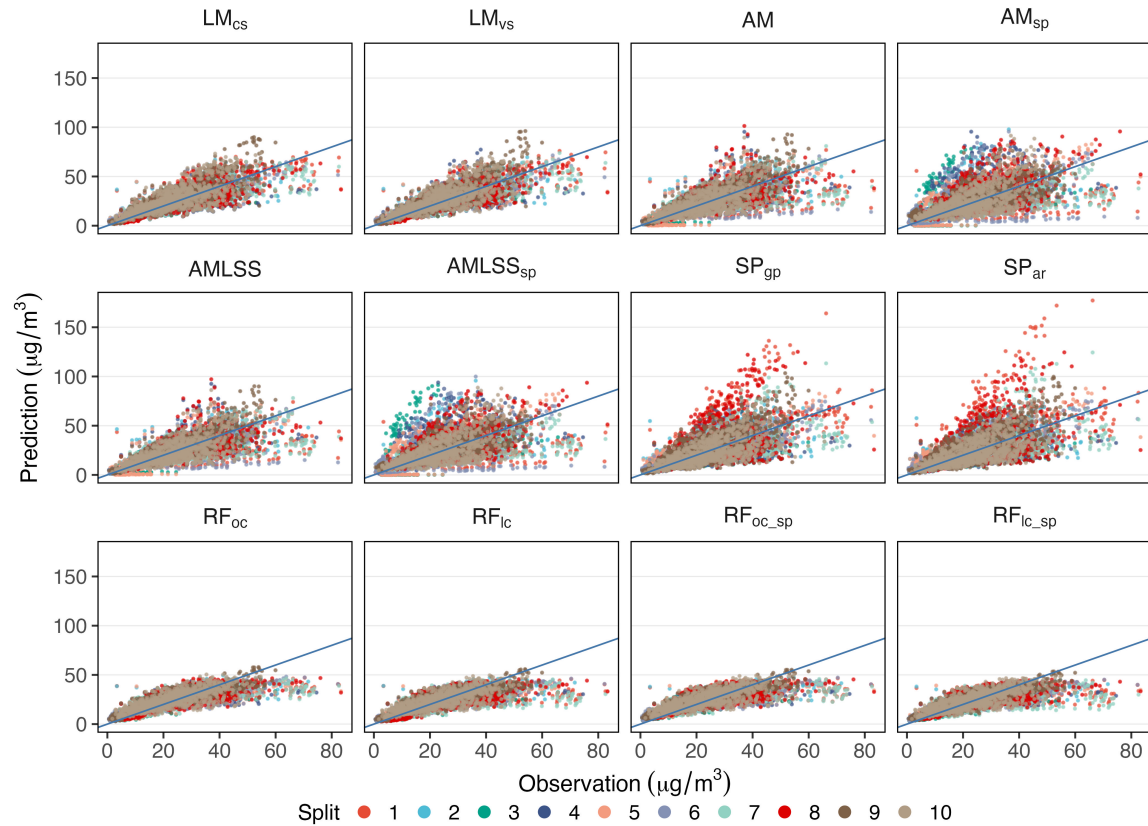


Figure 3.13: Scatter plots of the observed versus predicted  $\text{NO}_2$  concentrations across ten test sets.

corresponding seasonal spatial maps.

$\text{NO}_2$  exhibits a strong and consistent seasonal cycle across all five years, with concentrations approximately twice as high in winter as in summer, peaking around December–January and reaching a minimum around July. This reflects increased vehicular use and domestic heating in winter, and enhanced photochemical conversion of  $\text{NO}_2$  to ozone under stronger summer sunlight ([Air Quality Expert Group, 2004](#)). A general downward trend is present across the study period, most pronounced in 2020 and partly attributable to Covid-19 lockdowns, though the decline begins in January 2020 before lockdowns were imposed, suggesting longer-term improvements in vehicle emission standards might also contribute. Figure 3.17a shows that as expected there are higher mean  $\text{NO}_2$  concentrations across Scotland in winter compared to summer, which is consistent with the results from Figure 3.16 above. Urban areas appear to exhibit the biggest seasonal changes, with increases from around  $20 \mu\text{g}/\text{m}^3$  in summer to  $30 \mu\text{g}/\text{m}^3$  in winter. In contrast, rural areas exhibit minimal seasonal variation, with most grid squares changing by less than  $3 \mu\text{g}/\text{m}^3$  on average between

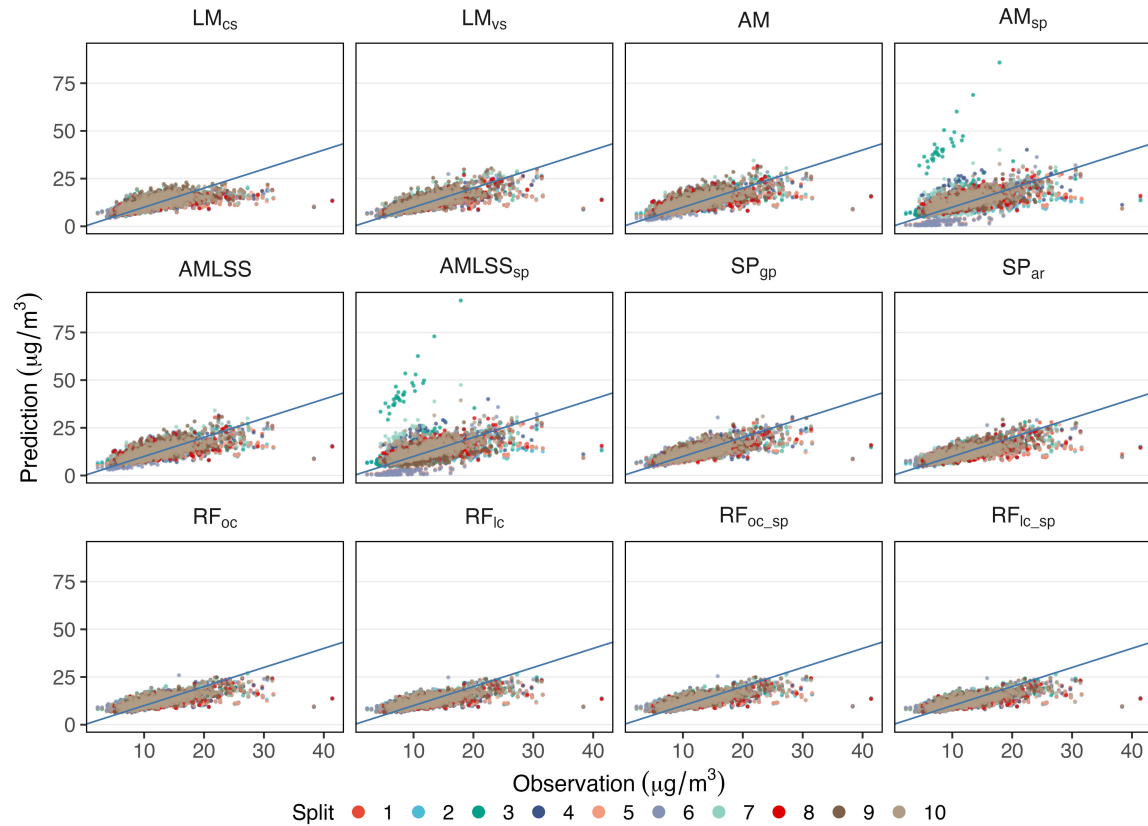


Figure 3.14: Scatter plots of the observed versus predicted  $\text{PM}_{10}$  concentrations across ten test sets.

summer and winter.

$\text{PM}_{10}$  shows a less regular seasonal signal than  $\text{NO}_2$ , with mean concentrations across Scotland ranging from  $5.79 \mu\text{g}/\text{m}^3$  in summer to  $8.11 \mu\text{g}/\text{m}^3$  in spring. The time series of  $\text{PM}_{10}$  shows a pronounced spike in early 2019, corresponding to two high pollution episodes in February and April driven by long-range transport of particulate matter from continental Europe ([Department for Environment, Food and Rural Affairs, 2020](#)). Concentrations in 2020 are generally lower than in previous years, consistent with reduced emissions during the lockdown period. The seasonal spatial maps reveal a persistent east–west divide across all seasons, with higher concentrations in urban areas and eastern regions reflecting both population density and trans-boundary pollution, and lower concentrations in the north-western Highlands. Spatial variability is greatest in spring, with a standard deviation of  $1.74 \mu\text{g}/\text{m}^3$ , and lowest in summer and autumn at  $1.33 \mu\text{g}/\text{m}^3$ .

$\text{PM}_{2.5}$  exhibits a moderate seasonality, with concentrations generally peaking in winter and spring and declining in summer. As with  $\text{PM}_{10}$ , the pollution episodes in

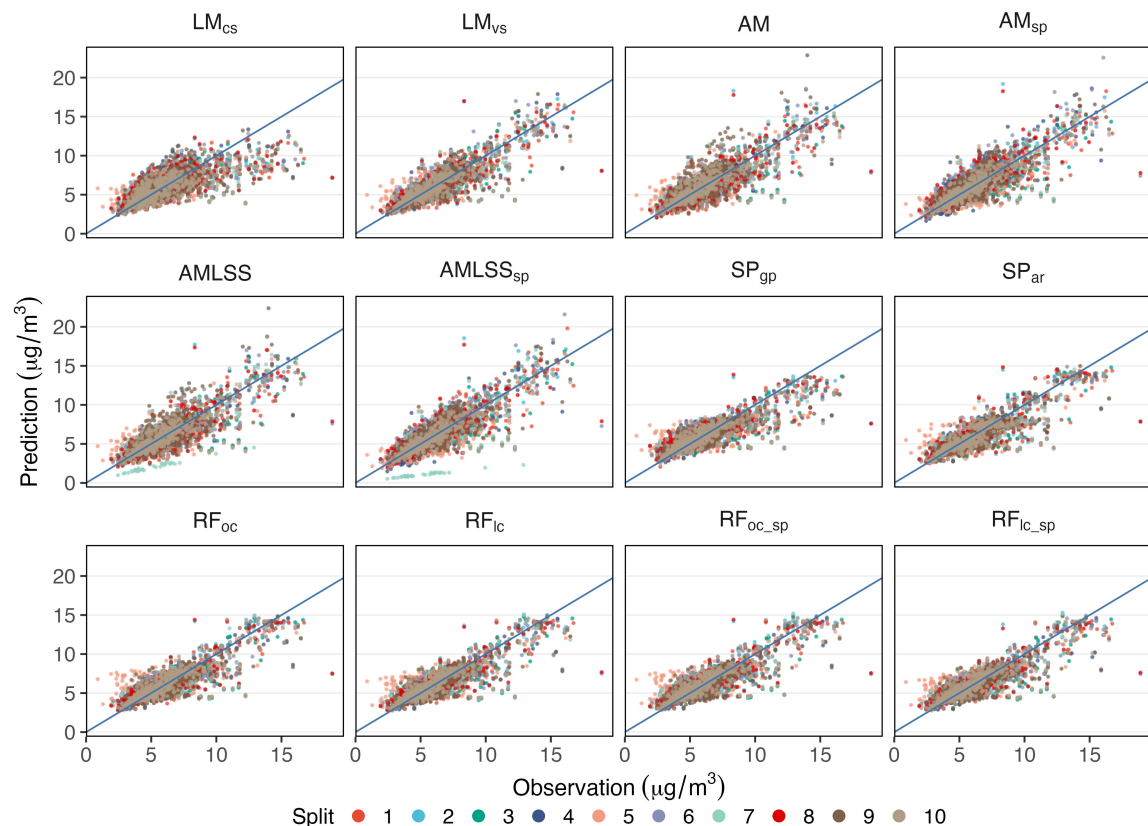


Figure 3.15: Scatter plots of the observed versus predicted  $\text{PM}_{2.5}$  concentrations across ten test sets.

2019 are clearly visible in the time series, and concentrations in 2020 are generally lower than preceding years. The seasonal spatial maps (see Figure 3.17c) show consistently elevated concentrations in the urban centres and along the eastern coast across all seasons, with the highest average concentrations occurring in spring. This spring peak is partly a consequence of the April 2019 episode, but also reflects elevated nitrate aerosol loading from agriculture operations across the UK and continental Europe during this period (Department for Environment, Food and Rural Affairs, 2023). In addition,  $\text{PM}_{2.5}$  shows the least regional variation in summer, with a spatial standard deviation of  $0.354 \mu\text{g}/\text{m}^3$  compared to  $0.924 \mu\text{g}/\text{m}^3$  in winter.

The predicted concentrations can also be compared against current regulatory standards. During the study period, the predicted five-year average concentrations of  $\text{NO}_2$  across Scotland range between  $4.05$  and  $22.9 \mu\text{g}/\text{m}^3$  in urban areas and between  $2.01$  and  $18.8 \mu\text{g}/\text{m}^3$  in rural ones, which compare to targets for annual mean concentrations (not 5-year averages) of  $40 \mu\text{g}/\text{m}^3$  from the National Air Quality Strategy (NAQS) and  $10 \mu\text{g}/\text{m}^3$  from the more recent World Health Organisation



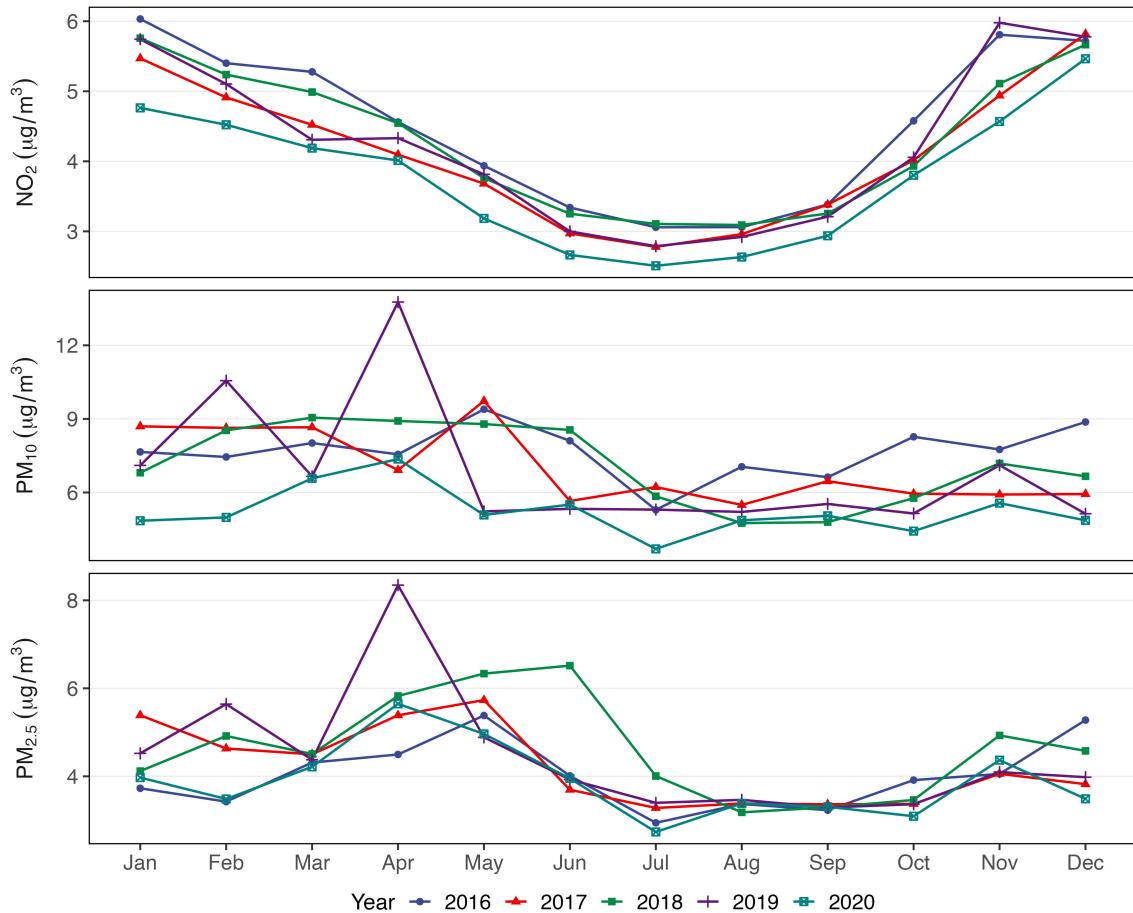


Figure 3.16: Line plots of the mean (over all  $1 \text{ km}^2$  grid squares in Scotland) monthly predicted concentrations of  $\text{NO}_2$ ,  $\text{PM}_{10}$  and  $\text{PM}_{2.5}$  between 2016 and 2020.

(WHO) guidelines that were enacted after the study period in 2021. For  $\text{PM}_{10}$ , these targets are  $18 \mu\text{g}/\text{m}^3$  (NAQS) and  $15 \mu\text{g}/\text{m}^3$  (WHO) respectively, and the five-year average concentrations during the study were below these levels across the country (see Figure 3.17b). Finally, for  $\text{PM}_{2.5}$  these targets are  $10 \mu\text{g}/\text{m}^3$  (NAQS) and  $5 \mu\text{g}/\text{m}^3$  (WHO), and while the former was achieved for all of Scotland on average over the study period, the latter is not going to be easy to achieve as 12.06% of the grid squares did not meet this target, albeit as a five-year rather than an annual average. Expressed as a probability of exposure, the five-year mean concentrations lie below the NAQS targets at every grid square for all three pollutants, so the probability of a location exceeding a NAQS target is effectively zero across Scotland. Against the more stringent 2021 WHO guidelines,  $\text{PM}_{10}$  also remains below its guideline everywhere, whereas the WHO guideline is exceeded over 12.06% of the country for  $\text{PM}_{2.5}$  and across the most heavily trafficked urban centres for  $\text{NO}_2$ , whose five-year



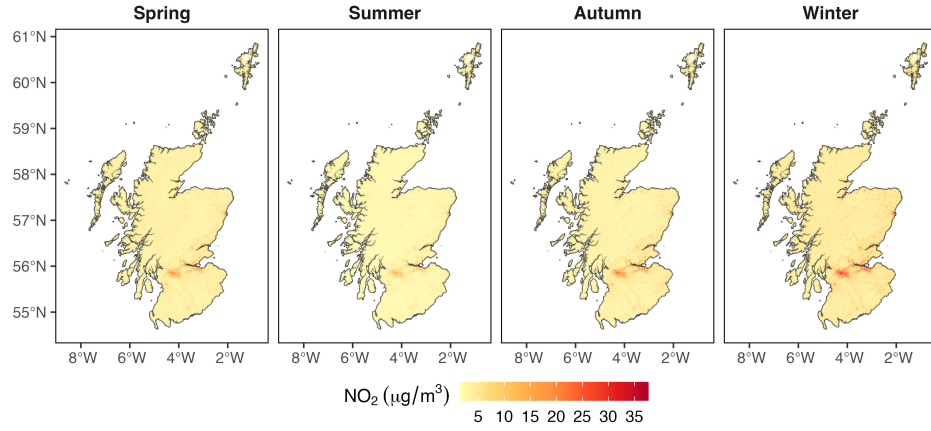
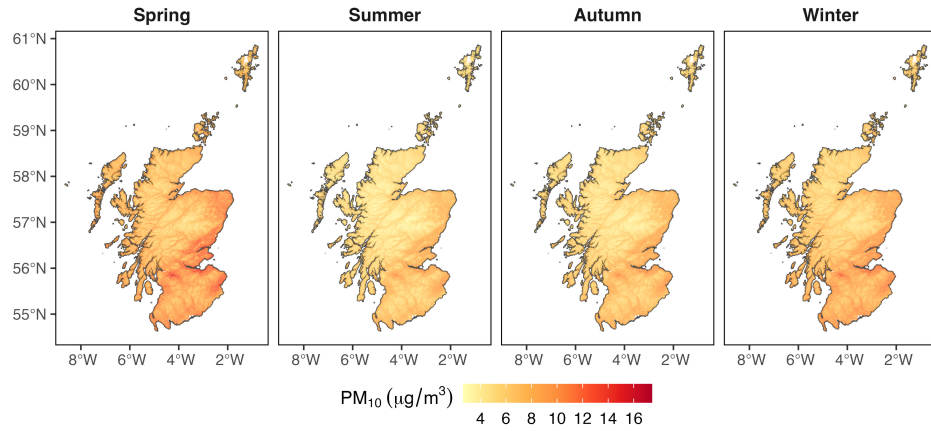
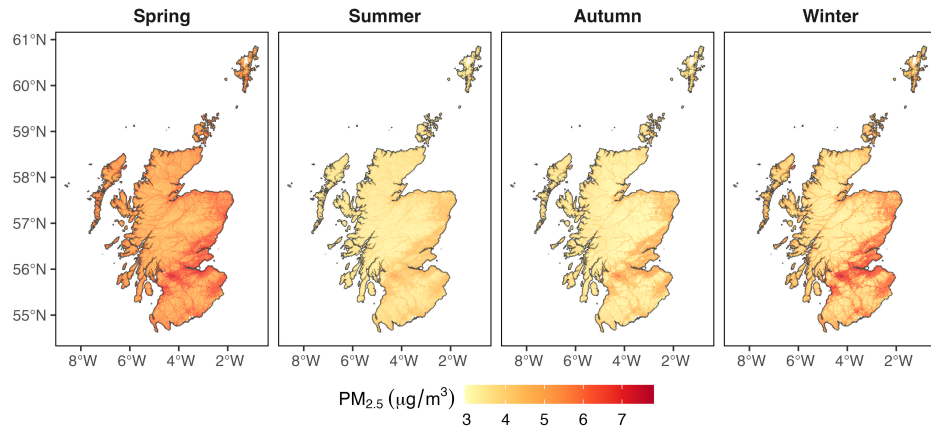
(a)  $\text{NO}_2$ (b)  $\text{PM}_{10}$ (c)  $\text{PM}_{2.5}$ 

Figure 3.17: Predicted seasonal mean concentrations on the  $1\text{ km}^2$  grid over 2016–2020 for (a)  $\text{NO}_2$ , (b)  $\text{PM}_{10}$  and (c)  $\text{PM}_{2.5}$ . Spring: {March, April, May}; summer: {June, July, August}; autumn: {September, October, November}; and winter: {December, January, February}.

mean reaches  $22.9 \mu\text{g}/\text{m}^3$  relative to the WHO guideline of  $10 \mu\text{g}/\text{m}^3$ .

### 3.7.2 Long-term changes in concentrations

To quantify the long-term temporal trend at each location, a pixel-specific normal linear model is fitted with the annual mean predicted concentration as the response and calendar year (scaled to  $t = 1, \dots, 5$ ) as the predictor. The estimated regression slope for each  $1 \text{ km}^2$  grid square is presented in Figure 3.18, where negative values indicate decreasing concentrations over the study period.

$\text{NO}_2$  concentrations have reduced across most of Scotland, with declines of between  $1 \mu\text{g}/\text{m}^3$  and  $3 \mu\text{g}/\text{m}^3$  per year in the major urban centres of Glasgow and Edinburgh. Some areas of the northern Highlands show little change or a marginal increase, though this should be interpreted cautiously given the near-absence of monitoring sites there. For  $\text{PM}_{10}$ , a universal decreasing trend is observed across all of Scotland, with reductions ranging between  $0.1 \mu\text{g}/\text{m}^3$  and  $1.0 \mu\text{g}/\text{m}^3$  per year. For  $\text{PM}_{2.5}$ , urban centres and eastern coastal areas exhibit decreasing trends with slopes between 0 and  $-0.4 \mu\text{g}/\text{m}^3$  per year, while the Highlands and south-western areas show increasing trends with slopes ranging between 0 and 0.2.

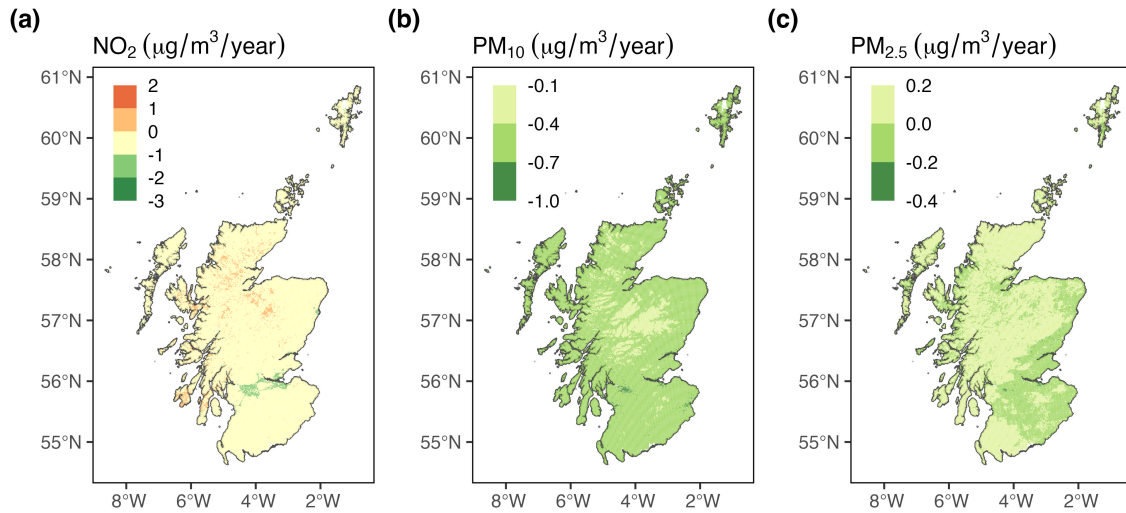


Figure 3.18: Estimated slopes of the temporal trend for  $\text{NO}_2$ ,  $\text{PM}_{10}$  and  $\text{PM}_{2.5}$ , obtained by regressing annual mean predicted concentration on normalised calendar year at each  $1 \text{ km}^2$  grid square. Negative values indicate decreasing concentrations.

### 3.7.3 Predictive uncertainty quantification

Figure 3.19 presents maps of the mean widths of the monthly 95% prediction intervals for each pollutant, averaged over all 60 months. For  $\text{NO}_2$ , predictive uncertainty is highest in the major urban centres, with mean interval widths ranging from  $4.88 \mu\text{g}/\text{m}^3$  in the most rural areas to  $55.20 \mu\text{g}/\text{m}^3$  in the most densely populated grid squares.  $\text{PM}_{10}$  follows a similar pattern, with mean interval widths ranging from  $2.72 \mu\text{g}/\text{m}^3$  to  $11.53 \mu\text{g}/\text{m}^3$  and greater uncertainty in urban areas and the eastern regions where concentrations are highest. The positive relationship between concentration level and predictive uncertainty is consistent for both  $\text{NO}_2$  and  $\text{PM}_{10}$ , which is likely to be due to the approximate log-normality of the data that causes the variances to increase with the mean. In contrast, predictive uncertainty for  $\text{PM}_{2.5}$  is generally larger in the rural areas where concentrations are lower but the monitoring sites are very sparse. Since  $\text{PM}_{2.5}$  concentrations vary relatively little across Scotland, the sparsity of the monitoring network in remote areas dominates over the concentration level as the primary driver of predictive uncertainty.

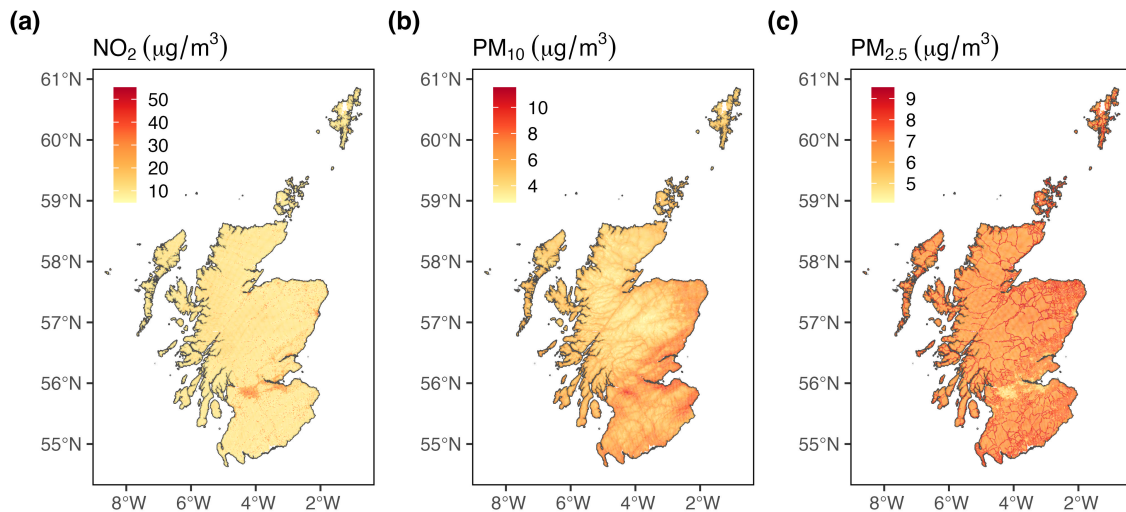


Figure 3.19: Average widths of the monthly 95% prediction intervals on the  $1 \text{ km}^2$  grid over the period 2016–2020 for  $\text{NO}_2$ ,  $\text{PM}_{10}$  and  $\text{PM}_{2.5}$ .

## 3.8 Discussion

This chapter has presented a comparison of statistical and machine learning approaches for predicting monthly air pollution concentrations in a Scottish context, and has provided a new data resource comprising monthly average predictions and

95% prediction intervals of  $\text{NO}_2$ ,  $\text{PM}_{10}$  and  $\text{PM}_{2.5}$  at a  $1 \text{ km}^2$  resolution for all of Scotland between 2016 and 2020. The candidate models range from normal linear models commonly used in a land use regression context to more complex approaches that allow for non-linear predictor-response relationships and spatio-temporal autocorrelations, as well as random forest algorithms. These models were applied in a temporal downscaling (Berrocal et al., 2010) and data-integration context, fusing pollutant concentration data from multiple sources with different spatio-temporal resolutions, alongside climate model outputs and geospatial inputs such as the road network and population density.

The results show that, overall, random forests had the best predictive performance, achieving optimal values of RMSE and MAE for  $\text{NO}_2$  and  $\text{PM}_{2.5}$  and near optimal values for  $\text{PM}_{10}$ . Linear and additive models also yielded strong predictive capabilities, with RMSE and MAE values that were only slightly above those from the random forests. These methods also provided appropriate uncertainty quantification, as their 95% prediction intervals had close to the nominal coverage levels. For  $\text{PM}_{10}$ , the linear model  $\text{LM}_{\text{vs}}$  was selected as the final model over alternatives with similar predictive performance on the grounds of parsimony. In contrast, the models with spatio-temporal smoothing components gave comparatively poor results, particularly for  $\text{NO}_2$  which is highly traffic dependent and hence more localised spatially with very short range spatial autocorrelations.

Collectively, these results could potentially be caused by a number of factors. The first is that the availability of informative spatially-structured predictors allowed the models to capture the majority of the spatial variation in pollutant concentrations through covariate effects, rather than relying on a spatially smooth covariance structure or spatial smoothing term. The suggestion is, therefore, that there was little spatial autocorrelation remaining in the data for the models with extra spatio-temporal structure to capture. Moreover, in the models that included explicit spatio-temporal effects (i.e., based on the spatial location of the monitoring sites), prediction at unmonitored locations in Scotland often requires extrapolating these effects well beyond the spatial range of the monitoring site data. Given that the monitoring network is sparse and unevenly distributed in the Scottish context, this extrapolation is likely to degrade predictive performance relative to models that rely solely on covariate effects.

The predictions from the best performing models show that elevated concentrations of  $\text{NO}_2$  are highly localised to the urban centres, while particulate matter also exhibits elevated concentrations on the eastern coast due to transboundary pollution

imported from continental Europe.

Figure 3.16 highlights that  $\text{NO}_2$  and  $\text{PM}_{2.5}$  concentrations generally decreased during the study period in Scotland, and more recent data from Air Quality in Scotland (<https://www.scottishairquality.scot/data/trends>) confirm that these decreasing trends have continued in the subsequent years, with statistically significant decreasing trends being observed at most monitoring sites. However, further reducing pollutant concentrations in the future will be increasingly challenging, as it requires reducing emissions from sources critical to human productivity and livelihoods, such as traffic and heating, as well as influencing trans-boundary emissions imported from continental Europe.

A challenge inherent in the UK monitoring context, rather than a shortcoming of this study, is the irregular spatial distribution of monitoring sites, which are predominantly located in urban areas. Where monitoring sites are absent, predictions rely entirely on covariate effects estimated from data in monitored areas, implicitly assuming that predictor-response relationships are universally applicable across Scotland. This concentration of sites in urban areas also means that model validation is dominated by urban performance, potentially under-stating prediction errors in rural regions. However, this concern may be partially mitigated by the fact that rural areas tend to exhibit greater spatial homogeneity in both pollutant concentrations and covariates, which reduces the complexity of the prediction task and suggests that covariate relationships estimated from limited rural monitoring data may remain broadly applicable in these less heterogeneous settings.

In summary, the results of this chapter provide a baseline for future research developments, which on the methodological side include the development of multivariate statistical and machine learning prediction models that borrow strength when making predictions across pollutants. Meanwhile, one of the primary motivations for this work is the production of monthly average pollutant concentration predictions with uncertainty quantification at a fine spatial resolution across Scotland, which will be used as exposure inputs in the epidemiological analyses presented in Chapter 4. Here, I aim to utilise monthly data on prescription rates in primary care split by general practice (GP) surgery, which are available from <https://www.opendata.nhs.scot/dataset/prescriptions-in-the-community>. A key challenge in that subsequent work will be producing representative pollution exposure estimates for the populations who attend each GP surgery, given that their exact locations are only partially known, as well as developing models that appropriately propagate the inherent uncertainty in the exposure estimates through to the health effect estimates.

## Chapter 4

# Quantifying the effects of air pollution on respiratory ill health treated in primary care when the locations of the populations at risk are unknown

### 4.1 Introduction

Air pollution is a complex mixture of particulate matter and gases, and is one of the most important environmental risk factors for non-communicable diseases, affecting global health and wellbeing. In 2019, ambient air pollution was estimated to have caused 4.2 million premature deaths worldwide ([World Health Organization, 2024](#)). In the United Kingdom, it remains the leading environmental risk to public health, contributing to between 28,000 and 36,000 deaths per year ([COMEAP, 2023](#)). The financial burden is also substantial, with air pollution expected to cost the National Health Service (NHS) and social care system £1.6 billion between 2017 and 2025 ([UK Government, 2022](#)). Research clearly shows that long-term exposure to air pollutants including NO<sub>2</sub>, PM<sub>10</sub> and PM<sub>2.5</sub> can affect lung function, trigger asthma and chronic obstructive pulmonary disease (COPD) exacerbations, and increase the risk of lung cancer, imposing a considerable social and economic burden ([Chen et al., 2023](#); [Sin et al., 2023](#); [Zhou et al., 2024](#)). The UK has one of the highest asthma prevalence rates globally, with approximately 5.4 million people receiving treatment ([NICE, 2025a](#)). COPD also affects around 3 million individuals in the UK and accounts for approximately 1.4 million general practitioner (GP) consultations per year ([NICE,](#)



2025b).

Most of the epidemiological research on air pollution and health has focused on severe health outcomes, such as hospital admissions (Lee et al., 2017; Huang et al., 2018; Caillaud et al., 2018) and mortality (Khomenko et al., 2021). Large international collaborative studies, including the World Health Organisation (WHO)-led *Estimating the Morbidity from Air Pollution and its Economic Costs (EMAPeC)* and *Health Risks of Air Pollution in Europe (HRAPIE-2)* projects, have provided expert advisory support for assessing the health effects of air pollution and the associated economic costs of morbidity. However, this almost exclusive focus on severe outcomes may underestimate the full burden of air pollution, because it overlooks its impact on non-hospitalised respiratory illness treated in primary care, where disease cases with comparatively mild symptoms are managed.

A small but growing literature has begun to address these milder outcomes using primary care data. Blangiardo et al. (2016) linked NO<sub>2</sub> concentrations to respiratory medication prescriptions dispensed in GP surgeries in England, using prescription rates as a proxy for the prevalence of respiratory illness not requiring hospitalisation. They found a very small association between NO<sub>2</sub> and GP prescribing rates, with the posterior relative risk of prescribing increasing by an average of 0.07% for a 10 µg/m<sup>3</sup> increase in NO<sub>2</sub>. In a related study, Lee (2018) jointly modelled prescription rates for medications that relieve and prevent chronic respiratory symptoms in Scotland and found that one standard deviation increases in both PM<sub>10</sub> and PM<sub>2.5</sub> were significantly associated with increases of around 1–3% in prescribing rates. More recently, Lee and Anderson (2023) extended this line of work by examining disease risk variation by severity, finding that the health inequalities are greatest for mild respiratory cases treated in primary care and decrease as severity increases, with hospitalisations and deaths being more evenly distributed across society.

Although these studies provide important evidence on the health effects that air pollutants have on non-hospitalised respiratory ill health, they have several limitations that motivate the work in this chapter. First, they relied on relatively short and now outdated time periods. For example, Blangiardo et al. (2016) analysed prescribing data from August 2010 to November 2012 in England, while Lee (2018) used data from October 2015 to August 2016 in Scotland. Updated epidemiological evidence based on more recent and longer-term data is needed to inform current public health policy. Second, these population-level studies related prescription rates for each GP surgery to estimated pollution concentrations at or near the GP surgery location. Blangiardo et al. (2016) estimated exposure at the GP surgery location, whereas Lee

(2018) used arbitrarily defined circular buffer zones around each surgery building, both of which fail to account for where patients who attend each surgery actually live. In reality, patient populations registered at a surgery reside in surrounding residential areas rather than at the surgery itself, and may not even attend their nearest surgery. As a result, exposure assessment based solely on GP surgery locations is inappropriate in this context and would lead to exposure misrepresentation and hence biased estimates of the pollution-health relationship. This problem is exacerbated by the fact that GP catchment areas have only partially known spatial extents. Patients are not recorded at exact geographical location. Instead, each registered patient is allocated only to the Data Zone in which they live. A catchment is therefore observed not as a boundary, but as a set of counts of how many of a surgery's patients reside in each Data Zone. Because patients need not attend their nearest surgery, different surgeries can draw patients from the same Data Zones, so their catchments overlap and no precise geographical boundary is available. Third, previous work has typically modelled residual spatial correlation using standard distance-based neighbourhood matrices (for example,  $D$ -nearest neighbours based on Euclidean distances between surgery locations). Such specifications only reflect the spatial configuration of the surgery buildings and do not directly incorporate information on where the registered patient populations actually live or the extent to which their catchment areas overlap. Consequently, they may fail to capture the true spatial dependence structure among GP-level prescription counts in settings where the locations of populations at risk are only partially observed through intersection data.

In this chapter I will address these limitations by conducting a new Scotland-wide study of the health effects of air pollution on non-hospitalised respiratory ill health using more recent and long-term data (January 2016–December 2020) at monthly resolution. To improve the spatial accuracy of the exposure estimates, I utilise the monthly average air pollution concentrations predicted in Chapter 3 on 1 km<sup>2</sup> grids covering Scotland, and realign them to GP surgery populations by explicitly accounting for where patients are likely to live. This is achieved by a two-stage spatial change-of-support procedure: air pollution and meteorological covariates that are at grid level are first aggregated to Data Zones using area-weighting, and then to GP surgery catchment populations using population intersection data between Data Zones and GP surgeries, rather than relying solely on GP surgery locations. The resulting exposure and covariate metrics are therefore constructed on the same spatial support as the prescription outcomes, while acknowledging that catchment boundaries are only partially known.



From a methodological perspective, this chapter develops and applies a Bayesian spatio-temporal negative binomial model with CAR-AR(1) random effects to quantify the association between air pollution and GP-level respiratory prescribing rates while accounting for overdispersion, spatial correlation and temporal dependence. Building on the framework of [Rushworth et al. \(2014\)](#), I introduce two novel neighbourhood matrices that characterise spatial closeness between GP surgery patient populations based on population overlaps and the adjacency of underlying Data Zones, rather than simple geographic distances between surgery locations. These matrices are combined with a Leroux-type conditional autoregressive (CAR) prior to model residual spatial structure in a way that better reflects observed catchment structure. The model is estimated using the integrated nested Laplace approximation (INLA, [Rue et al. 2009](#)) with a variational Bayes correction ([Van Niekerk et al., 2023](#)) implemented by the R-INLA software ([Blangiardo et al., 2013](#)), allowing full Bayesian inference and efficient uncertainty quantification.

The primary objective of this study is to quantify the effects that the three ambient air pollutants ( $\text{NO}_2$ ,  $\text{PM}_{10}$  and  $\text{PM}_{2.5}$ ) have on chronic respiratory prescription rates in Scotland, using medications prescribed to relieve and prevent asthma and COPD symptoms as proxies for the burden of non-hospitalised respiratory ill health treated in primary care. More specifically, this chapter aims to: (i) update epidemiological evidence on air pollution and primary-care respiratory medication outcomes using a longer and more recent Scotland-wide dataset, providing contemporary effect estimates that are directly relevant for current decision-making; (ii) improve the spatial modelling of GP-level datasets with partially unknown patient locations by developing a coherent spatial realignment strategy from gridded or Data Zone (DZ) level covariates to GP surgery populations using population intersection data, and constructing novel neighbourhood matrices that can capture spatial dependence among GP registered populations more realistically than standard distance-based structures; and (iii) assess whether Covid-19-related disruptions altered the pollution-health associations by comparing results from the pre-pandemic period (2016–2019) with those from the full study period (2016–2020), within a unified spatio-temporal modelling framework. In addition, I perform sensitivity analyses to examine the robustness of the findings to alternative neighbourhood matrix constructions and to the use of lagged pollution exposures.

The remainder of this chapter is organised as follows: Section [4.2](#) describes the study region and data sources, including prescription records, air pollution predictions and confounding variables. Section [4.3](#) details the spatial alignment procedure used

to link air pollution exposure and covariates to GP surgery patient populations when catchment areas are only partially known. Section 4.4 presents exploratory data analysis based on overdispersed negative binomial models, showing the presence of spatial and temporal correlation in the prescription data. Section 4.5 introduces the Bayesian spatio-temporal modelling framework, explaining the CAR-AR(1) structure, the novel neighbourhood matrices and prior specification. Section 4.6 presents the results of the study, focusing on how the newly proposed neighbourhood matrices capture residual spatial autocorrelation, and the estimated effects of air pollution on chronic respiratory prescription rates. Finally, Section 4.7 summarises the main findings, study contributions and directions for future research.

## 4.2 Study region and data

This chapter aims to estimate the impact of air pollution on asthma and chronic obstructive pulmonary disease (COPD), two of the most common chronic respiratory diseases in Scotland ([Scottish Government, 2023](#)). Most environmental epidemiology studies focus on hospitalisations and mortality, and hence underestimate the total burden of air pollution by ignoring non-hospitalised ill health that is managed in primary care. In Scotland, primary care is predominantly provided by general practitioners (GPs) working in GP surgeries, who are responsible for prescribing medications for respiratory disease. Consequently, this chapter investigates the impact of air pollution on respiratory ill health that is not severe enough to require secondary care, by analysing prescribing activity in primary care. Specifically, I will analyse and model GP prescription data for medications used to treat asthma and COPD, collected from a total of  $K = 1006$  surgeries that were operational for at least one month between January 2016 and December 2020, across the 14 health boards that provide health services for the local population and together cover the whole of Scotland.

In 2020, there were 928 GP surgeries in Scotland, decreasing from 964 in 2016, reflecting a broader trend for individual practices to merge into larger entities over time ([Public Health Scotland, 2022](#)). Figure 4.1 presents the spatial distribution of all surgeries that were operational in any month within the study period, colour-coded by health board. Most GP surgeries are located within the central belt of Scotland, where population density is highest, with particularly high concentrations around the major cities of Edinburgh and Glasgow. Due to ongoing establishment, closure and merging of surgeries, the number of active GP surgeries within each health board varies over time. Table 4.1 summarises the total number of surgeries that operated at

any point between 2016 and 2020 in each health board, with numbers ranging from as few as seven in Orkney to as many as 250 in Greater Glasgow and Clyde. This uneven distribution closely reflects underlying population density. Greater Glasgow and Clyde and Lothian, encompassing Glasgow and Edinburgh respectively, have the largest numbers of GP surgeries, corresponding to approximately 1.3 million and 0.98 million registered patients in 2020.

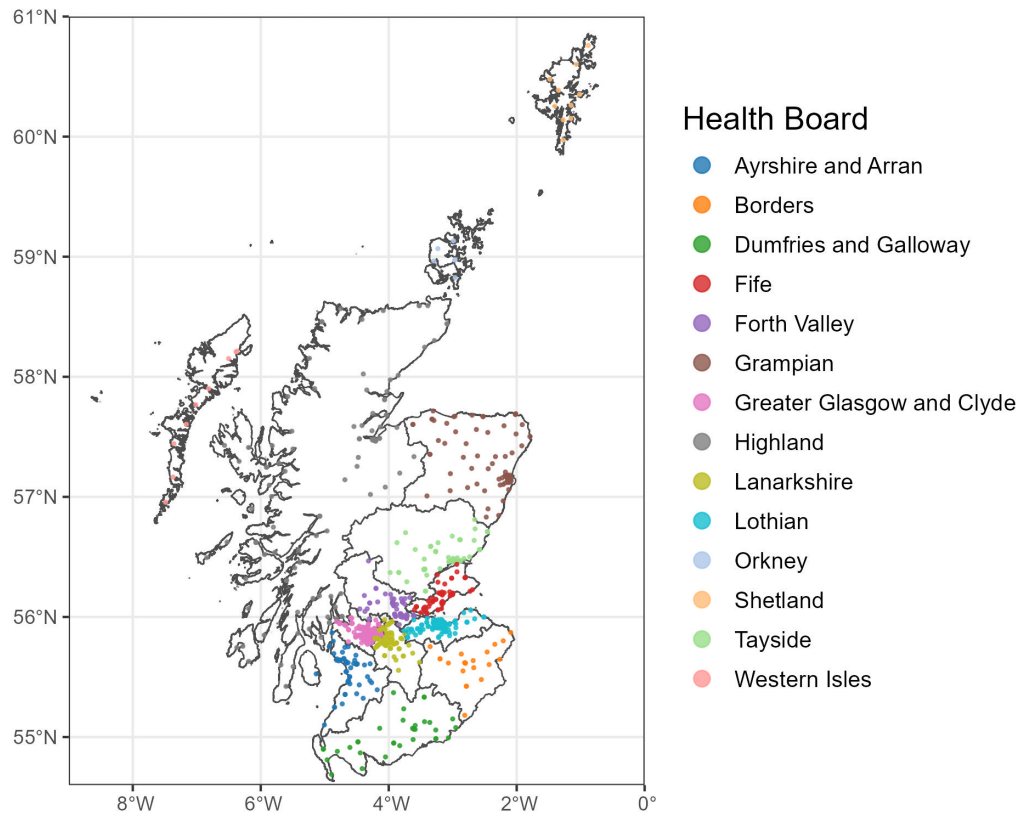


Figure 4.1: Map showing the distribution of GP surgeries across Scotland that were operational in any month between 2016 and 2020, colour-coded by Health Board.

#### 4.2.1 Data sources and spatial supports

The analysis combines prescription, exposure and confounder data that are recorded on three different spatial supports. Reconciling these supports onto a common scale is the central methodological problem of the chapter, so I introduce them here before describing the individual datasets. The three supports are as follows.

- **GP surgery catchments.** The prescription outcomes and the registered patient populations are observed at  $k = 1, \dots, K = 1006$  GP surgeries. Surgery  $k$  is

Table 4.1: The number of GP surgeries in each health board.

| Health Board              | Number of GP surgeries |
|---------------------------|------------------------|
| Ayrshire and Arran        | 57                     |
| Borders                   | 23                     |
| Dumfries and Galloway     | 35                     |
| Fife                      | 62                     |
| Forth Valley              | 61                     |
| Grampian                  | 81                     |
| Greater Glasgow and Clyde | 250                    |
| Highland                  | 104                    |
| Lanarkshire               | 111                    |
| Lothian                   | 129                    |
| Orkney                    | 7                      |
| Shetland                  | 10                     |
| Tayside                   | 67                     |
| Western Isles             | 9                      |

located at  $\mathbf{s}_k$ , but serves a patient catchment  $\mathcal{S}_k$  whose geographical boundary is not directly observed (Section 4.2.2).

- **Data Zones (DZs).** The deprivation and ethnicity confounders are reported on Data Zones, the standard small-area geography of Scotland. Scotland is partitioned into  $M = 6976$  non-overlapping Data Zones  $\{\mathcal{A}_1, \dots, \mathcal{A}_M\}$  each with a resident population of roughly 500–1000.
- **1 km<sup>2</sup> grid.** The air pollution predictions from Chapter 3 and the meteorological confounders are available on  $I = 4,708,065$  grid squares,  $\{\mathcal{H}_1, \dots, \mathcal{H}_I\}$ .

The catchments  $\{\mathcal{S}_k\}$  are known only through their population overlaps with Data Zones. The Data Zone geography therefore serves a dual role, including being the support on which several confounders are reported, and providing the link through which the partially observed catchments are defined. Bringing every variable onto the common GP-catchment support is a spatial change-of-support problem, which I address in Section 4.3. The datasets differ not only in their spatial support but also in their native temporal resolution, and each must be reconciled to the monthly resolution of the analysis. Table 4.2 summarises the source, spatial support and temporal resolution of each dataset. The key remaining difficulty is that the GP catchments are only partially observed, and that the datasets are recorded at differing spatial and temporal resolutions. The next sections describe each dataset in turn, beginning with the GP data (Section 4.2.2), before Section 4.3 sets out

the realignment procedure that places all variables on the common GP-catchment support.

Table 4.2: Datasets used in the health analysis, with their spatial support and temporal resolution. All variables are realigned onto the GP-catchment support and to monthly resolution before modelling (Section 4.3).

| Dataset                              | Spatial support        | Temporal resolution   |
|--------------------------------------|------------------------|-----------------------|
| Prescriptions (reliever, preventer)  | GP surgery             | Monthly               |
| Registered population by age and sex | GP surgery             | Quarterly             |
| Population intersections             | Catchment $\times$ DZ  | Annual                |
| Prevalence rates                     | National, by stratum   | Study-period constant |
| Air pollution predictions            | 1 km <sup>2</sup> grid | Monthly               |
| Meteorology (temperature, humidity)  | 1 km <sup>2</sup> grid | Monthly               |
| Deprivation (SIMD 2020)              | Data Zone              | Time-invariant        |
| Ethnicity (2022 census)              | Data Zone              | Time-invariant        |

## 4.2.2 GP data

This subsection describes the data recorded at GP surgeries, namely the monthly prescription counts together with the quarterly registered-population (age–sex strata) data used for indirect standardisation. The annual catchment data, i.e. the population intersections that link each surgery’s registered patients to Data Zones, were introduced in Section 4.2.1 and are used for the spatial realignment in Section 4.3.

In this study, monthly prescription data on asthma and COPD medications were obtained from the Prescribing Information System (PIS) database (<https://www.opendata.nhs.scot/dataset/prescriptions-in-the-community>) held by Public Health Scotland (PHS), spanning the period from January 2016 to December 2020. For each GP surgery and month, I aggregated the dispensed items for medications that relieve or prevent asthma and COPD symptoms.

Following [Blangiardo et al. \(2016\)](#) and [Lee \(2018\)](#), this study focuses on two broad categories of medications: bronchodilators (labelled Reliever, British National Formulary (BNF) Chapter 3.1) and inhaled corticosteroids (labelled Preventer, BNF Chapter 3.2). These medication groups are commonly used to manage chronic respiratory symptoms, and their prescribing rates provide a proxy measure of the burden of asthma and COPD managed in primary care. However, GP surgeries differ in both their total list size and their age-sex structure. Therefore, to obtain a comparable measure of disease risk across surgeries, I computed expected numbers of patients requiring asthma and COPD prescriptions using indirect standardisation

based on national age-sex specific prevalence rates, obtained from the Scottish Primary Care Information Resource (SPIRE) database and applied to the quarterly age–sex registered population counts for each surgery published by Public Health Scotland, to account for varying population demographics (age and sex) across GP surgeries.

The  $k$ -th GP surgery is located at  $\mathbf{s}_k$ , while the spatial extent of its patient catchment area is denoted by  $\mathcal{S}_k$ . Let  $\mathbb{K}^{(t)} \subset \{1, \dots, K = 1006\}$  be the index set of surgeries that are operating in month  $t$ , which changes over time. For the indirect standardisation, the registered population in each surgery is stratified by age and sex into  $J = 16$  strata like Male 0–4, Female 0–4, Male 5–14, Female 5–14, etc. The expected number of patients eligible for a prescription of respiratory disease medications for surgery  $\mathbf{s}_k$  in month  $t$  is then given by

$$e_t^*(\mathcal{S}_k) = \sum_{j=1}^J n_{kj}^{(t)} \times r_j, \quad k \in \mathbb{K}^{(t)} \subset \{1, \dots, K\},$$

where  $n_{kj}^{(t)}$  is the population size registered at surgery  $\mathbf{s}_k$  in stratum  $j$  during month  $t$ , and  $r_j$  is the national prevalence rate of asthma and COPD for stratum  $j$  over the study period. Here, the population data registered at each GP surgery by age (0–4, 5–14, 15–24, 24–44, 45–64, 65–74, 75–84 and 85+) and sex (male and female) were obtained from <https://www.opendata.nhs.scot/dataset/gp-practice-populations>. These data were published quarterly in January, April, July and October of each year. For months in which no population data are published, I use the data from the nearest available month, implicitly assuming that list sizes and age–sex structure vary smoothly over time. Finally, the expected counts were rescaled to  $e_t(\mathcal{S}_k) = \theta \times e_t^*(\mathcal{S}_k)$  as a proxy measure of expected count of prescriptions, with the rescaling coefficient given by

$$\theta = \frac{\sum_{t=1}^{60} \sum_{k \in \mathbb{K}^{(t)}} Y_t(\mathcal{S}_k)}{\sum_{t=1}^{60} \sum_{k \in \mathbb{K}^{(t)}} e_t^*(\mathcal{S}_k)}.$$

Here  $\theta$  is the national rate of prescriptions per eligible patient over the study period, so that multiplying by  $\theta$  converts the expected number of eligible patients  $e_t^*(\mathcal{S}_k)$  into an expected number of prescriptions. This ensures that  $\sum_{t=1}^{60} \sum_{k \in \mathbb{K}^{(t)}} Y_t(\mathcal{S}_k) = \sum_{t=1}^{60} \sum_{k \in \mathbb{K}^{(t)}} e_t(\mathcal{S}_k)$  so that both observed and expected counts of chronic respiratory medications are on the same scale over the study period.

The national prevalence rates  $\{r_j\}$  for asthma and COPD in Scotland were obtained from the Scottish Primary Care Information Resource (SPIRE) database, which includes data from over 65% of GP surgeries since April 2017. The national average

asthma and COPD prevalence by age and sex over the study period is displayed in Figure 4.2. Prevalence is initially higher in males than in females until adolescence, after which females exhibit higher rates. The sex differences in prevalence by age group are influenced by multiple factors, including socioeconomic status, nutrition, comorbidities, access to health care and occupational exposures (Chowdhury et al., 2021). Childhood asthma disparities can be partly attributed to differences in lung development, with boys showing more dysanaptic growth (Mead, 1980; Lee et al., 2016). Hormonal changes after adolescence are associated with a higher prevalence of asthma in women (Chowdhury et al., 2021). COPD, which is strongly influenced by smoking, is more commonly diagnosed in older male populations, thereby narrowing the sex gap in combined asthma and COPD prevalence after the age of 65. Interestingly, reported prevalence declines in the oldest age group (85+), which may reflect misreporting or underdiagnosis of symptoms, or a “healthy survivor” effect whereby those with the most severe chronic respiratory conditions are less likely to survive into very old age (Delmas et al., 2021).

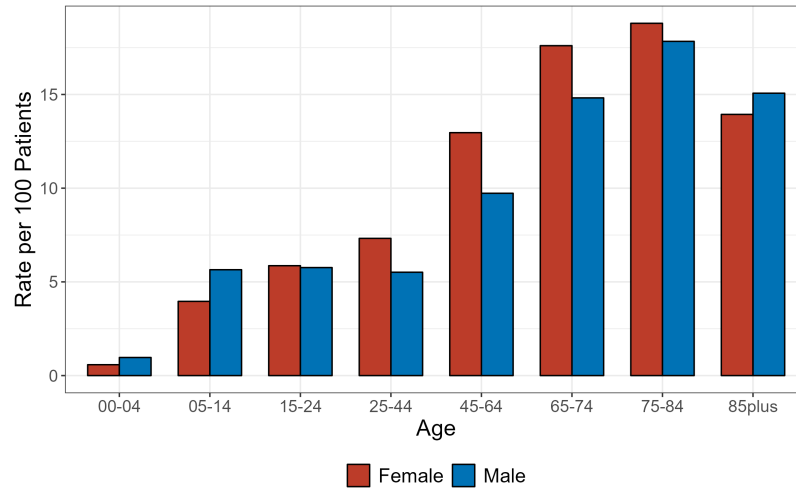


Figure 4.2: National average prevalence  $\{r_j\}$  of asthma and COPD by age and sex in Scotland, obtained from the SPIRE database.

Given the observed counts  $\{Y_t(S_k)\}$  and expected counts  $\{e_t(S_k)\}$  of chronic respiratory prescriptions, the standardised prescription ratio (SPR) is commonly used as a summary measure of disease risk and is given by

$$\text{SPR}_t(S_k) = \frac{Y_t(S_k)}{e_t(S_k)}, \quad k \in \mathbb{K}^{(t)}, \quad t = 1, \dots, 60.$$

Here an SPR of one indicates that the observed prescribing rate at a surgery is equal



to the national average over the study period. Values above one indicate higher risk, with an SPR of 1.2 corresponding to a 20% increased risk, while values below one indicate lower risk (e.g., an SPR of 0.9 corresponds to a 10% decrease).

Figure 4.3 shows the distribution of SPRs across all GP surgeries for each month of the study period. The SPRs exhibit relatively stable trends over time, with notable exceptions in December of each year and March 2020. The elevated SPRs in December are likely due to patients stocking up on medications before the Christmas holiday period, when GP surgeries are typically closed or operating reduced hours. The spike in March 2020 coincides with the onset of the first national Covid-19 lockdown, when Covid-19 was primarily perceived as a respiratory infection. This may have prompted patients with pre-existing asthma and COPD to ensure they had sufficient medication.

The temporal pattern is further illustrated in Figure 4.4, which shows the monthly mean SPR across all GP surgeries in Scotland. The temporal profile in 2020 is distinct from that in earlier years, with an apparent surge in mean SPR in March and April followed by lower-than-usual levels later in the year, consistent with stockpiling behaviour and subsequent reduced prescribing. Clear spikes are also visible in December of each year. The spatial distribution of the SPRs, averaged over the entire study period from January 2016 to December 2020, is shown in Figure 4.5. Surgeries are categorised into three groups (below average ( $< 0.9$ ), around average ( $[0.9, 1.1]$ ), and above average ( $> 1.1$ )) to aid interpretation of spatial patterns and inequalities in disease risk (Lee, 2018). Higher SPRs tend to be observed in parts of Glasgow, southern Scotland and along some coastal areas, indicating regional variation in the burden of asthma and COPD treated in primary care.

### 4.2.3 Air pollution data

The geographic region and monthly temporal resolution of this chapter aligns with the spatio-temporal resolution used in the air pollution modelling study described in Chapter 3. In this chapter, therefore, air pollution exposure is obtained using monthly average concentrations of nitrogen dioxide ( $\text{NO}_2$ ) concentrations, particulate matter with aerodynamic diameter  $\leq 10\mu\text{m}$  ( $\text{PM}_{10}$ ) and fine particulate matter with aerodynamic diameter  $\leq 2.5\mu\text{m}$  ( $\text{PM}_{2.5}$ ) predicted in Chapter 3. The aim of that work is to compare the predictive accuracy of a number of different methodologies, including simple linear and additive models, spatio-temporal smoothing models and machine learning algorithms. These models were fitted to monthly average concentrations measured at a network of fixed-site monitors (94  $\text{NO}_2$ , 93  $\text{PM}_{10}$



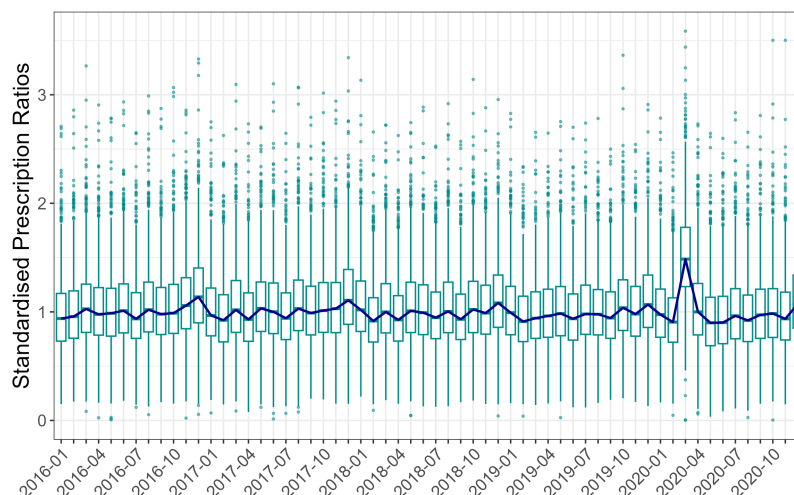


Figure 4.3: Box plots of standardised prescription ratios (SPRs) across all GP surgeries in Scotland for each month between January 2016 and December 2020.

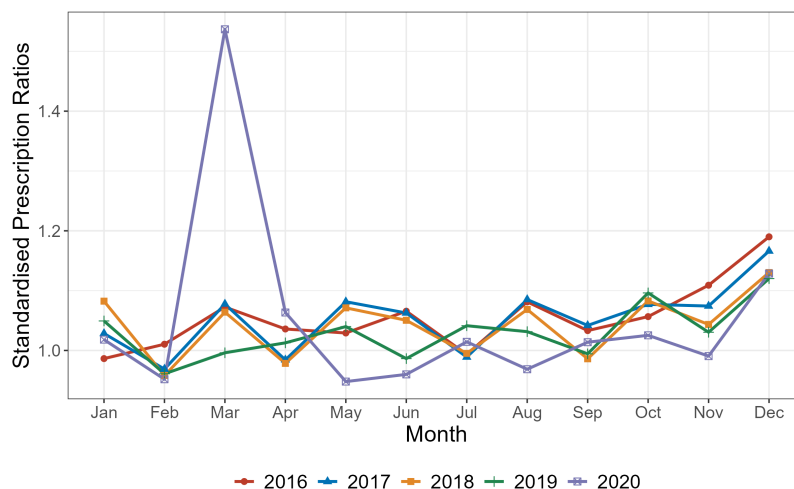


Figure 4.4: Line plot showing the monthly variation in the mean standardised prescription ratio (SPR) over all GP surgeries in Scotland, 2016–2020.

and 79  $\text{PM}_{2.5}$  sites over the 60 months) across Scotland. The covariates used in these models included numerical air quality model outputs, meteorological factors, land use, population density and distance to the nearest road. The best performing model for each pollutant was assessed by quantifying its out-of-sample predictive performance, which was achieved by splitting the measurement data into an 80% training set and a 20% test set by site. Each model was fitted to the training set (tuning parameters were optimised via a 10-fold cross validation procedure applied to this training set), before being used to predict observations in the test set. This procedure was repeated 10 times to ensure that the results were not biased by a

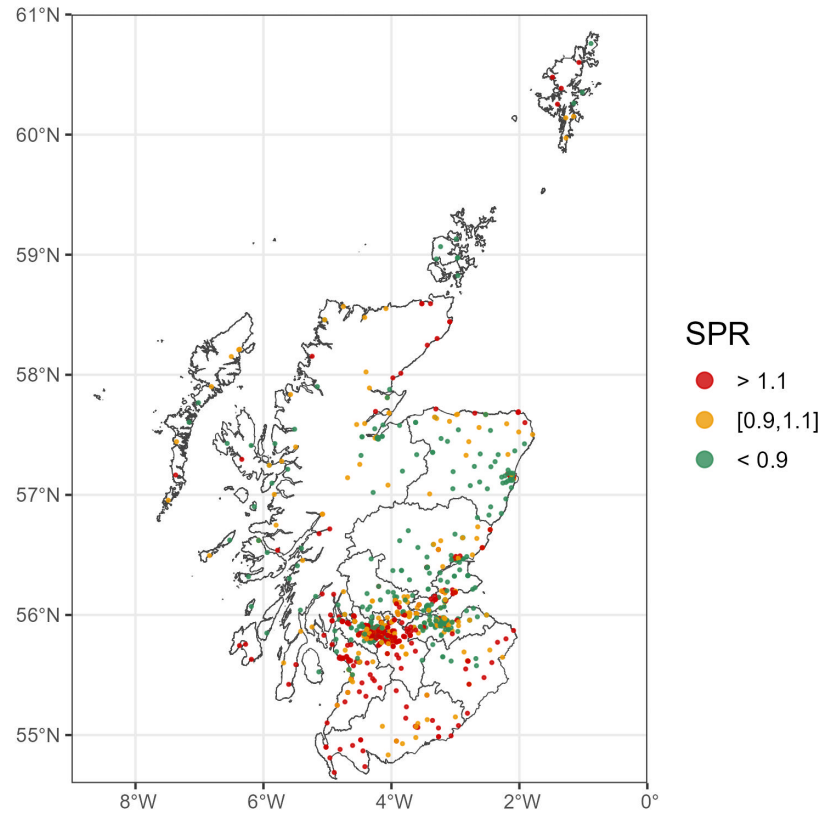


Figure 4.5: Spatial distribution of mean standardised prescription ratios (SPRs) for chronic respiratory medication prescriptions across GP surgeries in Scotland, averaged over the period 2016–2020.

single choice of training and test split.

The accuracy of the point predictions was measured by their root mean squared error (RMSE) and median absolute error (MAE), while the uncertainty was quantified using empirical coverage and average interval widths of their 95% prediction intervals. This model comparison resulted in a random forest being the best prediction algorithm for both  $\text{NO}_2$  and  $\text{PM}_{2.5}$ , while a simple linear model performed best for  $\text{PM}_{10}$ . The RMSE and MAE values from these best performing models are:  $\text{NO}_2$  – RMSE = 7.70, MAE = 4.36;  $\text{PM}_{10}$  – RMSE = 2.73, MAE = 1.44; and  $\text{PM}_{2.5}$  – RMSE = 1.15, MAE = 0.54; which compares to the original monthly average measurement data that have the following ranges:  $\text{NO}_2$  – [1.1, 76.5];  $\text{PM}_{10}$  – [2.8, 38.5];  $\text{PM}_{2.5}$  – [1.3, 23.9]. The above-mentioned best performing models were subsequently used to predict monthly average concentrations on a  $1 \text{ km}^2$  grid covering all of Scotland for all 60 months, which I use here as the basis for the air pollution exposure metrics. Further details on these predictions are given in Chapter 3.

#### 4.2.4 Socio-economic deprivation

To obtain unbiased estimates of the association between air pollution and chronic respiratory disease risk, it is essential to adjust for unmeasured confounding that may also be spatially correlated with air pollution. A key confounder in this context is socio-economic deprivation, which has been shown to strongly correlate with the risk of respiratory diseases (Lee, 2018, 2023). In Scotland, socio-economic deprivation is quantified using the Scottish Index of Multiple Deprivation (SIMD, <https://www.gov.scot/collections/scottish-index-of-multiple-deprivation-2020/>). The SIMD is constructed for 6,976 small non-overlapping areal units called Data Zones and includes indicators across seven domains such as income, employment, education, health, access to services, crime, and housing. In this study, I use the 2020 SIMD indicators, which are the closest available to the end of this study period and are assumed to suitably represent socio-economic conditions over 2016–2020. The SIMD indicators are temporally invariant, meaning that they remain constant over the 60-month study period, but they vary spatially across the DZs.

To avoid circular reasoning, the health domain is excluded, since using health-related indicators to model health outcomes could bias the estimated associations. I initially retained indicators from the remaining six domains, as summarised in Table 4.3. Note that the SIMD provides both count-based and rate-based indicators for similar quantities. To simplify the analysis and focus on interpretable, population-adjusted measures, all count-type indicators are excluded in favour of their corresponding percentage-, proportion- or rate-type equivalents.

Figure 4.6 presents histograms and kernel density estimates of the SIMD indicators at the DZ scale, highlighting that several indicators exhibit skewness and contain outliers. In particular, the education indicators *EduAttend* and *EduAttain* show bimodal distributions with a cluster of zeros. The minimum non-zero values are 0.4 for *EduAttend* and 2.890 for *EduAttain*, suggesting that zero values may not represent truly low education levels but instead reflect the absence of pupils or education records in some Data Zones. To avoid interpreting these zeros as extremely low education attainment, I replaced zeros in these two indicators with the corresponding minimum non-zero value, which represents the lowest observed level among Data Zones with valid data.

Figure 4.7 presents a correlation plot of the 20 SIMD indicators, where some pairs are highly correlated. For instance, *IncRate*, *EmpRate* and *EduNoQuals* have Pearson correlation coefficients exceeding 0.8, and indicators within the “access to services”

Table 4.3: Summary of the indicators included in the 2020 Scottish Index of Multiple Deprivation (SIMD) that are used in this study.

| Domain            | Description                                                                                                                                | Type               | Label      |
|-------------------|--------------------------------------------------------------------------------------------------------------------------------------------|--------------------|------------|
| <b>Income</b>     | Percentage of people who are income deprived.                                                                                              | Percentage         | IncRate    |
| <b>Employment</b> | Percentage of working age people who are employment deprived.                                                                              | Percentage         | EmpRate    |
| <b>Education</b>  | School pupil attendance.                                                                                                                   | Percentage         | EduAttend  |
|                   | Attainment of school leavers.                                                                                                              | Score              | EduAttain  |
|                   | Working age people with no qualifications.                                                                                                 | Standardised ratio | EduNoQuals |
|                   | Proportion of people aged 16-19 not participating in education, employment or training.                                                    | Proportion         | EduPartici |
|                   | Proportion of 17-21 year olds entering university.                                                                                         | Proportion         | EduUniver  |
| <b>Access</b>     | Average drive time to a petrol station.                                                                                                    | Time (minutes)     | GAccPetrol |
|                   | Average drive time to a GP surgery.                                                                                                        | Time (minutes)     | GAccDTGP   |
|                   | Average drive time to a post office.                                                                                                       | Time (minutes)     | GACCDTPost |
|                   | Average drive time to a primary school.                                                                                                    | Time (minutes)     | GACCDTPsch |
|                   | Average drive time to a retail centre.                                                                                                     | Time (minutes)     | GAccDTRet  |
|                   | Average drive time to a secondary school.                                                                                                  | Time (minutes)     | GAccDTSsch |
|                   | Public transport travel time to a GP surgery.                                                                                              | Time (minutes)     | GAccPTGP   |
|                   | Public transport travel time to a post office.                                                                                             | Time (minutes)     | GAccPTPost |
|                   | Public transport travel time to a retail centre.                                                                                           | Time (minutes)     | GAccPTRet  |
|                   | Percentage of premises without access to superfast broadband (at least 30Mb/s download speed).                                             | Percentage         | GAccBrdbnd |
| <b>Crime</b>      | Recorded crimes of violence, sexual offences, domestic housebreaking, vandalism, drugs offences, and common assault per 10,000 population. | Rate               | CrimeRate  |
| <b>Housing</b>    | Percentage of people in overcrowded households.                                                                                            | Percentage         | HouseOCrat |
|                   | Percentage of people in households without heating.                                                                                        | Percentage         | HouseNCrat |

domain are also strongly correlated with one another. Such multicollinearity can lead to biased regression coefficients and inflated uncertainty in the estimated effects. To address this, I applied a variance inflation factor (VIF)-based selection procedure, which is shown in Section 4.4. Note that the SIMD indicators are at the DZ scale, which will be spatially re-aligned to the GP population catchment scale using the approach described in Section 4.3.

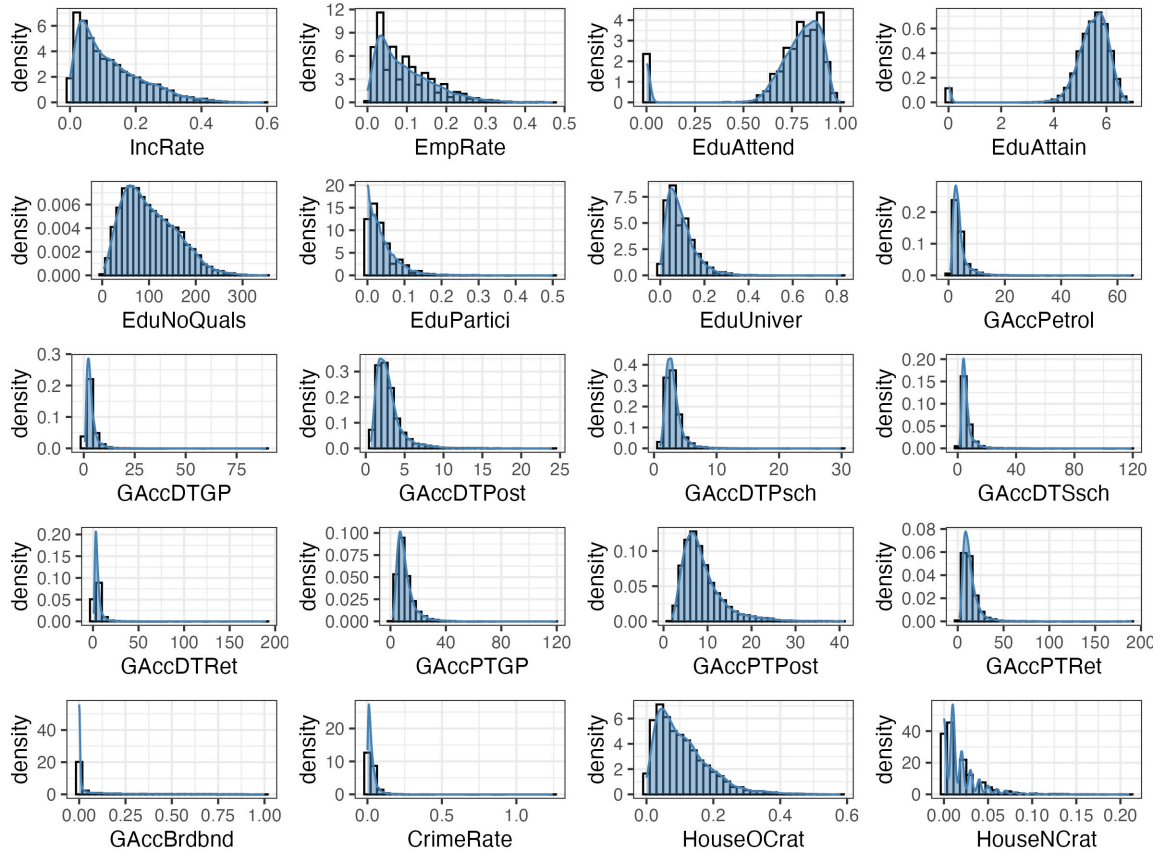


Figure 4.6: Histograms and density plots of the SIMD indicators at the DZ scale.

#### 4.2.5 Meteorological data

Epidemiological studies have consistently found that both extreme heat and cold are associated with increased respiratory morbidity and mortality in asthma and COPD cases (Lin et al., 2009; Hansel et al., 2016). Inhalation of hot, humid air can lead to bronchoconstriction, directly affecting the respiratory system (Hayes Jr et al., 2012). Cold temperatures also cause bronchoconstriction, inflammation, and damage to lung function (D'Amato et al., 2018), and induce mucous hypersecretion, which has been proposed as a mediator in respiratory disease exacerbations (Li et al., 2011). Although humidity is not typically considered a primary health risk factor, high humidity can promote mould growth, producing spores that, when inhaled, can trigger respiratory symptoms such as coughing and wheezing, particularly in individuals with mould allergies (Kosmidis et al., 2023). This may exacerbate asthma and COPD symptoms and increase the risk of disease exacerbations (Mu et al., 2017; Caillaud et al., 2018).

Motivated by this evidence, I included temperature and humidity as meteorologi-

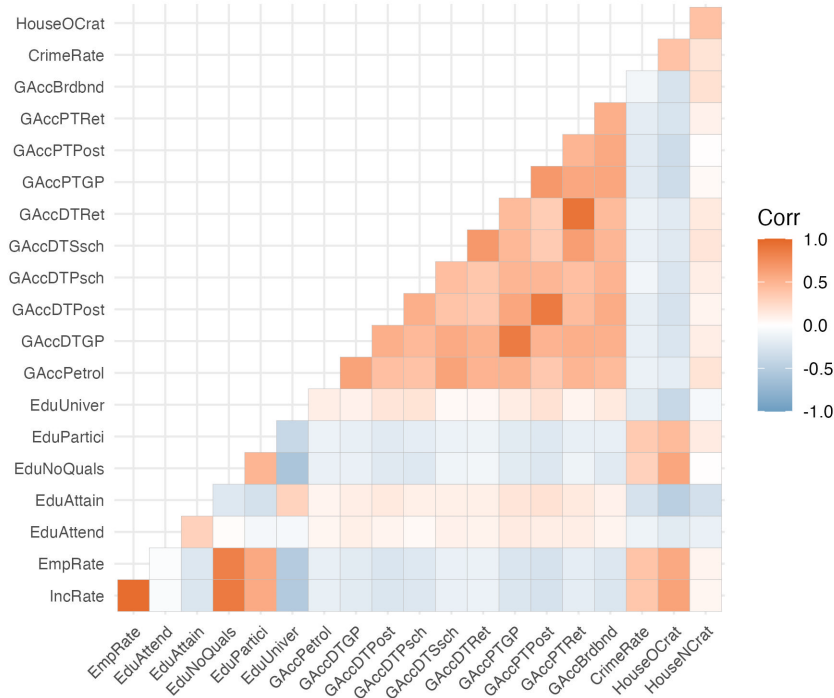


Figure 4.7: Correlation plot of the SIMD indicators at the DZ scale.

cal confounders in the health model. These data come from the same source as the meteorological covariates used in the air pollution modelling in Chapter 3, namely the HadUK-Grid dataset. Monthly averages are provided on a 1 km<sup>2</sup> grid covering all of Scotland. As with the air pollution predictions, these gridded meteorological data are first aggregated to the DZ scale and then realigned to the GP scale using the procedure described in Section 4.3.

#### 4.2.6 Ethnicity

Differences in the prevalence and severity of asthma and COPD across ethnic groups have been documented, with evidence that white populations tend to have a lower risk of asthma but a higher risk of COPD compared with some minority ethnic groups (Tran et al., 2011; Gilkes et al., 2017).

In this study, ethnicity data are sourced from the 2022 Scottish census (the closest year of data available). For each Data Zone, the ethnicity data include the percentage of the population belonging to the following ethnic groups, as recorded in the census: (i) White; (ii) South Asian (Indian/Pakistani/Bangladeshi) descent; (iii) Chinese descent; and (iv) Black. Figure 4.8 shows the correlations among these four proportions. The percentage of the White population is highly negatively

correlated with the percentages of each of the other ethnic groups. In contrast, the correlations among the other three ethnic groups (Chinese, South Asian, and Black) are relatively lower. Because the four proportions are almost linearly dependent (they sum to nearly one in most Data Zones), I excluded the White proportion and instead included the percentages of the Chinese, South Asian and Black populations as ethnicity confounders. The omitted White group then serves as the reference category.

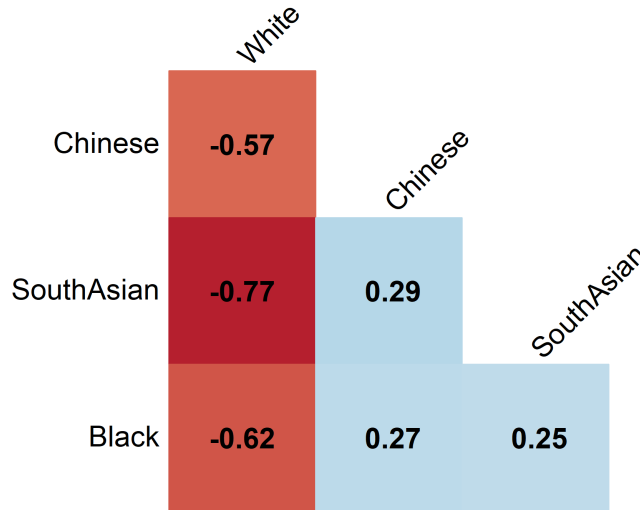


Figure 4.8: Correlation plot of the ethnic groups: White, South Asian descent, Chinese descent, and Black at the DZ scale.

### 4.3 Spatial realignment

The study consists of data for  $t = 1, \dots, T = 60$  consecutive months between January 2016 and December 2020, and collectively the prescription, air pollution and covariate data relate to three different spatial scales. Prescription data were recorded at  $k = 1, \dots, K = 1,006$  GP surgeries, but recall that not all surgeries were operational for all 60 months, leading to different numbers of spatial data points for each month. The  $k$ -th GP surgery is located at  $\mathbf{s}_k$ , while the spatial extent of its patient catchment area is denoted by  $\mathcal{S}_k$ . Patients do not necessarily attend their nearest GP surgery, which means that the set of patient catchment areas  $\{\mathcal{S}_1, \dots, \mathcal{S}_K\}$  constitutes a set of overlapping areal units. In addition,  $\{\mathcal{S}_k\}$  are partially unknown, because the only information available to define them are their population intersections with each

Data Zone. Here, I denote the  $m = 1, \dots, M = 6,976$  Data Zones by  $\{\mathcal{A}_1, \dots, \mathcal{A}_M\}$ . The specific population intersection data provided by Public Health Scotland are denoted by  $\{p_t(\mathcal{A}_m \cap \mathcal{S}_k)\}$ , which quantifies the number of people who live in Data Zone  $\mathcal{A}_m$  and are registered at the  $k$ -th GP surgery in month  $t$ . Note, each person can only attend one GP surgery. Thus, while these data provide some information on the spatial extents of the GP surgery catchment areas, they do not give an exact geographical boundary as it is unknown where in each Data Zone each patient lives.

The SIMD and ethnicity confounders are available at the DZ scale, while the air pollution concentrations and the meteorological confounders are available at a third spatial scale, namely the set of  $I = 4,708,065$   $1 \text{ km}^2$  grid squares across Scotland denoted here by  $\{\mathcal{H}_1, \dots, \mathcal{H}_I\}$ . As these grid squares and the Data Zones have fixed geographical boundaries, I computed their areas of intersection, which for grid square  $\mathcal{H}_i$  and Data Zone  $\mathcal{A}_m$  is denoted by  $a(\mathcal{H}_i \cap \mathcal{A}_m)$ . Therefore, the first statistical challenge I face when modelling these data is the *spatial change of support problem*, as I need all the variables to be on a common spatial scale. The spatial scale used for inference is the GP population catchment scale, because that is the scale at which the disease count data are available. The realignment proceeds in two steps, following the flow from the finest scale to the inference scale. The gridded variables are first aggregated to the DZ scale, and all DZ-level covariates are then realigned to the GP catchment scale. First, the gridded pollution exposures and meteorological confounders were re-scaled to the DZ scale using area-weighted averaging via

$$x_t(\mathcal{A}_m) = \sum_{i=1}^I \frac{a(\mathcal{H}_i \cap \mathcal{A}_m)}{\sum_{j=1}^I a(\mathcal{H}_j \cap \mathcal{A}_m)} x_t(\mathcal{H}_i).$$

Second, the covariates at the DZ scale  $\{x_t(\mathcal{A}_m)\}$ , comprising the SIMD and ethnicity confounders together with the spatially aggregated pollution and meteorological variables, were realigned to the GP scale  $\{x_t(\mathcal{S}_k)\}$  by population weighted averaging, i.e., by computing

$$x_t(\mathcal{S}_k) = \sum_{m=1}^M \frac{p_t(\mathcal{A}_m \cap \mathcal{S}_k)}{\sum_{r=1}^M p_t(\mathcal{A}_r \cap \mathcal{S}_k)} x_t(\mathcal{A}_m).$$

Note, the population intersection data  $\{p_t(\mathcal{A}_m \cap \mathcal{S}_k)\}$  are only available on an annual basis, so I assume that these intersections apply for all months within the year.



## 4.4 Exploratory analysis using a negative binomial model

In this study, the response variables consist of non-negative spatially aggregated prescription counts at the GP surgery level, thus I first fitted Poisson log-linear models. However, the diagnostics revealed notable overdispersion, indicating that the variance of the counts exceeds the mean, which is a common challenge in count data analysis. This violation of the Poisson mean-variance equality can lead to underestimated standard errors and biased inference if not explicitly addressed. Overdispersion in this context is likely to arise from unobserved heterogeneity between surgeries and over time, as well as residual spatial and temporal dependence that is not captured by a simple Poisson likelihood structure.

To address this, I therefore considered two commonly used alternatives: quasi-Poisson and negative binomial regression, which are widely available in statistical software. Both approaches extend the Poisson framework by introducing a second parameter to relax the restriction of the mean-variance relationship, but they differ in how the variance depends on the mean. In the quasi-Poisson model, the variance is specified as a scalar multiple of the mean, whereas in the negative binomial model the variance is a quadratic function of the mean (Ver Hoef and Boveng, 2007; Lindén and Mäntyniemi, 2011). There is no universal rule as to which approach is preferable, so in this setting I fitted both quasi-Poisson and negative binomial models to the prescription data and compared their performance by examining the residual plots (standardised residuals versus fitted values). The results indicated that the negative binomial models produced a more stable residual scatter distribution, suggesting a better overall fit to the data. Consequently, I adopted the overdispersed negative binomial model as the baseline framework for the exploratory analyses and models.

### 4.4.1 Data likelihood model

Recall that for each surgery  $s_k$ ,  $Y_t(\mathcal{S}_k)$  denotes the observed number of prescriptions in month  $t$ , and  $e_t(\mathcal{S}_k)$  denotes the indirectly standardised expected count, which accounts for differences in population size and age-sex structure as described in Section 4.2.2. The model is specified as

$$Y_t(\mathcal{S}_k) \sim \text{NegBin}(\mu_t(\mathcal{S}_k), \xi), \quad k \in \mathbb{K}^{(t)}, \quad t = 1, \dots, T,$$

with the quadratic mean-variance relationship

$$\begin{aligned}\mathbb{E}(Y_t(\mathcal{S}_k)) &= \mu_t(\mathcal{S}_k) = e_t(\mathcal{S}_k)\theta_t(\mathcal{S}_k), \\ \text{Var}(Y_t(\mathcal{S}_k)) &= \mu_t(\mathcal{S}_k) + \xi\mu_t^2(\mathcal{S}_k), \quad \xi > 0,\end{aligned}$$

where  $\xi$  is the overdispersion parameter. The overdispersion factor  $1 + \xi\mu_t(\mathcal{S}_k)$ , the amount in excess of  $\mu_t(\mathcal{S}_k)$ , represents the multiplicative inflation of the variance relative to the Poisson model.

The relative risk of prescribing  $\theta_t(\mathcal{S}_k)$  is modelled on the log scale as a function of a single pollutant and potential confounders:

$$\ln(\theta_t(\mathcal{S}_k)) = \mathbf{z}_t(\mathcal{S}_k)^\top \boldsymbol{\delta} + x_t(\mathcal{S}_k)\beta + f(t),$$

where  $\mathbf{z}_t(\mathcal{S}_k)$  is a  $p \times 1$  vector of socio-economic, ethnic and meteorological confounders, and  $x_t(\mathcal{S}_k)$  denotes the exposure to a single pollutant (NO<sub>2</sub>, PM<sub>10</sub> or PM<sub>2.5</sub>), realigned to the GP scale as described in Section 4.3. Thus  $(\boldsymbol{\delta}, \beta)$  respectively represent the effects of the confounders and air pollutant concentrations on respiratory prescribing rates, which are estimated by the maximum likelihood estimation (MLE) method. The function  $f(t)$  denotes a set of temporal trend terms, including indicator variables for year (five levels) and month-of-year (12 levels), and an additional indicator for March 2020. This specification adjusts for year-to-year temporal variation and seasonality. It explicitly allows for the anomalous spike in prescriptions observed in the first month of the Covid-19 lockdown (see Figure 4.3), which was likely due to people stocking up on medications in the run-up to the national lockdown introduced on 24 March 2020. Note, the data do not support the addition of a longer-term Covid effect indicator for the months succeeding March 2020, which is probably due to the fact that after stocking up in March there was no need to stock up on further medications again in subsequent months.

#### 4.4.2 Selecting the confounders

As described in Section 4.2.2, two main prescription types are considered in this study: bronchodilators (Reliever) and inhaled corticosteroids (Preventer). Therefore, I will fit negative binomial models for each of these medication types, as well as a combined category (Both) obtained by summing reliever and preventer prescriptions. In addition, to assess potential impacts of the Covid-19 pandemic, each analysis is carried out for two time periods: (i) the full study period (2016–2020); and (ii) the

pre-pandemic period (2016–2019). This yields six datasets in total (three medication outcomes  $\times$  two time periods). For each dataset, the initial set of confounders included the SIMD indicators, ethnicity proportions, and meteorological variables (temperature and humidity).

Because some of these confounding variables are correlated, there is a risk of multicollinearity. Therefore, I applied a variance inflation factor (VIF)-based procedure (see Section 2.1.3) applied to negative binomial models that included only the offset and the full set of candidate confounders, but no pollutant and no temporal trend terms. This was done separately for the two time periods considered in the analysis. Because the VIF depends only on the design matrix of covariates and not on the response, this procedure yielded the same set of selected confounders for all three prescription outcomes within a given time period.

Specifically, starting from the full confounder set, I identified covariates with VIF greater than 5, which is a commonly used threshold indicating problematic multicollinearity. I then sequentially removed the confounder with the largest VIF above this threshold, re-computed the VIFs, and repeated this process until all remaining confounders had  $\text{VIF} \leq 5$ . In practice, this led to the removal of six confounders in all cases. The confounders removed included those summarising employment deprivation and no school qualifications (correlated with income deprivation), housing overcrowding (correlated with no central heating in housing) and public transport travel times to a post office, GP surgery and retail centre (correlated with the drive time equivalents).

#### 4.4.3 Measuring residual spatial correlation

To assess how well the negative binomial models capture spatial dependence in the prescription data, I examined spatial correlation in the model residuals  $\{r_{1t}, \dots, r_{Kt}\}$  for each month  $t$  using Moran's I statistic (Moran, 1950), which, based on a generic neighbourhood matrix  $\mathbf{W} = \{w_{kj}\}$ , is given by

$$I_t = \frac{K \sum_{k=1}^K \sum_{j=1}^K w_{kj} (r_{kt} - \bar{r}_t)(r_{jt} - \bar{r}_t)}{\left( \sum_{k=1}^K \sum_{j=1}^K w_{kj} \right) \sum_{k=1}^K (r_{kt} - \bar{r}_t)^2} \in [-1, 1],$$

where  $\bar{r}_t = (1/K) \sum_{k=1}^K r_{kt}$ . A value of zero corresponds to spatial independence, while values respectively above and below zero correspond to positive and negative spatial autocorrelation. To assess the statistical significance of  $I_t$ , I use a permutation-based

hypothesis test (based on 10,000 random permutations of the residuals), where the null hypothesis is  $H_0 : I_t = 0$ . Depending on the choice of alternative, I consider both a one-sided test  $H_1 : I_t > 0$  (for positive spatial autocorrelation) and a two-sided test  $H_1 : I_t \neq 0$  (for either positive or negative autocorrelation). Further details on Moran's I and the associated testing procedure are provided in Chapter 2. Here I summarise residual spatial dependence using three metrics:

- **Positive Spatial Autocorrelation Indicator (PSAI)** – The percentage of the  $T$  months that have statistically significant Moran's I values at the 5% level based on  $H_1 : I_t > 0$ .
- **General Spatial Autocorrelation Indicator (GSAI)** – The percentage of the  $T$  months that have statistically significant Moran's I values at the 5% level based on  $H_1 : I_t \neq 0$ .
- **Mean Absolute Moran's I (MAI)** – The mean absolute value of Moran's I over all  $T$  months computed by  $MAI = \frac{1}{T} \sum_{t=1}^T |I_t|$ .

The PSAI and GSAI measure how often the residuals exhibit statistically significant spatial autocorrelation. If a model can adequately capture the spatial structure, it is expected that both PSAI and GSAI to be close to the nominal 5% level, reflecting only occasional false positives under approximate spatial independence. Chronic respiratory disease data typically exhibit positive spatial correlation due to shared environmental exposures and socio-economic conditions among neighbouring areas, but the spatial random effects smoothing model proposed in Section 4.5 can in some cases over-smooth the spatial field and induce negative spatial autocorrelation in the residuals. In such situations, PSAI may be low, while GSAI remains elevated because the two-sided test also detects significant negative autocorrelation.

The MAI complements PSAI and GSAI by quantifying the overall magnitude of spatial dependence, irrespective of its sign or statistical significance. A low MAI indicates that the residuals are close to spatially independent in most months, whereas a high MAI suggests that substantial spatial structure remains. Taken together, PSAI, GSAI and MAI provide a comprehensive diagnostic summary of residual spatial dependence and are used in this chapter to compare model specifications and, later in Section 4.6, alternative neighbourhood structures.

#### 4.4.4 Results from the negative binomial models

The negative binomial models fitted in this section provide an initial, non-spatial assessment of the association between air pollution and respiratory prescribing rates. These models assume that, after adjustment for air pollution exposure, confounders and temporal terms, the residuals are approximately independent across space and time. To evaluate this assumption, I examined residual spatial correlation using Moran's I statistic together with PSAI, GSAI and MAI defined in Section 4.4.3, and residual temporal correlation using the lag-1 autocorrelation function (ACF(1)). Here I focus on models where NO<sub>2</sub> is included as the pollutant, and the corresponding results for PM<sub>10</sub> and PM<sub>2.5</sub> are similar and are not shown for brevity.

Residual spatial correlation was assessed monthly using Moran's I statistic, where a binary neighbourhood matrix based on the 5 nearest neighbours method (Euclidean distances were calculated between GP surgery locations  $\{\mathbf{s}_k\}$ ) was used to compute the indicator. Figure 4.9 displays the resulting Moran's I time series for the three medication outcomes, for the full period (2016–2020, blue lines) and the pre-pandemic period (2016–2019, red lines), together with PSAI, GSAI and MAI summaries. For all six medication–period pairs, both PSAI and GSAI equal 100%, indicating that every month exhibits statistically significant positive and general spatial autocorrelation at the 5% level. These results confirm that strong residual spatial dependence remains after fitting the non-spatial negative binomial models, and hence that more flexible spatial models are required to adequately capture the spatial structure in the prescription data.

Residual temporal correlation was assessed using the lag-1 autocorrelation function, computed separately for each GP surgery from its time series of residuals. Figure 4.10 presents the distribution of ACF(1) values across surgeries for each medication type and time period, together with approximate 95% significance bounds under the null hypothesis of no temporal autocorrelation. The percentages in red in each panel report the proportion of surgeries with ACF(1) lying outside the bounds, and hence exhibiting statistically significant lag-1 temporal autocorrelation. Roughly one quarter of surgeries exhibit significant residual temporal dependence, indicating that the temporal term  $f(t)$  included in the negative binomial models does not fully capture the underlying temporal dynamics. In summary, the strong residual spatial correlation and the existing residual temporal correlation motivate the Bayesian spatio-temporal smoothing models developed in Section 4.5, which incorporate explicitly structured spatial and temporal random effects.

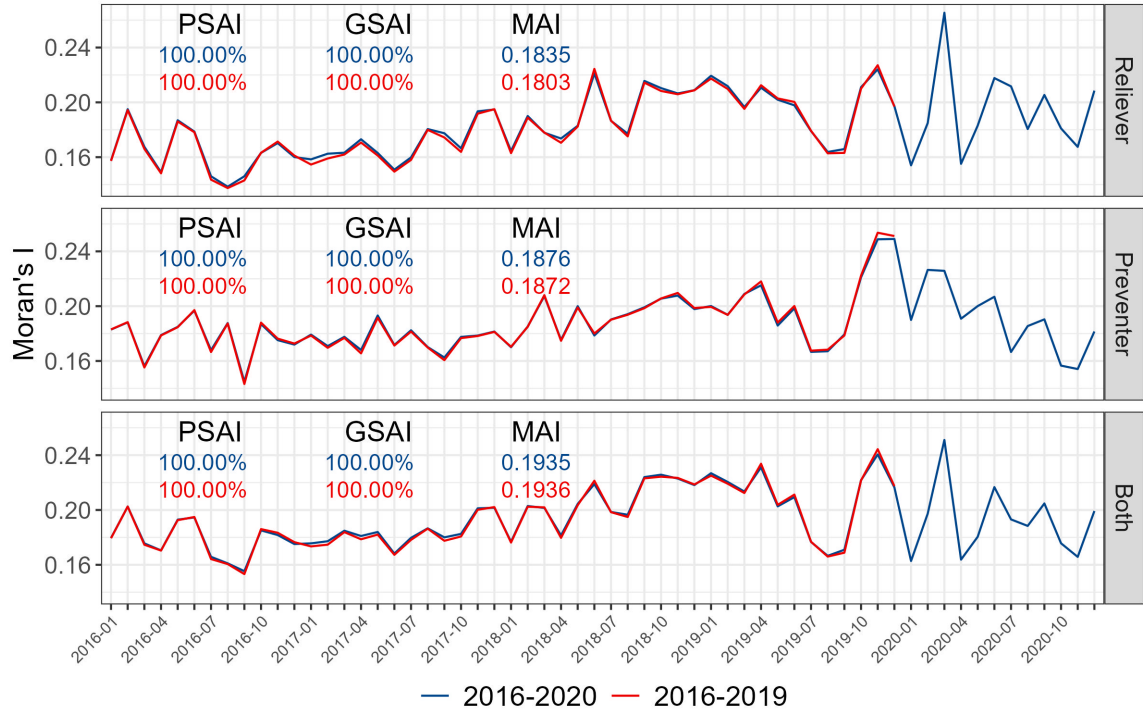


Figure 4.9: Moran's I values for the standardised residuals from the negative binomial models with  $\text{NO}_2$  as the pollutant, by medication type (rows) and time period (blue: 2016–2020; red: 2016–2019). A 5-nearest-neighbour spatial weights matrix based on GP surgery locations is used to compute Moran's I.

## 4.5 Spatio-temporal smoothing model with the CAR-AR(1) structure

In this section I propose a novel Bayesian hierarchical spatio-temporal model for quantifying the effects of air pollution on GP prescribing rates, where the novel contribution concerns how to construct a spatial neighbourhood matrix for capturing the spatial closeness between the patient populations registered at each pair of GP surgeries. Inference for this model is based on the integrated nested Laplace approximation with a Bayes correction (Rue et al., 2009; Van Niekerk et al., 2023), implemented using the R package INLA. The model is presented in stages below, beginning with the data likelihood component.

### 4.5.1 Data likelihood model

Motivated by the exploratory analysis in Section 4.4, which showed clear spatial and temporal dependences in the model residuals, I model the prescription data using





Figure 4.10: Lag-1 autocorrelation (ACF(1)) of the negative binomial model residuals for  $\text{NO}_2$ , by medication type (rows) and time period (columns). Each point represents ACF(1) for a single GP surgery. The dashed red lines indicate approximate 95% significance bounds under the null of no temporal autocorrelation, and the red text gives the proportion of surgeries with significant ACF(1) values.

a negative binomial likelihood with a spatio-temporal random effects component to capture residual correlation in space and time. Let  $\{1, \dots, K = 1,006\}$  index the complete set of GP surgeries that appear at least once throughout the  $T = 60$  month study period, and let  $\mathbb{K}^{(t)} \subset \{1, \dots, K\}$  denote the set of surgeries that are operational in month  $t$ . The data likelihood model is given by

$$\begin{aligned}
 Y_t(\mathcal{S}_k) &\sim \text{NegBin}(e_t(\mathcal{S}_k)\theta_t(\mathcal{S}_k), \xi) \quad \text{for } k \in \mathbb{K}^{(t)} \text{ and } t = 1, \dots, T, \\
 \ln(\theta_t(\mathcal{S}_k)) &= \mathbf{z}_t(\mathcal{S}_k)^\top \boldsymbol{\delta} + x_t(\mathcal{S}_k)\beta + f(t) + \phi_t(\mathcal{S}_k).
 \end{aligned}
 \tag{4.1}$$

Here  $Y_t(\mathcal{S}_k)$  denotes the number of prescriptions dispensed by the GP surgery  $\mathbf{s}_k$  with population catchment area  $\mathcal{S}_k$  in month  $t$ , while  $e_t(\mathcal{S}_k)$  is the indirectly standardised expected count defined in Section 4.2.2, which is included as a fixed offset. The expected value of the negative binomial distribution is  $e_t(\mathcal{S}_k)\theta_t(\mathcal{S}_k)$ , where  $\theta_t(\mathcal{S}_k)$  is

the relative risk of prescribing, and  $\xi$  is the overdispersion parameter.

The term  $\mathbf{z}_t(\mathcal{S}_k)$  is a vector of confounders (socio-economic, ethnic and meteorological variables), while  $x_t(\mathcal{S}_k)$  is the pollution exposure (one pollutant for each model, i.e.  $\text{NO}_2$ ,  $\text{PM}_{10}$  or  $\text{PM}_{2.5}$ ). The function  $f(t)$  denotes a temporal effect that is common to all GP surgeries and captures long-term trends and seasonality. In line with the exploratory analysis,  $f(t)$  is specified using factor terms for year (5 levels) and month-of-year (12 levels), together with an additional indicator for March 2020 to account for the spike in prescribing observed at the start of the Covid-19 lockdown. The remaining unobserved space-time variation is modelled by the spatio-temporal random effect  $\phi_t(\mathcal{S}_k)$  described in Section 4.5.2.

The regression coefficients for the confounders and pollution exposure ( $\delta, \beta$ ) are assigned weakly informative independent Gaussian priors with mean 0 and precision 0.001, so that the data primarily dominate their values. In addition,  $\log(\xi)$  is assigned a penalised complexity Gamma prior distribution (`pc.mgamma`, default in the INLA software), which is again chosen to be weakly informative and hence let the data inform the level of overdispersion.

### 4.5.2 Spatio-temporal random effects model

In the data likelihood Equation (4.1),  $\{\phi_t(\mathcal{S}_k)\}$  represents spatio-temporally correlated random effects that account for residual spatial correlation and temporal dependence not explained by the fixed effects. A range of spatial and spatio-temporal prior structures have been proposed for such random effects, and here I use the multivariate first-order autoregressive (AR(1)) process with a CAR spatial structure proposed by [Rushworth et al. \(2014\)](#). The structure is given by

$$\begin{aligned}\boldsymbol{\phi}_1 &\sim \text{N}(\mathbf{0}, [\tau\mathbf{Q}(\mathbf{W}, \rho)]^{-1}), \\ \boldsymbol{\phi}_t | \boldsymbol{\phi}_{t-1} &\sim \text{N}(\alpha\boldsymbol{\phi}_{t-1}, [\tau\mathbf{Q}(\mathbf{W}, \rho)]^{-1}) \quad \text{for } t = 2, \dots, T,\end{aligned}\tag{4.2}$$

where  $\boldsymbol{\phi}_t = (\phi_t(\mathcal{S}_1), \dots, \phi_t(\mathcal{S}_K))^\top$  denotes the complete vector of spatial random effects for all  $K$  GP surgeries in month  $t$ . Because the precision matrix  $\mathbf{Q}(\mathbf{W}, \rho)$  is time invariant, the length of  $\boldsymbol{\phi}_t$  must be the same for all months. For any month in which surgery  $\mathcal{S}_k$  is inactive, the observation  $Y_t(\mathcal{S}_k)$  is treated as missing in the Equation (4.1), but the corresponding random effect  $\phi_t(\mathcal{S}_k)$  is still included in  $\boldsymbol{\phi}_t$ , maintaining a consistent spatial random effects structure over time.

Temporal correlation is introduced via the autoregressive mean term  $\alpha\boldsymbol{\phi}_{t-1}$ , so that  $\alpha$  controls the dependence between consecutive months. Here  $\alpha = 0$  corresponds to



temporal independence, while  $|\alpha|$  close to one indicates strong temporal dependence. Spatial correlation is introduced through the covariance matrix  $[\tau \mathbf{Q}(\mathbf{W}, \rho)]^{-1}$ , where  $\tau$  is a precision parameter governing the marginal variability of the random effects and  $\mathbf{Q}(\mathbf{W}, \rho)$  is the Leroux CAR precision matrix (Leroux et al., 2000) defined as

$$\mathbf{Q}(\mathbf{W}, \rho) = \rho(\text{diag}(\mathbf{W}\mathbf{1}) - \mathbf{W}) + (1 - \rho)\mathbf{I}, \quad (4.3)$$

where  $\mathbf{1}$  denotes a  $K \times 1$  vector of ones,  $\mathbf{I}$  is the  $K \times K$  identity matrix, and  $\mathbf{W}$  is a spatial neighbourhood matrix whose entries encode the spatial closeness between GP surgeries (defined in Section 4.5.3).

The spatial dependence parameter  $\rho \in [0, 1]$  determines the overall level of spatial smoothing. When  $\rho = 0$  the random effects are independent across surgeries, whereas  $\rho$  close to one induces strong spatial autocorrelation. The spatial closeness between each pair of GP surgeries is captured by the non-negative  $K \times K$  neighbourhood matrix  $\mathbf{W}$ , where typically  $w_{kj} > 0$  if the patient populations represented by surgeries  $(S_k, S_j)$  are close together, and  $w_{kj} = 0$  otherwise ( $w_{kk} = 0 \forall k$ ). A binary specification is commonly assumed for this matrix, because it encourages sparsity in  $\mathbf{Q}(\mathbf{W}, \rho)$  and hence fast model fitting. Under either a binary or a non-binary construction, multivariate Gaussian theory shows that the partial correlations implied by the Leroux CAR model are given by

$$\text{Corr}(\phi_t(S_k), \phi_t(S_j) | \phi_t(-S_k, S_j)) = \frac{\rho w_{kj}}{\sqrt{(\rho \sum_{i=1}^K w_{ki} + 1 - \rho)(\rho \sum_{l=1}^K w_{jl} + 1 - \rho)}},$$

where  $\phi_t(-S_k, S_j)$  denotes all the random effects at time period  $t$  except  $\{\phi_t(S_k), \phi_t(S_j)\}$ . Thus, if  $w_{kj} = 0$  then  $\{\phi_t(S_k), \phi_t(S_j)\}$  are modelled as conditionally independent given the remaining random effects at time  $t$ , while if  $w_{kj} > 0$  then they are modelled as marginally correlated.

### 4.5.3 Neighbourhood matrix

In constructing the neighbourhood matrix  $\mathbf{W}$  for the CAR prior, one must choose a specification that reflects the spatial closeness between GP-registered populations. Standard approaches for areal data define  $\mathbf{W}$  either via border-sharing rule (two regions are neighbours if they share a common geographic boundary) or via a  $D$  nearest neighbours rule based on Euclidean distances between region centroids. These are not directly applicable here because GP surgery catchment areas  $\{S_k\}$

overlap with partially unknown boundaries. In addition, the underlying health risk is determined by where patients reside rather than by the locations of the surgery buildings, and patients do not necessarily attend their nearest surgery. Thus, defining neighbours purely in terms of surgery locations may misrepresent the epidemiological closeness between populations.

In this study I consider three neighbourhood matrix constructions. First, I used a conventional  $D$  nearest neighbours matrix based on surgery locations as a baseline. Second, I proposed a novel matrix, denoted by  $\mathbf{W}^{(O)}$  ('O' for overlap), that captures direct overlaps between GP-registered populations from two GP surgeries living in the same Data Zones. I also constructed another matrix, denoted by  $\mathbf{W}^{(S)}$  ('S' for spatial), that extends this idea by also incorporating overlaps in neighbouring Data Zones. These two population-based matrices aim to better represent the spatial interactions between patient populations than a purely distance-based approach. Figure 4.11 illustrates schematically how these three constructions can assign different neighbours to the same set of surgeries.

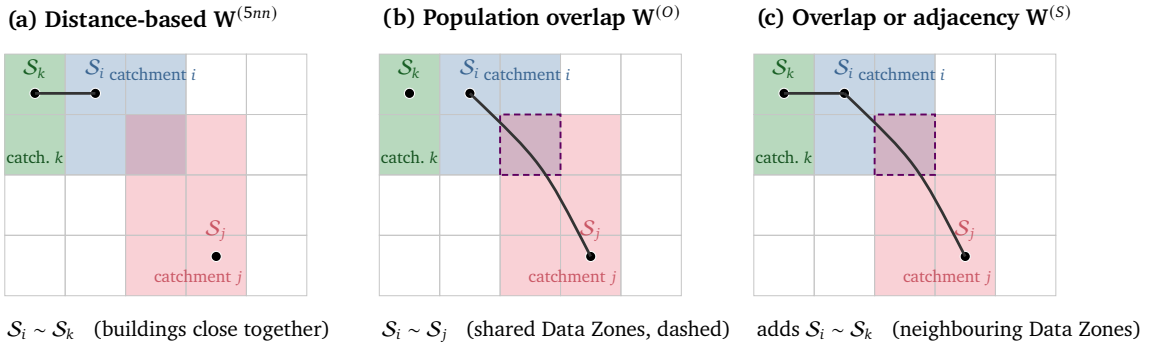


Figure 4.11: Schematic comparison of the three neighbourhood constructions. Patients live across the Data Zones that form each catchment rather than at the surgery buildings, and the catchments of  $S_i$  and  $S_j$  share the Data Zone outlined in dashed purple. The distance-based  $\mathbf{W}^{(5nn)}$  links  $S_i$  and  $S_k$  because their buildings are close together (a), whereas the population-overlap  $\mathbf{W}^{(O)}$  instead links  $S_i$  and  $S_j$  because their patient populations share Data Zones (b), and  $\mathbf{W}^{(S)}$  links both pairs since it also counts catchments lying in neighbouring Data Zones (c).

### Neighbourhood matrix based on the $D$ nearest neighbours method

In the  $D$  nearest neighbours (DNN) approach,  $w_{kj} = 1$  if the  $j$ -th surgery  $s_j$  is one of the  $D$  nearest neighbours of the  $k$ -th surgery in Euclidean distance, and  $w_{kj} = 0$  otherwise. Symmetry is then imposed by setting  $w_{jk} = 1$  whenever  $w_{kj} = 1$  but  $w_{jk} = 0$  under the initial construction. This ensures that the spatial neighbourhood matrix

$\mathbf{W}$  and hence the precision matrix  $\mathbf{Q}(\mathbf{W}, \rho)$  are symmetric, which is required by the Leroux CAR prior construction.

To yield a sparse, well-connected spatial graph, in this chapter I explored several values  $D \in \{3, 5, 10\}$ , and found that the estimated relative risks and pollutant effects were very similar across this range, demonstrating that the model results are robust to the choice of  $D$ . However, if  $D$  is chosen too small, the resulting graph can misrepresent the local structure, especially in rural areas where GP surgeries are sparse. In those regions, the closest  $D$  neighbours may still be graphically far away, so induced “neighbourhood” does not necessarily correspond to genuine local interaction, and the graph may become weakly connected. Conversely, a large value may induce excessive connection between neighbourhoods and hence over-smooth the spatial random effects. I therefore chose  $D = 5$  and denote the resulting neighbourhood matrix by  $\mathbf{W}^{(5nn)}$ , as this tends to match the average numbers of neighbours that each areal unit would have under the commonly used border sharing rule. This choice provides a reasonable balance between capturing local spatial dependence and avoiding over-smoothing. Nevertheless, this DNN construction defines spatial closeness solely in terms of surgery locations, not where patients live, which motivates the development of the population-based matrices described next.

### Neighbourhood matrix based on population intersection products

To better capture the spatial closeness among registered patient populations, I propose two novel binary specifications for  $\mathbf{W}$  denoted by  $\mathbf{W}^{(O)}$  and  $\mathbf{W}^{(S)}$ , which account for the extent to which the patient populations from two GP surgeries overlap and are spatially close together. The only data available to quantify these population overlaps are at the DZ level, and comprise population counts  $\{p_t(\mathcal{A}_m \cap \mathcal{S}_k)\}$  defined earlier. The first matrix  $\mathbf{W}^{(O)}$  (‘O’ for overlap) only considers direct population overlaps, i.e., the extent to which the patient populations from two GP surgeries live in the same Data Zones. For the GP surgeries with catchment areas  $(\mathcal{S}_k, \mathcal{S}_j)$  the  $kj$ -th element of the matrix is based on

$$w_{kj}^{*(O)} = \sum_{m=1}^M \left( \frac{1}{T_k} \sum_{t: p_t(\mathcal{S}_k) > 0} \frac{p_t(\mathcal{A}_m \cap \mathcal{S}_k)}{p_t(\mathcal{S}_k)} \times \frac{1}{T_j} \sum_{t: p_t(\mathcal{S}_j) > 0} \frac{p_t(\mathcal{A}_m \cap \mathcal{S}_j)}{p_t(\mathcal{S}_j)} \right), \quad (4.4)$$

where  $p_t(\mathcal{S}_k) = \sum_{r=1}^M p_t(\mathcal{A}_r \cap \mathcal{S}_k)$  denotes the total size of the patient population for the  $k$ th GP surgery. Here,  $T_k$  and  $T_j$  denote the number of time periods that GP surgeries  $\mathcal{S}_k$  and  $\mathcal{S}_j$  are operational. This construction first computes the product of

the (time-averaged) proportions of the two GP surgery's patient populations who live in the same Data Zone  $\mathcal{A}_m$ , before summing these products over all Data Zones in Scotland. Note, most of these time averaged products will be zero. Thus, two GP surgeries will have a higher value of  $w_{kj}^{*(O)}$  if their populations are generally located in the same Data Zones, and lower values if they are distributed across different Data Zones. In the extreme, if the patient populations from the GP surgeries located at  $(\mathbf{s}_k, \mathbf{s}_j)$  are both all located in the same Data Zone for all time periods, then  $w_{kj}^{*(O)} = 1$ . Conversely if they have no DZ-level population overlaps then  $w_{kj}^{*(O)} = 0$ .

The above construction only considers patient population overlaps in the same Data Zone, and thus our second novel specification  $\mathbf{W}^{(S)}$  ('S' for spatial) considers population overlaps in the same and spatially neighbouring Data Zones. Here, spatially neighbouring means that two Data Zones share a common border. This construction naturally extends Equation (4.4) and is given by

$$w_{kj}^{*(S)} = \sum_{m=1}^M \left( \frac{1}{T_k} \sum_{t: p_t(\mathcal{S}_k) > 0} \sum_{i \in \mathbb{R}_m} \frac{p_t(\mathcal{A}_i \cap \mathcal{S}_k)}{p_t(\mathcal{S}_k)} \times \frac{1}{T_j} \sum_{t: p_t(\mathcal{S}_j) > 0} \sum_{i \in \mathbb{R}_m} \frac{p_t(\mathcal{A}_i \cap \mathcal{S}_j)}{p_t(\mathcal{S}_j)} \right), \quad (4.5)$$

where  $\mathbb{R}_m$  denotes the set of indices that include the target Data Zone  $\mathcal{A}_m$ , along with all Data Zones that share a border with  $\mathcal{A}_m$ . Thus, population overlaps between two GP surgeries are aggregated over each individual Data Zone and its spatial neighbours, giving a more comprehensive picture of the spatial closeness between two GP surgery's patient populations. This construction is particularly important in rural areas, where Data Zones are large and sparsely populated, and where two neighbouring surgeries may share relatively few patients in exactly the same Data Zone but many living in adjacent ones.

The constructions  $(w_{kj}^{*(O)}, w_{kj}^{*(S)})$  are inherently non-binary, and also contain a large number of very small non-zero values corresponding to two GP surgeries that only have a small number of patients living in the same or neighbouring Data Zones. These very small non-zero values may not be representative of the closeness of the two surgeries in question, and will lead to a less sparse neighbourhood matrix which will increase the model fitting time. In addition, initial analyses that will be presented in Section 4.6.2 showed that a non-binary matrix performed much worse than a binary one constructed from  $(w_{kj}^{*(O)}, w_{kj}^{*(S)})$  in the motivating study (see Section 4.6.1), which is likely to be due to a number of reasons. Firstly, a non-binary matrix would lead to some of the row sums  $\sum_{j=1}^K w_{kj}^{*(O)}$  being very small (minimum = 0.0001, mean = 0.12, maximum = 0.997), and in CAR models this results in division by

very small numbers in the mean and variance of the conditional distribution of  $\{\phi_t(\mathcal{S}_k) | \phi_t(\mathcal{S}_1), \dots, \phi_t(\mathcal{S}_{k-1}), \phi_t(\mathcal{S}_{k+1}), \dots, \phi_t(\mathcal{S}_K)\}$ , which would lead to numerical instability and unreliable smoothing. Moreover, the degree of spatial smoothing would be highly uneven, because some  $(w_{kj}^{*(O)}, w_{kj}^{*(S)})$  values are close to one while others are close zero, meaning that a few large weights dominate the smoothing process which will pull some random effects strongly toward only one neighbour while effectively ignoring the others. To address these limitations I created binary neighbourhood matrices based on  $(w_{kj}^{*(O)}, w_{kj}^{*(S)})$ , and additionally applied a threshold ( $\varepsilon$  or  $\omega$ ) so that very small values are set to zero. Specifically, I define the elements in the neighbourhood matrices  $(\mathbf{W}^{(O)}, \mathbf{W}^{(S)})$  as follows:

$$w_{kj}^{(O)} = \begin{cases} 1 & \text{if } w_{kj}^{*(O)} > \varepsilon \text{ and } k \neq j \\ 0 & \text{otherwise} \end{cases}, \quad w_{kj}^{(S)} = \begin{cases} 1 & \text{if } w_{kj}^{*(S)} > \omega \text{ and } k \neq j \\ 0 & \text{otherwise} \end{cases}. \quad (4.6)$$

The thresholds considered are  $\varepsilon \in \{0.0001, 0.0005, 0.001, 0.005, 0.01\}$  for  $\mathbf{W}^{(O)}$  and  $\omega \in \{0.005, 0.01, 0.05, 0.1, 0.5\}$  for  $\mathbf{W}^{(S)}$ , which are different because the elements  $w_{kj}^{*(S)}$  are typically larger than the corresponding elements  $w_{kj}^{*(O)}$  because of the additional spatial overlaps considered in the construction of the former. Tables 4.4 and 4.5 provide summary statistics for the distribution of the numbers of neighbours (that is  $w_{kj}^{(O)} = 1$  or  $w_{kj}^{(S)} = 1$ ) that each GP surgery has at different threshold values in the neighbourhood matrices  $\mathbf{W}^{(O)}$  and  $\mathbf{W}^{(S)}$  respectively.

When  $\varepsilon = 0$  and  $\omega = 0$ , which means no threshold is applied, each GP surgery has on average over 100 spatial connections with other GP surgeries. This suggests a highly interconnected network where a substantial proportion of GPs are considered neighbours of each other, which goes against the notion of a neighbourhood matrix capturing local spatial proximity. This high connectivity may lead to over-smoothing, given the induced correlations from the CAR model between GP surgeries that are actually far apart. On the other hand, the maximum threshold values of  $\{\varepsilon, \omega\}$  result in only 3 or 4 connections on average for each GP surgery, meaning that some spatially close GP surgeries will be treated as conditionally independent in the modelling. Intermediate values of  $\{\varepsilon, \omega\}$  thus seem to provide a more reasonable compromise between connectivity and sparsity, but as discussed before, the optimal values will be chosen as the ones that minimise the level of residual spatial correlation.

Table 4.4: Summary statistics for the number of non-zero weights in the neighbourhood matrix  $\mathbf{W}^{(O)}$  at different threshold values of  $\varepsilon$  for the data spans 2016–2020 and 2016–2019.

| $\varepsilon$ | 2016–2020 |        |        |       |       |      | 2016–2019 |        |        |       |       |      |
|---------------|-----------|--------|--------|-------|-------|------|-----------|--------|--------|-------|-------|------|
|               | 0         | 0.0001 | 0.0005 | 0.001 | 0.005 | 0.01 | 0         | 0.0001 | 0.0005 | 0.001 | 0.005 | 0.01 |
| Min.          | 5         | 1      | 1      | 1     | 1     | 1    | 3         | 1      | 1      | 1     | 1     | 1    |
| 1st Qu.       | 63        | 10     | 6      | 5     | 3     | 2    | 57        | 10     | 6      | 5     | 2     | 1    |
| Median        | 110       | 17     | 12     | 10    | 5     | 3    | 101       | 17     | 12     | 10    | 5     | 3    |
| Mean          | 135       | 27     | 17     | 13    | 6     | 3    | 125       | 27     | 17     | 13    | 6     | 3    |
| 3rd Qu.       | 199       | 30     | 23     | 18    | 9     | 5    | 183       | 31     | 23     | 18    | 8     | 5    |
| Max           | 419       | 159    | 85     | 61    | 20    | 13   | 384       | 157    | 84     | 63    | 20    | 13   |

Table 4.5: Summary statistics for the number of neighbours for each GP surgery in the binary matrix  $\mathbf{W}^{(S)}$  at different threshold values of  $\omega$  for the data spans 2016–2020 and 2016–2019.

| $\omega$ | 2016–2020 |       |      |      |     |     | 2016–2019 |       |      |      |     |     |
|----------|-----------|-------|------|------|-----|-----|-----------|-------|------|------|-----|-----|
|          | 0         | 0.005 | 0.01 | 0.05 | 0.1 | 0.5 | 0         | 0.005 | 0.01 | 0.05 | 0.1 | 0.5 |
| Min.     | 6         | 1     | 1    | 1    | 1   | 1   | 3         | 1     | 1    | 1    | 1   | 1   |
| 1st Qu.  | 137       | 23    | 19   | 12   | 9   | 2   | 123       | 22    | 19   | 12   | 9   | 2   |
| Median   | 241       | 33    | 28   | 18   | 13  | 4   | 214       | 32    | 28   | 18   | 13  | 4   |
| Mean     | 245       | 42    | 35   | 20   | 15  | 4   | 225       | 41    | 35   | 20   | 15  | 4   |
| 3rd Qu.  | 356       | 56    | 48   | 26   | 20  | 6   | 332       | 56    | 47   | 26   | 20  | 6   |
| Max      | 613       | 177   | 139  | 67   | 44  | 16  | 580       | 176   | 138  | 67   | 44  | 16  |

#### 4.5.4 Selecting priors for the spatio-temporal CAR-AR(1) component

In this study, the spatio-temporal CAR-AR(1) model is built in a Bayesian framework and hence I need to specify priors for the spatio-temporal random effects  $\{\phi_t\}$  in Equation (4.2). A half-normal prior is specified for the precision parameter  $\tau$  on the standard deviation scale (i.e., for  $\sigma = 1/\sqrt{\tau}$ ) following the recommendation from [Gelman \(2006\)](#). This reparameterisation is more intuitive, and the half-normal distribution is constrained to positive values, aligning with the physical interpretation that  $\sigma$  cannot be negative. Specifically, the probability density function of  $\sigma$  is given by

$$\sigma \sim \text{HN}(0, \sigma_0^2), \quad f_\sigma(\sigma|\sigma_0) = \frac{\sqrt{2}}{\sigma_0\sqrt{\pi}} \exp\left(-\frac{\sigma^2}{2\sigma_0^2}\right).$$

Here, larger values of  $\sigma$  allow greater uncertainty and hence variation in the spatio-temporal random effects, leading to greater variation in the disease risks  $\{\theta_t(S_k)\}$ .

There is no universally “correct” choice for  $\sigma_0$ , but extreme spatial variability is uncommon for disease risk on the natural log linear predictor scale. In this study  $\sigma_0$  is set to 1.5, which offers flexibility to capture variation in log-relative risks while avoiding implausibly high values. As shown in Figure 4.12 (a), the prior density for  $\sigma$  drops sharply beyond 5, reflecting a low probability of extreme spatial variability.

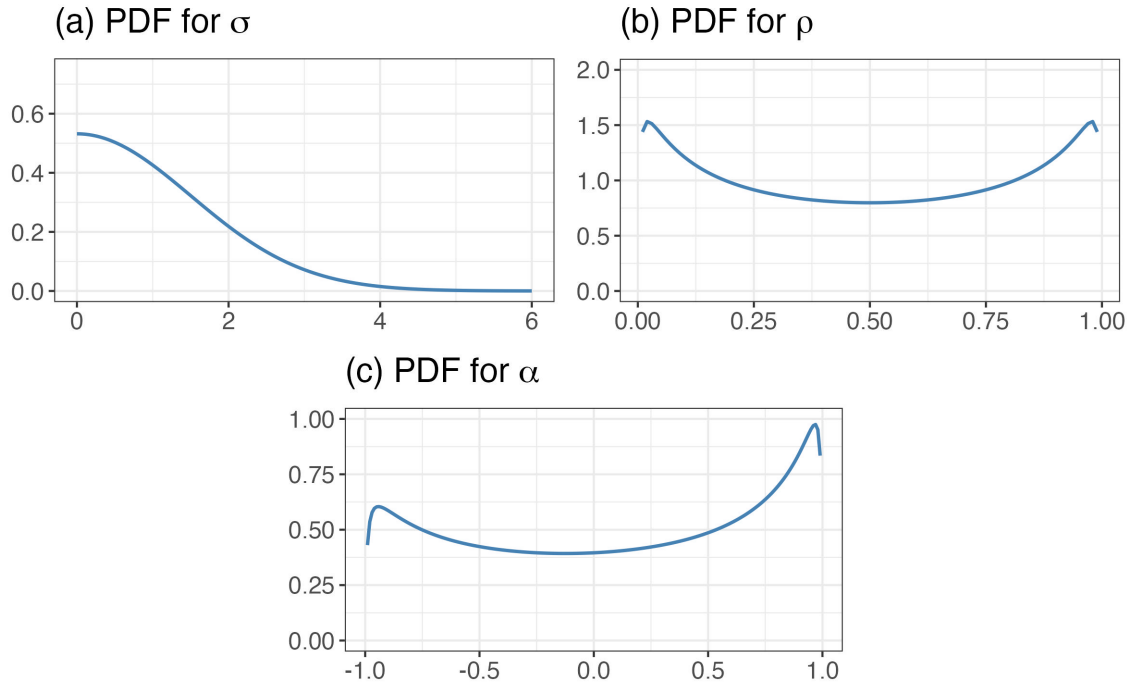


Figure 4.12: Probability density function (PDF) curves for the prior distributions assumed for the following parameters: (a) the standard deviation  $\sigma$ ; (b) the spatial dependence  $\rho$ ; and (c) the temporal correlation  $\alpha$ .

Within the INLA software, a prior is required on the log precision scale, that is,  $\theta = \log(\tau)$ . The pdf of  $\tau$  induced by the half-normal distribution of  $\sigma$  is thus given by

$$\begin{aligned}
 f_{\tau}(\tau) &= f_{\sigma}\left(\frac{1}{\sqrt{\tau}}\right) \left| \frac{\partial \sigma}{\partial \tau} \right| \\
 &= \frac{\sqrt{2}}{\sigma_0 \sqrt{\pi}} \exp\left(-\frac{1}{2\sigma_0^2 \tau}\right) \cdot \frac{1}{2} \tau^{-\frac{3}{2}} \\
 &= \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{1}{2\sigma_0^2 \tau}\right) \tau^{-\frac{3}{2}}.
 \end{aligned}$$

Then using the change of variables formula again, the induced prior on the log

precision  $\theta$  is given by

$$\begin{aligned} f_{\theta}(\theta) &= f_{\tau} \{ \tau = \exp(\theta) \} \left| \frac{\partial \tau}{\partial \theta} \right| \\ &= \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp \left( -\frac{\exp(-\theta)}{2\sigma_0^2} \right) \exp \left( -\frac{3}{2}\theta \right) \exp(\theta) \\ &= \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp \left( -\frac{\theta}{2} \right) \exp \left( -\frac{\exp(-\theta)}{2\sigma_0^2} \right). \end{aligned}$$

Then, taking a natural log transformation the log prior density is given by

$$\ln\{f_{\theta}(\theta)\} = -\frac{1}{2} \ln(2\pi) - \ln(\sigma_0) - \frac{\theta}{2} - \frac{\exp(-\theta)}{2\sigma_0^2}.$$

For the spatial dependence parameter  $\rho$ , a Gaussian prior with mean  $\mu_{\gamma}$  and variance  $\sigma_{\gamma}^2$  is assigned as the prior distribution on the logit scale  $\gamma = \log\left(\frac{\rho}{1-\rho}\right)$  by the INLA software, ensuring that  $\rho$  remains bounded between 0 and 1. Specifically,

$$\gamma = \text{logit}(\rho) = \log \left( \frac{\rho}{1-\rho} \right) \sim N(\mu_{\gamma}, \sigma_{\gamma}^2),$$

that is,

$$f_{\gamma}(\gamma) = \frac{1}{\sqrt{2\pi\sigma_{\gamma}^2}} \exp \left( -\frac{(\gamma - \mu_{\gamma})^2}{2\sigma_{\gamma}^2} \right).$$

Then the corresponding pdf of  $\rho$  is again given by

$$\begin{aligned} f_{\rho}(\rho) &= f_{\gamma} \left\{ \gamma = \log \left( \frac{\rho}{1-\rho} \right) \right\} \left| \frac{\partial \gamma}{\partial \rho} \right| \\ &= \frac{1}{\sqrt{2\pi\sigma_{\gamma}^2}} \exp \left( -\frac{\left( \log \left( \frac{\rho}{1-\rho} \right) - \mu_{\gamma} \right)^2}{2\sigma_{\gamma}^2} \right) \left| \frac{\partial \gamma}{\partial \rho} \right| \\ &= \frac{1}{\sqrt{2\pi\sigma_{\gamma}^2}} \exp \left( -\frac{\left( \log \left( \frac{\rho}{1-\rho} \right) - \mu_{\gamma} \right)^2}{2\sigma_{\gamma}^2} \right) \frac{1}{\rho(1-\rho)}. \end{aligned}$$

Here I specify  $\mu_{\gamma} = 0$  and  $\sigma_{\gamma} = 2$ , and Figure 4.12 (b) presents the probability density curve for  $\rho$ . The resulting curve is weakly informative and exhibits a slight U-shape, with a higher density at the extremes near zero and one. In cases where the covariates



comprehensively capture all spatial dependence, spatial independence ( $\rho = 0$ ) would be reasonable. Conversely, in scenarios lacking essential covariates, a strong spatial correlation might be evident. In this study, some covariates included exhibit inherent spatial structures but are insufficient to capture all of the spatial dependence in the response, indicating that some extents of residual spatial correlation are still present. This necessitates a prior that accommodates a broad range of potential  $\rho$  values, from independence to strong correlation. The chosen prior does not impose strong assumptions on the distribution of  $\rho$ , allowing the data itself to exert more influence on estimating  $\rho$ . Given the substantial size of the dataset, it is expected that the data will primarily drive the posterior estimates of  $\rho$ , rather than the prior.

Finally, the temporal autocorrelation parameter  $\alpha$  is also assigned a Gaussian prior with mean  $\mu_\eta$  and variance  $\sigma_\eta^2$  on a transformed scale  $\eta = \log\left(\frac{1+\alpha}{1-\alpha}\right)$  by the INLA software. This transformation and the subsequent normal distribution ensure that  $\alpha$  stays within the bounds of -1 and 1, maintaining the feasibility of the correlation parameter. Specifically,

$$\eta = \log\left(\frac{1+\alpha}{1-\alpha}\right) \sim N(\mu_\eta, \sigma_\eta^2),$$

then the pdf is given by

$$f_\alpha(\alpha) = \frac{1}{\sqrt{2\pi}\sigma_\eta} \exp\left(-\frac{\left(\log\left(\frac{1+\alpha}{1-\alpha}\right) - \mu_\eta\right)^2}{2\sigma_\eta^2}\right) \frac{2}{1-\alpha^2}.$$

Here I specify  $\mu_\eta = 0.25$  and  $\sigma_\eta = 2$ , and the pdf curve of  $\alpha$  is given in Figure 4.12 (c). This setup is fairly weakly informative and captures the dual possible nature of temporal dynamics in the context of respiratory disease medication patterns. Temporal autocorrelation in medication data may manifest in contrasting ways, particularly influenced by seasonal behaviours and healthcare practices. For instance, in the winter with a spike in drug use, the increase in respiratory disease often leads to higher medication stock by patients, suggesting a positive autocorrelation where high medication use one month may lead to high use in subsequent months. Conversely, there may be another explanation that if individuals stock up sufficiently in one month, there might be a decrease in purchases the following month, indicating a potential negative autocorrelation coefficient. The prior specification for  $\alpha$  needs to accommodate these possibilities, and the chosen parameters allow the prior to encompass a wide range of  $\alpha$  values. In addition, historical data and time series analysis often show a tendency towards positive temporal autocorrelation in similar

contexts, which supports the slight offset of  $\mu_\eta$  above zero. The scale parameter  $\sigma_\eta = 2$  ensures that the prior is sufficiently non-restrictive, allowing the data to primarily influence estimating  $\alpha$ .

These priors are selected to enable the spatio-temporal smoothing models built in this study to deliver reliable estimates even in the presence of complex spatial and temporal structures in the data. By accommodating a broad range of potential values, these priors ensure that the model remains sufficiently flexible to capture the underlying data dynamics without being overly restrictive.

## 4.6 Results of the motivating study

The results of the study are presented below, and focus on how well each of the proposed neighbourhood matrices captures residual spatial correlation in the prescribing rates (Section 4.6.1). I also present their non-binary counterparts to justify the choice of a binary specification in Section 4.6.2. Having identified suitable neighbourhood matrix structures, I present the estimated pollution-health associations for all three pollutants and medication types (Section 4.6.3), and then investigate the robustness of these findings by examining lagged PM<sub>10</sub> effects (Section 4.6.4) and by propagating uncertainty in the PM<sub>10</sub> predictions through to the health effect estimates (Section 4.6.5).

### 4.6.1 Assessment of the residual spatial correlation

Figures 4.13 to 4.15 present the levels of spatial correlation in the model residuals across all GP surgeries when using  $\mathbf{W}^{(5nn)}$  (Figure 4.13),  $\mathbf{W}^{(O)}$  (Figure 4.14) and  $\mathbf{W}^{(S)}$  (Figure 4.15) for each medication type and study period. These plots are based on the models fitted including NO<sub>2</sub> as the single pollutant, and the results when including the other two pollutants are almost identical and not shown for brevity. Each plot presents the Moran's I statistic for each month of the study and for each threshold, and the PSAI / GSAI / MAI metrics are given in each panel to summarise these residual correlations.

Figure 4.13 highlights that for both study periods (red – 2016–2019; blue – 2016–2020)  $\mathbf{W}^{(5nn)}$  fails to adequately capture the spatial correlation in all three medication types. The models typically induce negative spatial correlation into the residuals as a result of over-smoothing, which is evident because almost all of the Moran's I values are negative. The exception to this is during the start of the Covid-19 pandemic in

March 2020, where there is still positive residual correlation for all three medication types. The induced negative spatial correlations are generally statistically significant, as the GSAI indicator ranges between 15% and 32% across the medication types and study periods, when under independence it should be close to 5%. Conversely, the newly proposed binary neighbourhood matrices  $\{\mathbf{W}^{(O)}, \mathbf{W}^{(S)}\}$  are much better at capturing the spatial correlation in the data, because in both cases at least one threshold  $(\varepsilon, \omega)$  gives a GSAI value relatively close to 5% for each medication type and study period (see Figures 4.14 and 4.15).

For  $\mathbf{W}^{(O)}$ , small threshold values seem to provide the best model fit, with  $\varepsilon = 0.0001$  typically yielding the lowest MAI values between 0.0150 and 0.0213. In contrast, larger values of  $\varepsilon$  tend to either over- or under-smooth the data, resulting in either more positive or negative correlations in the residuals. For  $\mathbf{W}^{(S)}$ , the middle threshold values (e.g.  $\omega = \{0.01, 0.05, 0.1\}$ ) seem to most appropriately capture the spatial correlation in the data, because they result in the smallest MAI values between 0.0153 and 0.0229 across all data sets. Therefore,  $\varepsilon = 0.0001$  and  $\omega = 0.1$  seem to be the optimal thresholds for  $\mathbf{W}^{(O)}$  and  $\mathbf{W}^{(S)}$ , because they typically minimise the GSAI and offer the lowest MAI values thus providing the best control for residual spatial correlation. Based on these optimal thresholds,  $\mathbf{W}^{(O)}$  and  $\mathbf{W}^{(S)}$  perform similarly, having similar GSAI, PSAI and MAI values. They both, however, provide a substantial improvement on  $\mathbf{W}^{(5nn)}$ , because their MAI values range between ( $\mathbf{W}^{(O)}$ : 0.0150–0.0213) and ( $\mathbf{W}^{(S)}$ : 0.0158–0.0208) compared to (0.0230–0.0308) for  $\mathbf{W}^{(5nn)}$ . I also investigated using non-binary versions of  $\{\mathbf{W}^{(O)}, \mathbf{W}^{(S)}\}$ , but these performed very poorly and the results are displayed in the next Section 4.6.2.

### 4.6.2 Results of the non-binary neighbourhood matrices

In the initial model development, I considered non-binary versions of  $\{\mathbf{W}^{(O)}, \mathbf{W}^{(S)}\}$ , because I thought that this would better represent the population overlaps between GP surgeries and hence the spatial correlation structure in the data. These non-binary matrices are constructed from Equations (4.4) and (4.5) by

$$w_{kj}^{(O)} = \begin{cases} w_{kj}^{*(O)} & \text{if } w_{kj}^{*(O)} > \varepsilon \text{ and } k \neq j \\ 0 & \text{otherwise} \end{cases}, \quad w_{kj}^{(S)} = \begin{cases} w_{kj}^{*(S)} & \text{if } w_{kj}^{*(S)} > \omega \text{ and } k \neq j \\ 0 & \text{otherwise} \end{cases},$$

which is analogous to Equation (4.6) except that the non-zero values are not fixed at one. When fitting the spatio-temporal CAR-AR(1) model using these non-binary neighbourhood matrices, I fixed  $\rho = 1$ , which corresponds to the intrinsic CAR prior

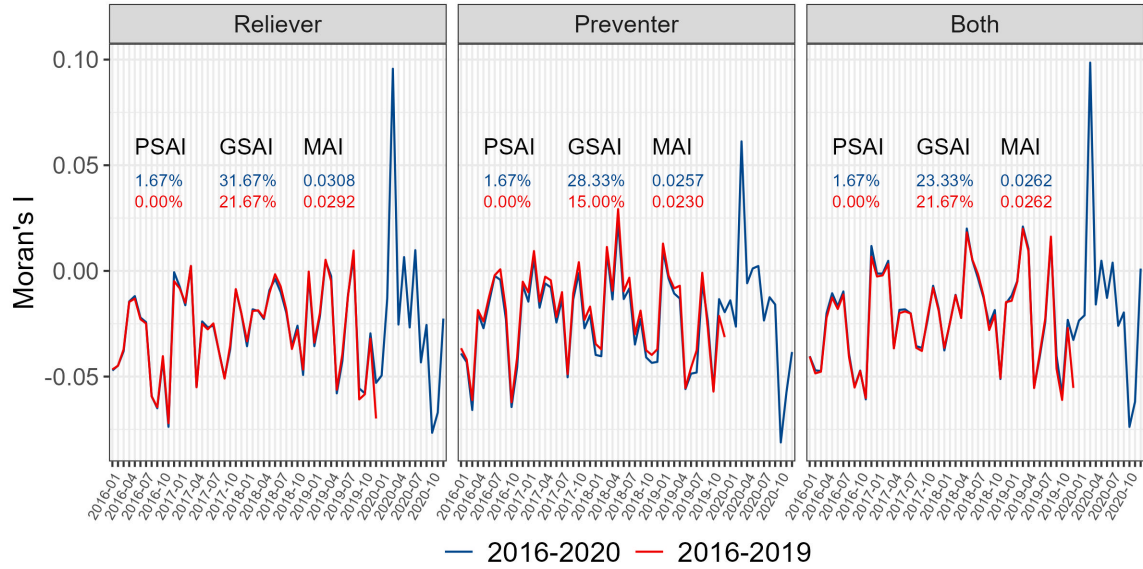


Figure 4.13: Moran's I statistics quantifying the level of spatial correlation in the residuals for each month when using  $\mathbf{W}^{(5nn)}$  as the neighbourhood matrix. Each panel represents a different disease outcome, while red lines relate to the 2016–2019 study period and the blue lines relate to 2016–2020. PSAI, GSAI and MAI indicators are listed in each case.

proposed by [Besag et al. \(1991\)](#). The reason for this is that, when fitting the CAR prior proposed by [Leroux et al. \(2000\)](#) using the R package INLA, I used the `besagproper2` latent model, but this only fits binary matrices. In contrast, the intrinsic CAR prior can be fitted with non-binary neighbourhood matrices using the `generic0` latent model, and as most values of  $\rho$  were very close to one when fitting the binary neighbourhood matrices this restriction will have little impact on the results.

The levels of residual spatial correlation from using these non-binary neighbourhood matrices in the model are summarised in [Figure 4.16](#) and [Figure 4.17](#), which relate to  $\mathbf{W}^{(o)}$  and  $\mathbf{W}^{(s)}$ , respectively. Compared with their binary counterparts, the non-binary matrices lead to consistently higher PSAI, GSAI, and MAI values across almost all threshold settings. For example, when modelling both medication types between 2016 and 2019 using  $\mathbf{W}^{(o)}$ , the lowest MAI values over the five thresholds are 0.0162 for the binary matrix and 0.0951 for the non-binary one. This indicates that the binary specifications are better able to capture the spatial correlation structure amongst the GP surgery prescription counts, and are hence used when quantifying the effects of each air pollutant on each disease outcome.

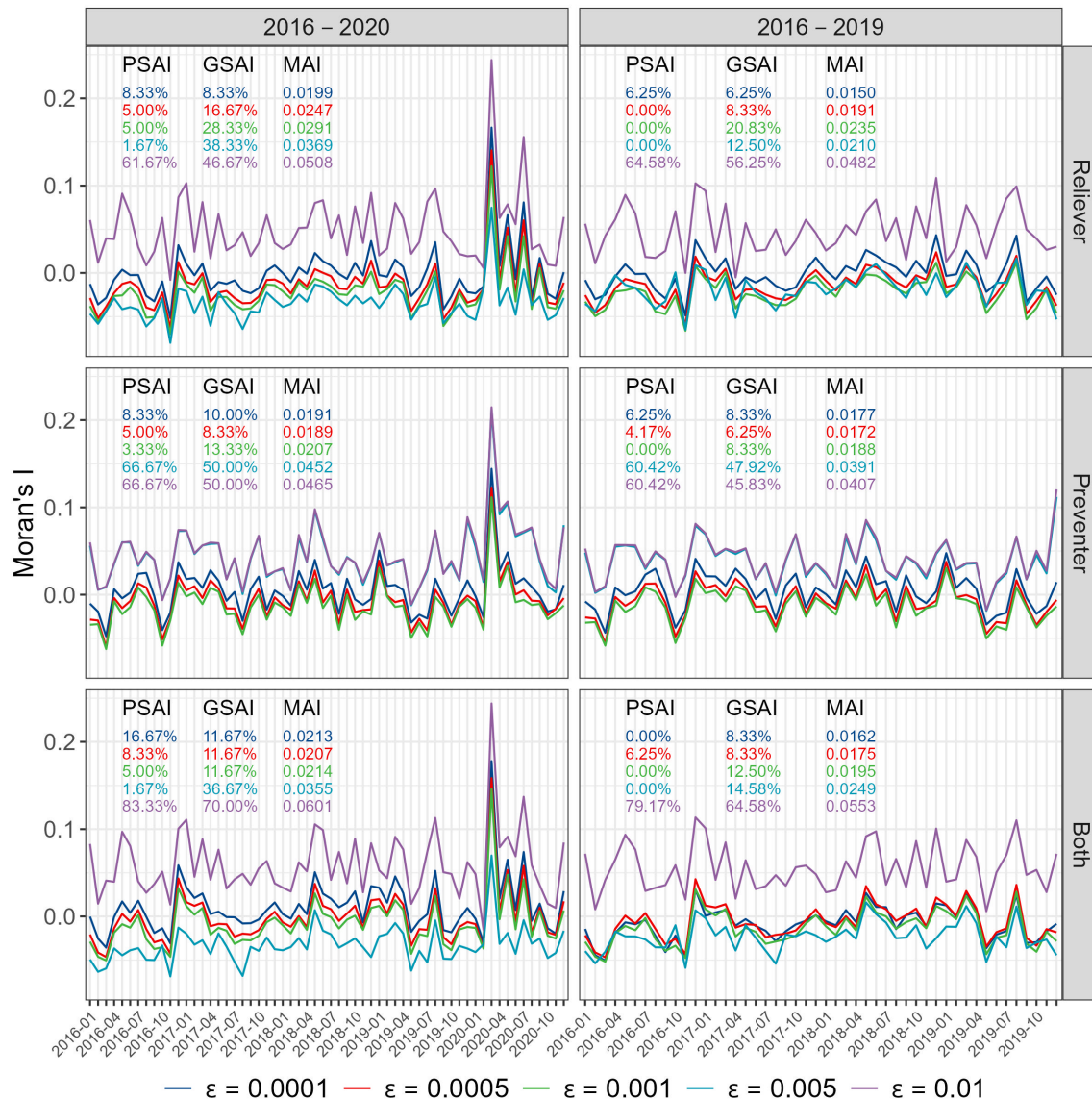


Figure 4.14: Moran's I statistics quantifying the level of spatial correlation in the residuals for each month when using  $W^{(0)}$  as the neighbourhood matrix. Each panel represents a different disease outcome and time period, while the coloured lines relate to the different thresholds  $\epsilon$ . PSAI, GSAI and MAI indicators are listed in each case.

### 4.6.3 Pollution-health associations

The effect of each pollutant on each medication outcome is presented as a relative risk (RR), where a value of one corresponds to no association between air pollution concentrations and respiratory prescription rates. An RR greater than one indicates that increased concentrations are associated with an increased rate of respiratory prescriptions, while an RR less than one indicates a protective effect. Each relative risk



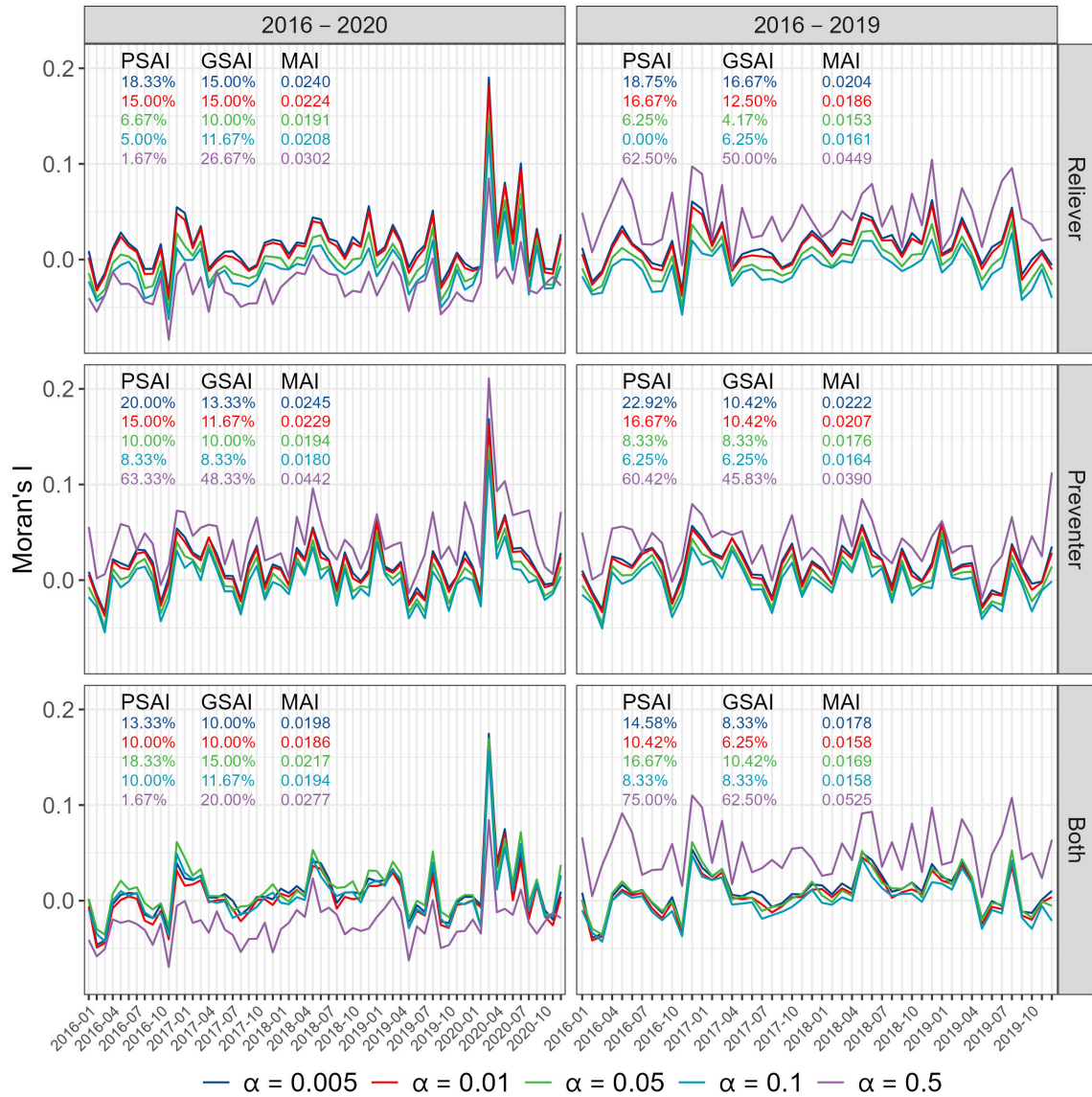


Figure 4.15: Moran's I statistics quantifying the level of spatial correlation in the residuals for each month when using  $W^{(S)}$  as the neighbourhood matrix. Each panel represents a different disease outcome and time period, while the coloured lines relate to the different thresholds  $\omega$ . PSAI, GSAI and MAI indicators are listed in each case.

presented here is computed for a one standard deviation increase in each pollutant, because this represents a realistic increase in the concentrations that could be observed in the present study. Specifically, these standard deviations are:  $\text{NO}_2 - 6.34 \mu\text{g}/\text{m}^3$ ,  $\text{PM}_{10} - 2.77 \mu\text{g}/\text{m}^3$ , and  $\text{PM}_{2.5} - 1.68 \mu\text{g}/\text{m}^3$ . The relative risk results for  $\text{NO}_2$ ,  $\text{PM}_{10}$  and  $\text{PM}_{2.5}$  are displayed in Tables 4.6 to 4.8 respectively. Each table presents the results from 18 model runs, which include all combinations of the three medication

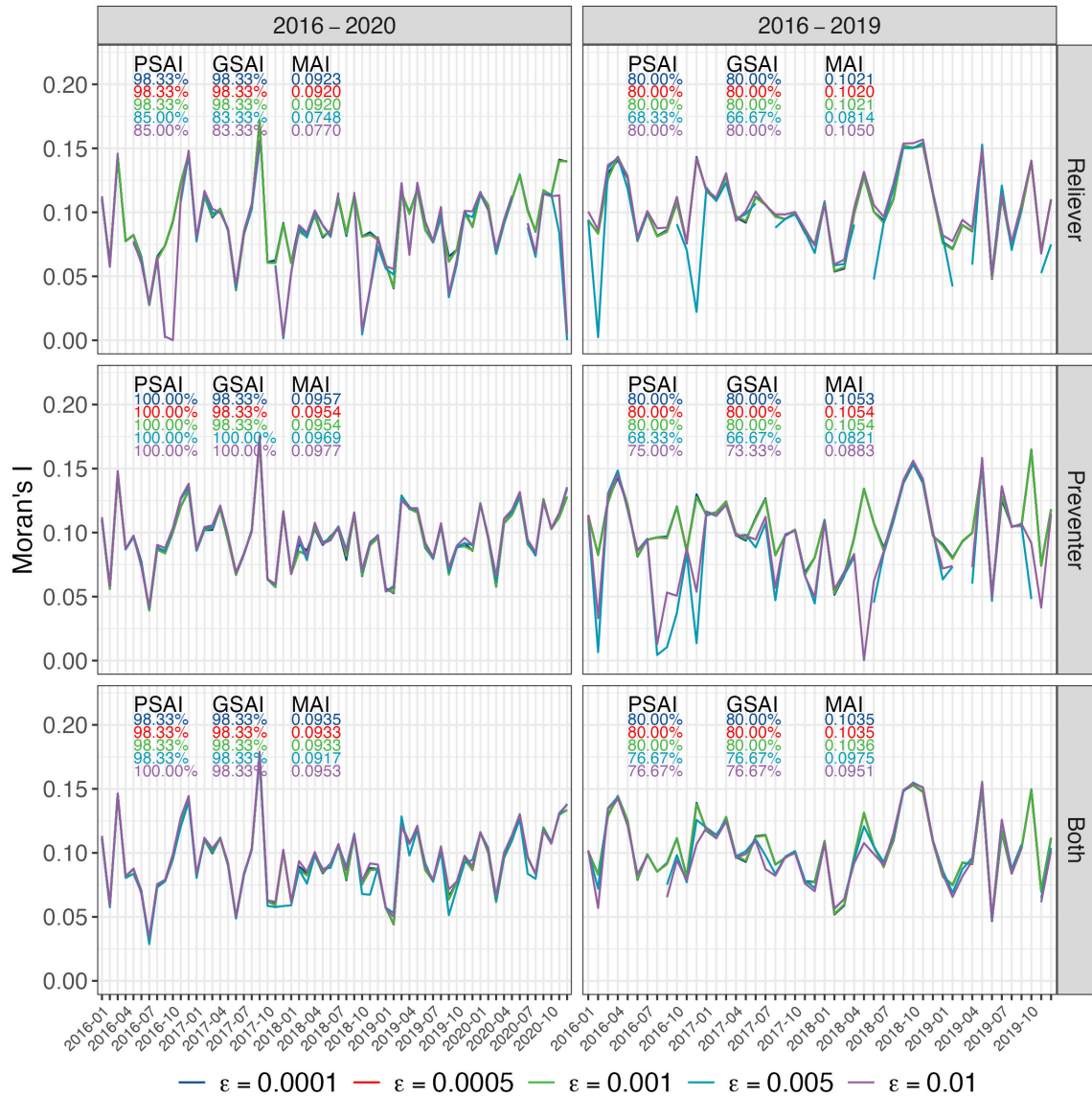


Figure 4.16: Moran's I statistics quantifying the level of spatial correlation in the residuals for each month when using a non-binary version of  $\mathbf{W}^{(O)}$  as the neighbourhood matrix. Each panel represents a different disease outcome and time period, while the coloured lines relate to the different thresholds  $\varepsilon$ . PSAI, GSAI and MAI indicators are listed in each case.

types, three neighbourhood matrices ( $\mathbf{W}^{(5nn)}$ ,  $\mathbf{W}^{(O)}$  and  $\mathbf{W}^{(S)}$ , the latter two using the optimal thresholds) and two time periods. The results for each model run comprise the posterior median relative risk, the exceedance probability (EP) that the estimated relative risk is greater than one, a 95% credible interval, and the DIC and WAIC model fit metrics.

Table 4.6 shows that  $\text{NO}_2$  does not exhibit a strong association with respiratory

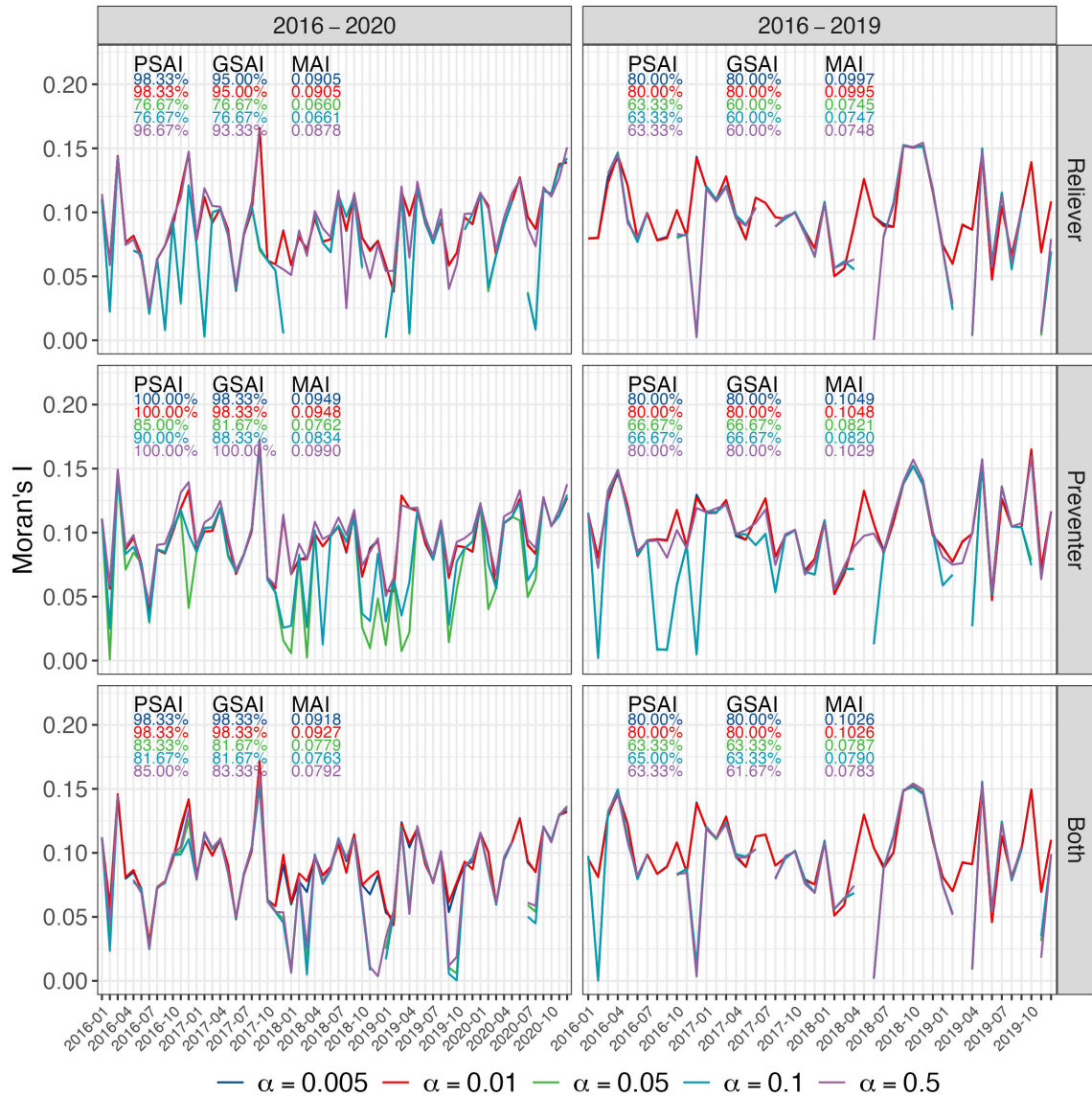


Figure 4.17: Moran's I statistics quantifying the level of spatial correlation in the residuals for each month when using a non-binary version of  $W^{(s)}$  as the neighbourhood matrix. Each panel represents a different disease outcome and time period, while the coloured lines relate to the different thresholds  $\omega$ . PSAI, GSAI and MAI indicators are listed in each case.

medication prescriptions, because all RR estimates are very close to one across all medication types and time periods, while all but one of the 95% credible intervals include the null RR of one. The table also shows that for the pre-pandemic period (2016–2019) all relative risks are above one, while for the full time period (2016–2020) they are all below one. While the absolute differences in these posterior median relative risk are small between the two time periods the exceedance probabilities are



typically substantially different, ranging between 0.6980 and 0.9824 for 2016–2019 and between 0.0054 and 0.4902 for 2016–2020.

Table 4.6: Posterior median relative risk estimates, 95% credible intervals and exceedance probabilities from models based on  $\{\mathbf{W}^{(5nn)}, \mathbf{W}^{(O)}, \mathbf{W}^{(S)}\}$  quantifying the impacts of  $\text{NO}_2$  (based on an increase of  $6.34 \mu\text{g}/\text{m}^3$ ) on respiratory prescription rates for all three medication types and both study periods. The thresholds used for each of  $\{\mathbf{W}^{(O)}, \mathbf{W}^{(S)}\}$  are also given.

| Medication | Matrix               | Period    | Threshold | RR     | EP     | 95% CI           | DIC      | WAIC     |
|------------|----------------------|-----------|-----------|--------|--------|------------------|----------|----------|
| Reliever   | $\mathbf{W}^{(5nn)}$ | 2016–2020 | –         | 0.9967 | 0.1150 | (0.9912, 1.0021) | 477965.5 | 491837.1 |
|            |                      | 2016–2019 | –         | 1.0040 | 0.9189 | (0.9984, 1.0097) | 382085.9 | 392721.4 |
|            | $\mathbf{W}^{(O)}$   | 2016–2020 | 0.0001    | 0.9943 | 0.0054 | (0.9900, 0.9987) | 476169.3 | 489441.9 |
|            |                      | 2016–2019 | 0.0001    | 1.0012 | 0.6980 | (0.9967, 1.0056) | 381224.5 | 390015.6 |
|            | $\mathbf{W}^{(S)}$   | 2016–2020 | 0.05      | 0.9956 | 0.0271 | (0.9912, 1.0001) | 475787.2 | 489265.1 |
|            |                      | 2016–2019 | 0.05      | 1.0031 | 0.9094 | (0.9986, 1.0076) | 381067.6 | 390084.6 |
| Preventer  | $\mathbf{W}^{(5nn)}$ | 2016–2020 | –         | 0.9974 | 0.1797 | (0.9918, 1.0030) | 437902.3 | 448002.2 |
|            |                      | 2016–2019 | –         | 1.0033 | 0.8627 | (0.9974, 1.0091) | 350102.6 | 356375.1 |
|            | $\mathbf{W}^{(O)}$   | 2016–2020 | 0.0005    | 0.9994 | 0.4000 | (0.9945, 1.0043) | 438203.7 | 448005.4 |
|            |                      | 2016–2019 | 0.0005    | 1.0045 | 0.9574 | (0.9994, 1.0096) | 350422.7 | 356842.3 |
|            | $\mathbf{W}^{(S)}$   | 2016–2020 | 0.1       | 0.9999 | 0.4902 | (0.9952, 1.0047) | 438041.6 | 447854.9 |
|            |                      | 2016–2019 | 0.1       | 1.0053 | 0.9824 | (1.0004, 1.0103) | 350226.7 | 356665.0 |
| Both       | $\mathbf{W}^{(5nn)}$ | 2016–2020 | –         | 0.9976 | 0.1678 | (0.9926, 1.0025) | 525300.0 | 543534.2 |
|            |                      | 2016–2019 | –         | 1.0039 | 0.9317 | (0.9988, 1.0091) | 419565.5 | 433303.9 |
|            | $\mathbf{W}^{(O)}$   | 2016–2020 | 0.0005    | 0.9972 | 0.1063 | (0.9929, 1.0016) | 526202.0 | 542255.4 |
|            |                      | 2016–2019 | 0.0001    | 1.0029 | 0.9063 | (0.9986, 1.0072) | 417564.3 | 436106.4 |
|            | $\mathbf{W}^{(S)}$   | 2016–2020 | 0.01      | 0.9959 | 0.0278 | (0.9917, 1.0001) | 521707.8 | 545464.6 |
|            |                      | 2016–2019 | 0.01      | 1.0034 | 0.9582 | (0.9995, 1.0073) | 417731.7 | 435337.5 |

The results for  $\text{PM}_{10}$  are shown in Table 4.7, and demonstrate that there are significant effects of  $\text{PM}_{10}$  on preventer medications for all models and time periods as all of the 95% credible intervals are wholly greater than one. The posterior median relative risks for this pollutant-outcome pair range between 1.0045 and 1.0075, suggesting that, if  $\text{PM}_{10}$  increases by around  $2.77 \mu\text{g}/\text{m}^3$ , then the rate of preventer medication usage will increase by around 0.5%. This significant effect is reinforced by the exceedance probabilities, which are all above 98%. In contrast, there are no significant effects of  $\text{PM}_{10}$  on reliever medications, as all of the 95% credible intervals contain the null relative risk of one. However, there is some evidence of a potential effect, as the exceedance probabilities range between 0.3576 and 0.9738. Moreover, considering both medications jointly shows results in the middle of the previous two as expected, with relative risks ranging between 1.0017 and 1.0063 and exceedance probabilities ranging between 0.7632 and 0.9981. But in this case only two of the six estimates are statistically significant at the 0.05 level. Comparing the

two data spans shows that the relative risks for the entire time period (2016–2020) are generally larger than those for the shorter time period (2016–2019), which is the opposite result observed for  $\text{NO}_2$ . However, these differences do not generally affect the statistical significance.

Table 4.7: Posterior median relative risk estimates, 95% credible intervals and exceedance probabilities from models based on  $\{\mathbf{W}^{(5nn)}, \mathbf{W}^{(O)}, \mathbf{W}^{(S)}\}$  quantifying the impacts of  $\text{PM}_{10}$  (based on an increase of  $2.77 \mu\text{g}/\text{m}^3$ ) on respiratory prescription rates for all three medication types and both study periods. The thresholds used for each of  $\{\mathbf{W}^{(O)}, \mathbf{W}^{(S)}\}$  are also given.

| Medication | Matrix               | Period    | Threshold | RR     | EP     | 95% CI           | DIC      | WAIC     |
|------------|----------------------|-----------|-----------|--------|--------|------------------|----------|----------|
| Reliever   | $\mathbf{W}^{(5nn)}$ | 2016–2020 | –         | 1.0042 | 0.9479 | (0.9991, 1.0092) | 477974.7 | 491835.1 |
|            |                      | 2016–2019 | –         | 1.0031 | 0.9225 | (0.9988, 1.0074) | 382073.1 | 392574.9 |
|            | $\mathbf{W}^{(O)}$   | 2016–2020 | 0.0001    | 0.9990 | 0.3576 | (0.9934, 1.0045) | 476083.9 | 489553.2 |
|            |                      | 2016–2019 | 0.0001    | 1.0006 | 0.5876 | (0.9956, 1.0055) | 381222.0 | 389984.1 |
|            | $\mathbf{W}^{(S)}$   | 2016–2020 | 0.05      | 1.0042 | 0.9738 | (1.0000, 1.0084) | 475780.4 | 489327.3 |
|            |                      | 2016–2019 | 0.05      | 1.0030 | 0.9402 | (0.9992, 1.0067) | 381064.9 | 390094.5 |
| Preventer  | $\mathbf{W}^{(5nn)}$ | 2016–2020 | –         | 1.0062 | 0.9942 | (1.0014, 1.0110) | 437915.3 | 448050.3 |
|            |                      | 2016–2019 | –         | 1.0045 | 0.9812 | (1.0003, 1.0087) | 350110.8 | 356376.4 |
|            | $\mathbf{W}^{(O)}$   | 2016–2020 | 0.0005    | 1.0075 | 0.9997 | (1.0033, 1.0118) | 438202.9 | 448008.2 |
|            |                      | 2016–2019 | 0.0005    | 1.0054 | 0.9971 | (1.0016, 1.0092) | 350430.4 | 356814.9 |
|            | $\mathbf{W}^{(S)}$   | 2016–2020 | 0.1       | 1.0066 | 0.9991 | (1.0024, 1.0107) | 438036.2 | 447858.2 |
|            |                      | 2016–2019 | 0.1       | 1.0046 | 0.9926 | (1.0009, 1.0084) | 350226.9 | 356666.0 |
| Both       | $\mathbf{W}^{(5nn)}$ | 2016–2020 | –         | 1.0042 | 0.9536 | (0.9993, 1.0090) | 525307.9 | 543525.7 |
|            |                      | 2016–2019 | –         | 1.0031 | 0.9222 | (0.9988, 1.0073) | 419567.1 | 433306.3 |
|            | $\mathbf{W}^{(O)}$   | 2016–2020 | 0.0005    | 1.0063 | 0.9981 | (1.0020, 1.0105) | 526212.8 | 542237.1 |
|            |                      | 2016–2019 | 0.0001    | 1.0017 | 0.7632 | (0.9970, 1.0063) | 417577.8 | 436097.6 |
|            | $\mathbf{W}^{(S)}$   | 2016–2020 | 0.01      | 1.0041 | 0.9768 | (1.0001, 1.0081) | 521715.1 | 545465.9 |
|            |                      | 2016–2019 | 0.01      | 1.0025 | 0.9112 | (0.9989, 1.0060) | 417706.1 | 435341.8 |

The results for  $\text{PM}_{2.5}$  are shown in Table 4.8, and are similar but slightly weaker than those for  $\text{PM}_{10}$ . The table shows significant effects of  $\text{PM}_{2.5}$  on preventer medications, with three of the six 95% credible intervals being equal to or above the null relative risk of one, as well as exceedance probabilities ranging between 0.801 and 0.985. However, the sizes of the posterior median relative risks are less than those from the corresponding models for  $\text{PM}_{10}$ , indicating that a one standard deviation increase in  $\text{PM}_{2.5}$  concentrations would lead to a smaller percentage increase in prescription rates for preventing chronic respiratory diseases. For the reliever and both medication types there are no significant effects, as all of the 95% credible intervals contain the null relative risk of one.

The tables of results also display the DIC and WAIC model fit metrics for each neighbourhood matrix. The tables show that for the reliever and both medication

Table 4.8: Posterior median relative risk estimates, 95% credible intervals and exceedance probabilities from models based on  $\{\mathbf{W}^{(5nn)}, \mathbf{W}^{(O)}, \mathbf{W}^{(S)}\}$  quantifying the impacts of  $\text{PM}_{2.5}$  (based on an increase of  $1.68 \mu\text{g}/\text{m}^3$ ) on respiratory prescription rates for all three medication types and both study periods. The thresholds used for each of  $\{\mathbf{W}^{(O)}, \mathbf{W}^{(S)}\}$  are also given.

| Medication | Matrix               | Period    | Threshold | RR     | EP     | 95% CI           | DIC      | WAIC     |
|------------|----------------------|-----------|-----------|--------|--------|------------------|----------|----------|
| Reliever   | $\mathbf{W}^{(5nn)}$ | 2016–2020 | –         | 1.0010 | 0.7101 | (0.9974, 1.0047) | 477970.1 | 491848.2 |
|            |                      | 2016–2019 | –         | 1.0019 | 0.8689 | (0.9986, 1.0051) | 382087.8 | 392730.3 |
|            | $\mathbf{W}^{(O)}$   | 2016–2020 | 0.0001    | 1.0001 | 0.5176 | (0.9966, 1.0035) | 476096.6 | 489527.8 |
|            |                      | 2016–2019 | 0.0001    | 1.0008 | 0.6773 | (0.9975, 1.0040) | 381213.3 | 390015.4 |
|            | $\mathbf{W}^{(S)}$   | 2016–2020 | 0.05      | 1.0019 | 0.8958 | (0.9989, 1.0049) | 475781.5 | 489325.6 |
|            |                      | 2016–2019 | 0.05      | 1.0023 | 0.9517 | (0.9996, 1.0050) | 381067.5 | 390072.2 |
| Preventer  | $\mathbf{W}^{(5nn)}$ | 2016–2020 | –         | 1.0016 | 0.8011 | (0.9980, 1.0051) | 437906.1 | 448018.1 |
|            |                      | 2016–2019 | –         | 1.0021 | 0.8884 | (0.9987, 1.0053) | 350103.3 | 356373.0 |
|            | $\mathbf{W}^{(O)}$   | 2016–2020 | 0.0005    | 1.0035 | 0.9823 | (1.0002, 1.0068) | 438202.8 | 448008.5 |
|            |                      | 2016–2019 | 0.0005    | 1.0033 | 0.9851 | (1.0003, 1.0063) | 350427.0 | 356812.3 |
|            | $\mathbf{W}^{(S)}$   | 2016–2020 | 0.1       | 1.0030 | 0.9728 | (0.9999, 1.0061) | 438038.9 | 447859.5 |
|            |                      | 2016–2019 | 0.1       | 1.0029 | 0.9755 | (1.0000, 1.0057) | 350226.3 | 356669.1 |
| Both       | $\mathbf{W}^{(5nn)}$ | 2016–2020 | –         | 1.0004 | 0.5835 | (0.9969, 1.0038) | 525302.4 | 543533.5 |
|            |                      | 2016–2019 | –         | 1.0012 | 0.7639 | (0.9980, 1.0043) | 419566.2 | 433312.6 |
|            | $\mathbf{W}^{(O)}$   | 2016–2020 | 0.0005    | 1.0024 | 0.9331 | (0.9993, 1.0055) | 526211.4 | 542228.1 |
|            |                      | 2016–2019 | 0.0001    | 1.0006 | 0.6423 | (0.9975, 1.0036) | 417560.9 | 436085.0 |
|            | $\mathbf{W}^{(S)}$   | 2016–2020 | 0.01      | 1.0017 | 0.8813 | (0.9989, 1.0045) | 521701.3 | 545480.6 |
|            |                      | 2016–2019 | 0.01      | 1.0018 | 0.9214 | (0.9993, 1.0043) | 417695.3 | 435350.6 |

types  $\mathbf{W}^{(O)}$  and  $\mathbf{W}^{(S)}$  generally outperform  $\mathbf{W}^{(5nn)}$  in terms of both DIC and WAIC, while for the preventer medication type the results are much closer and  $\mathbf{W}^{(5nn)}$  typically performs better. Thus, while the proposed neighbourhood matrices do not always provide the best model fit they do so in the majority of situations, which taken together with their improved modelling of spatial correlation suggests that they are superior in this application compared to using the 5 nearest neighbours method.

#### 4.6.4 Results for $\text{PM}_{10}$ lagged by one month

To explore the potential delayed effects of air pollution, I conducted an additional analysis by refitting the models using  $\text{PM}_{10}$  concentrations lagged by one month, with this pollutant being chosen because it is the one that exhibits the strongest relationships with respiratory prescription rates. Table 4.9 summarises the posterior median relative risks, exceedance probabilities and 95% credible intervals from the lag-1 analysis, which can be compared to the non-lagged (lag-0) results presented in Table 4.7. The tables show that at lag-1 the  $\text{PM}_{10}$  relative risks are smaller than

those obtained using lag-0. For example, between 2016–2020 the estimated RRs for preventer prescriptions using  $\mathbf{W}^{(O)}$  decrease from 1.0075 (95% CI: 1.0033–1.0118) at lag-0 to 1.0015 (95% CI: 1.0004–1.0026) at lag-1. Similar attenuation is observed across other medication types and spatial structures. However, there is still evidence that  $\text{PM}_{10}$  is significantly associated with preventer medication at lag-1, with three of the 95% credible intervals being equal to or above the null RR of one.

Table 4.9: Posterior median relative risk estimates, 95% credible intervals and exceedance probabilities from models based on  $\{\mathbf{W}^{(5nn)}, \mathbf{W}^{(O)}, \mathbf{W}^{(S)}\}$  quantifying the impacts of Lag-1  $\text{PM}_{10}$  (based on an increase of  $2.77 \mu\text{g}/\text{m}^3$ ) on respiratory prescription rates for all 3 medication types and both study periods. The thresholds used for each of  $\{\mathbf{W}^{(O)}, \mathbf{W}^{(S)}\}$  are also given.

| Medication | Matrix               | Period    | Threshold | RR     | EP     | 95% CI           | DIC      | WAIC     |
|------------|----------------------|-----------|-----------|--------|--------|------------------|----------|----------|
| Reliever   | $\mathbf{W}^{(5nn)}$ | 2016–2020 | –         | 1.0014 | 0.9929 | (1.0003, 1.0025) | 481256.8 | 489984.7 |
|            |                      | 2016–2019 | –         | 0.9984 | 0.0019 | (0.9974, 0.9995) | 380160.9 | 387005.9 |
|            | $\mathbf{W}^{(O)}$   | 2016–2020 | 0.0001    | 1.0017 | 0.9999 | (1.0008, 1.0026) | 479224.6 | 493091.0 |
|            |                      | 2016–2019 | 0.0005    | 0.9985 | 0.0009 | (0.9976, 0.9995) | 378583.3 | 386839.0 |
|            | $\mathbf{W}^{(S)}$   | 2016–2020 | 0.01      | 1.0000 | 0.4730 | (0.9990, 1.0009) | 475280.9 | 489056.2 |
|            |                      | 2016–2019 | 0.1       | 0.9985 | 0.0008 | (0.9976, 0.9994) | 378266.5 | 386998.4 |
| Preventer  | $\mathbf{W}^{(5nn)}$ | 2016–2020 | –         | 1.0016 | 0.9977 | (1.0005, 1.0027) | 436697.5 | 444753.8 |
|            |                      | 2016–2019 | –         | 0.9977 | 0.0000 | (0.9966, 0.9989) | 345821.3 | 350250.8 |
|            | $\mathbf{W}^{(O)}$   | 2016–2020 | 0.0001    | 1.0015 | 0.9955 | (1.0004, 1.0026) | 436944.8 | 445331.7 |
|            |                      | 2016–2019 | 0.005     | 0.9977 | 0.0000 | (0.9966, 0.9988) | 346111.8 | 350060.6 |
|            | $\mathbf{W}^{(S)}$   | 2016–2020 | 0.05      | 1.0015 | 0.9963 | (1.0004, 1.0026) | 436995.3 | 445833.3 |
|            |                      | 2016–2019 | 0.05      | 0.9977 | 0.0000 | (0.9966, 0.9988) | 345954.3 | 350779.6 |
| Both       | $\mathbf{W}^{(5nn)}$ | 2016–2020 | –         | 1.0012 | 0.9882 | (1.0002, 1.0023) | 528234.6 | 537449.0 |
|            |                      | 2016–2019 | –         | 0.9979 | 0.0000 | (0.9969, 0.9989) | 417530.0 | 426135.5 |
|            | $\mathbf{W}^{(O)}$   | 2016–2020 | 0.005     | 1.0012 | 0.9885 | (1.0002, 1.0023) | 528699.8 | 536190.5 |
|            |                      | 2016–2019 | 0.005     | 0.9979 | 0.0000 | (0.9969, 0.9989) | 417709.6 | 426737.8 |
|            | $\mathbf{W}^{(S)}$   | 2016–2020 | 0.005     | 0.9997 | 0.2387 | (0.9989, 1.0005) | 522955.6 | 548082.5 |
|            |                      | 2016–2019 | 0.01      | 0.9981 | 0.0000 | (0.9974, 0.9989) | 416726.9 | 433938.9 |

#### 4.6.5 Results for $\text{PM}_{10}$ incorporating exposure uncertainty

Here, I also assessed the influence of prediction uncertainty in the modelled pollution concentrations on the estimated pollution-health associations. My focus is on the association between  $\text{PM}_{10}$  and prescriptions of preventer medications, because this is the pollutant-outcome pair that exhibits a consistent significant association in the results presented in Table 4.7. To quantify the uncertainty in the predicted  $\text{PM}_{10}$  concentrations, I simulated 100 draws from its predictive distribution. As described in Chapter 3, the best prediction model for  $\text{PM}_{10}$  is a linear regression

model fitted on the natural log scale. One hundred sets of predicted pollution surfaces are sampled from this predictive distribution, and are subsequently back transformed to the original (un-logged) scale.

To quantify the impact that propagating pollution uncertainty into the health model has on the estimated pollution-health association, I followed the approach proposed by [Blangiardo et al. \(2016\)](#) who also modelled prescription rate data. Specifically, I refitted the spatio-temporal model for each of the 100 simulated  $PM_{10}$  surfaces, and generated 1,000 posterior samples of the  $PM_{10}$  association for each model run. This results in 100,000 posterior samples of the pollution-health association, which allow for the uncertainty in the predicted  $PM_{10}$  concentrations. The results are presented in Table 4.10 for both study time periods and the three neighbourhood matrices. The statistical significance of the relative risks remains unchanged after accounting for prediction uncertainty, because all six of the 95% credible intervals are wholly above the null risk of one. However, the posterior median relative risks are slightly bigger in magnitude, ranging between 1.0073 and 1.0094 which compares to a range of 1.0045 to 1.0075 when pollution uncertainty is ignored. Thus, while the absolute size of the relative risks has changed slightly, their statistical significance and hence the main conclusion from the study have not changed.

Table 4.10: Posterior median relative risk estimates, 95% credible intervals and exceedance probabilities from models quantifying the effect of  $PM_{10}$  (based on an increase of  $2.77 \mu g/m^3$ ) on preventer medications while incorporating uncertainty in the  $PM_{10}$  concentrations.

| Matrix      | Period    | Threshold | RR     | EP | 95% CI           |
|-------------|-----------|-----------|--------|----|------------------|
| $W^{(5nn)}$ | 2016–2020 | –         | 1.0094 | 1  | (1.0075, 1.0117) |
|             | 2016–2019 | –         | 1.0073 | 1  | (1.0055, 1.0095) |
| $W^{(O)}$   | 2016–2020 | 0.0005    | 1.0092 | 1  | (1.0073, 1.0114) |
|             | 2016–2019 | 0.0005    | 1.0073 | 1  | (1.0055, 1.0094) |
| $W^{(S)}$   | 2016–2020 | 0.1       | 1.0093 | 1  | (1.0074, 1.0115) |
|             | 2016–2019 | 0.1       | 1.0073 | 1  | (1.0055, 1.0094) |

## 4.7 Discussion

This study has developed spatio-temporal models to quantify the effects of air pollution on the volume of medications prescribed to treat chronic respiratory ill health like asthma and COPD within a primary care setting, and applied the models in a new

study based in Scotland. This study extends the work of [Blangiardo et al. \(2016\)](#) and [Lee \(2018\)](#) to a longer and more recent time period, while examining multiple respiratory medications (reliever, preventer and both) and pollutants ( $\text{NO}_2$ ,  $\text{PM}_{10}$  and  $\text{PM}_{2.5}$ ) to provide a comprehensive assessment of how air pollution influences chronic respiratory ill health that is managed in a non-hospitalised setting. This study also estimates population-level pollution exposures based on where the GP surgery patient populations are likely to live, rather than simply estimating the concentration at each GP surgery itself.

From a methodological standpoint, this study adjusts existing spatial areal unit smoothing models for the situation where the spatial extents of the areal units (i.e., GP surgery catchment areas) are partially unknown and overlapping. The challenge here is how to determine the spatial adjacency between these areal units, and this study proposes novel spatial adjacency structures based on population overlap data obtained on a second set of areal units (i.e., Data Zones) that have known geographical boundaries. I propose two specifications here, which measure spatial closeness based on whether the patient populations from two GP surgeries occupy the same ( $\mathbf{W}^{(O)}$ ) or the same/spatially neighbouring ( $\mathbf{W}^{(S)}$ ) Data Zones. These neighbourhood matrices thus consider where each GP surgery's patient population actually live, rather than just considering the exact locations of the surgeries themselves. As a result, these neighbourhood matrices are better able to capture spatial dependencies and prevent over-smoothing of the spatial random effects than a commonly used 5 nearest neighbours approach is, which is evidenced by lower MAI and GSAI values. For example, the MAI values reduce by between 20.99% and 48.63% when using  $\mathbf{W}^{(O)}$  or  $\mathbf{W}^{(S)}$  in the model compared with using  $\mathbf{W}^{(5nn)}$ . For other researchers modelling the type of data considered here, I recommend fitting models with both  $\mathbf{W}^{(O)}$  and  $\mathbf{W}^{(S)}$  initially, so that the consistency of the results can be observed. Here, the results are fairly consistent, but this may not always be the case. In practice, I recommend choosing between  $\mathbf{W}^{(O)}$  and  $\mathbf{W}^{(S)}$  using the MAI, as the role of the spatial random effects is to capture the unmeasured correlation in the data.

This study provides robust evidence that  $\text{PM}_{10}$  has a consistent significant association with prescriptions of medications used to prevent respiratory ill health events, as the exceedance probabilities exceed 98% across all spatial structures and both data spans (2016–2019 and 2016–2020). The posterior median relative risks suggest that the rate of prescriptions of such medications increases on average by 0.5% when  $\text{PM}_{10}$  increases by  $2.77 \mu\text{g}/\text{m}^3$ . This effect estimate should be interpreted as likely conservative due to attenuation bias from random prediction error in the

Ch3 exposure estimates (see Discussion). Indeed, when the predictive uncertainty in the  $PM_{10}$  concentrations is propagated through the health model (Section 4.6.5), the posterior median relative risks rise to between 0.73% and 0.94%, approaching a 1% increase in prescription rates for the same increment in  $PM_{10}$ , while remaining statistically significant. Although this estimated effect size is small, the population at risk of air pollution is large. For example, over the study period about 2,002,720 items of this type of medication were prescribed on average across the entire country each year, which when multiplied by the above relative risk equates to an additional 10,014 items of preventer medication each year. In contrast, there are no consistent associations between medications used to relieve the symptoms of respiratory ill health and  $PM_{10}$ , while the results when both medications are combined show weak evidence of an effect as their results are a compromise between those for the two medication types. The effects of  $PM_{2.5}$  on respiratory prescription rates are also weaker than the corresponding effects of  $PM_{10}$ , with smaller exceedance probabilities and posterior median relative risks in most cases. In contrast, across all models and medication types  $NO_2$  consistently demonstrates no association with prescription rates, with relative risk estimates being close to one and 95% credible intervals that almost always include one.

Comparing the two data spans highlights the potential influence of the Covid-19 pandemic on the relationship between air pollution and respiratory prescription rates. The most notable difference is observed for  $NO_2$ , where the exceedance probabilities are substantially higher for the pre-pandemic period compared to the whole time period, while the corresponding relative risks are generally above one and below one respectively. That said, in all cases the associations are not significant at the 0.05 level. For  $PM_{10}$ , the differences between the two data spans are smaller, with slightly weaker associations in the pre-pandemic period, although this does not alter the statistical significance of the findings. In contrast,  $PM_{2.5}$  exhibits almost no differences, suggesting that its impact on prescriptions remained stable despite pandemic-related disruptions. There are a number of possible reasons for these pandemic related differential health impacts, including changes in the lifestyle and healthcare-seeking behaviours which may have affected the respiratory prescription counts. For example, the spike in prescription rates in March 2020 (see Figure 4.3) suggests that people stocked up on medications just before the first national lockdown. The pandemic also led to a reduction in pollution levels as a result of national lockdowns and other travel restrictions that reduced vehicle and other emissions. For example, based on pollution predictions used in this study, the mean concentrations across Scotland in March



2020 reduced by  $0.12 \mu\text{g}/\text{m}^3$  ( $\text{NO}_2$ ),  $0.07 \mu\text{g}/\text{m}^3$  ( $\text{PM}_{10}$ ) and  $0.16 \mu\text{g}/\text{m}^3$  ( $\text{PM}_{2.5}$ ) compared to the corresponding concentrations for March 2019. These reductions were still evident later that year, as a result of the additional national and regional lockdowns imposed by the Scottish Government.

The health effect estimates presented in this chapter are necessarily dependent on the accuracy of the predicted  $\text{NO}_2$ ,  $\text{PM}_{10}$  and  $\text{PM}_{2.5}$  concentrations obtained in Chapter 3. Systematic bias in those predictions (for example, under-estimating concentrations in urban areas or over-estimating them in rural areas) would propagate to the estimated relative risks, and random error would attenuate the estimated pollution-health associations toward the null. The best performing prediction models from Chapter 3 achieved RMSE values of 7.70, 2.73 and  $1.15 \mu\text{g}/\text{m}^3$  for  $\text{NO}_2$ ,  $\text{PM}_{10}$  and  $\text{PM}_{2.5}$  respectively, which are non-negligible relative to the corresponding standard deviations used to scale the relative risks. It is therefore plausible that attenuation due to prediction error has contributed to the non-significant sizes observed, particularly for  $\text{NO}_2$  and  $\text{PM}_{2.5}$ . To investigate this formally for the pollutant-outcome pair with the strongest and most consistent association, Section 4.6.5 propagated the full predictive uncertainty of the linear regression model for  $\text{PM}_{10}$  through to the health model. The results show that, while accounting for prediction uncertainty slightly inflates the posterior median relative risks, the associations remain statistically significant. This confirms that the key finding is robust to the level of uncertainty inherent in the pollution predictions, and that the health conclusions are not an artefact of treating those predictions as exact.

A limitation of this study is that the proposed approach to modelling the spatial correlation structure depends on additional population distribution data that may not be available in all countries. For instance in developing countries like China, where primary healthcare is less structured, patients often visit multiple hospitals or pharmacies without adequate registration systems, making it challenging to construct a meaningful neighbourhood matrix based on GP registration data. A further limitation is that the spatio-temporal smoothing model used assumes that the length of the random effects vector and the neighbourhood structure captured by  $\mathbf{W}$  remains constant throughout the 60-month study period, and thus does not dynamically adjust to potential changes in GP surgery locations or shifts in patient registration patterns over time. Therefore, in future work we will explore alternative random effects models with adaptive neighbourhood structures that can evolve over time to reflect changes in patient distribution and differing numbers of GP surgeries. The study also assumes that the spatial distribution of the set of registered patients from



each GP surgery to the surrounding Data Zones matches the distribution of the subset of patients actually receiving respiratory related prescriptions. This assumption may lead to potential misclassification, since disease prevalence and thus prescription rates differ across Data Zones served by a GP surgery. In the future, I will explore a more flexible Bayesian hierarchical modelling approach, which will allow the proportions of patients receiving prescriptions from a GP surgery in each Data Zone to be treated as an unknown random vector and modelled using a Dirichlet distribution, at a latent level. This random vector can be informed by the observed proportions of registered patients in each Data Zone, with the mean of the Dirichlet distribution depending on these proportions.

A further limitation concerns the GP prescribing data themselves. The prescription counts capture medications prescribed in primary care only, so respiratory medications issued through secondary care, for example during hospital admissions or by specialist clinics, are not included. In addition, the national prevalence rates used for the indirect standardisation are derived from the SPIRE database, which covers just over 65% of GP surgeries in Scotland, so the resulting rates may not be fully representative of all practices. Both factors mean that the burden of respiratory ill health managed in primary care could be underestimated. However, because the expected counts are rescaled by the national prescription rate  $\theta$ , any such national-level underestimation affects all surgeries approximately proportionally, and is therefore unlikely to materially bias the estimated pollution–health associations, although it does add uncertainty to the absolute scale of the estimated risks.

# Chapter 5

## Estimating small-area respiratory disease burden from spatially misaligned primary care data

### 5.1 Introduction

Chronic respiratory diseases such as asthma and chronic obstructive pulmonary disease (COPD) cause a substantial burden on individuals, health services and society. In the United Kingdom, around 5.4 million people receive treatment for asthma and around three million live with COPD ([NICE, 2025a,b](#)). In Scotland, respiratory disease accounts for a disproportionate share of the national disease burden, and prevalence rates are higher in the most deprived communities ([Scottish Government, 2023](#)). Their prevalence varies across space and time, and understanding this variation is essential for public health surveillance, the planning of primary care services, and the targeting of interventions towards communities with the greatest need. Disease mapping provides a statistical framework for estimating spatio-temporal variation in disease burden from population-level summary data observed on non-overlapping areal units over consecutive time periods. In particular, disease mapping analyses can be used to address questions on which communities carry the highest disease burden, how disease patterns evolve over time, and the extent of spatial and socio-economic inequalities in disease prevalence.

In practice, however, health outcomes and contextual determinants are often recorded on different spatial supports. In Scotland, routinely collected indicators of respiratory ill health treated in primary care are available as counts of prescriptions at

the level of general practitioner (GP) surgeries, while socio-economic and demographic covariates are provided for Data Zones (DZs), which are small-area geographical units designed for statistical reporting. Because GP catchment areas overlap and patients are not necessarily registered at their nearest surgery, the residential distribution of a surgery's registered population is only partially observed. This spatial misalignment creates a "change-of-support" problem ([Gelfand et al., 2001](#)), in which the target inferential scale for disease mapping and inequity assessment is naturally the Data Zone, but the health outcome is observed at the GP level.

A naive approach to misalignment is deterministic area- or population-weighted interpolation of GP-level risks to the DZ scale ([Flowerdew and Green, 1992](#)). Such approaches implicitly assume homogeneity of risk within each GP catchment and do not provide principled model-based uncertainty quantification, although resampling schemes such as bootstrapping can produce approximate intervals. Recent methodological developments in spatial statistics address misaligned areal data more rigorously by defining a latent risk surface on the target spatial support and linking observations at other spatial scales through a probabilistic aggregation mechanism. [Lee and Anderson \(2023\)](#) defined a latent risk surface on a chosen spatial scale and linked primary care prescriptions, hospital admissions and deaths to this surface through data augmentation and a Bayesian hierarchical model. [Gramatica et al. \(2021\)](#) developed a multiple-membership (MM) approach in which conditional autoregressive (CAR) random effects on one areal scale are transformed via a membership matrix to induce structured covariance on another scale. A follow-up paper ([Gramatica et al., 2024](#)) examined the theoretical properties of this MM-CAR construction, clarifying conditions under which the transformation preserves desirable features such as sparsity and identifiability.

This chapter develops a Bayesian disease mapping framework for GP-level prescription counts that is similar in spirit to the MM-CAR construction of [Gramatica et al. \(2021, 2024\)](#) and the joint-realignment formulation of [Lee and Anderson \(2023\)](#). All three approaches link observations on one areal support to a latent risk surface on another through a population-weighted linkage. This study differs in two concrete respects. First, the aggregation in the change-of-support mechanism here operates on the relative-risk scale rather than on the log-linear predictor scale used in [Gramatica et al. \(2021\)](#), so that each GP-level relative risk is a convex combination of DZ-level relative risks. Second, the framework extends the MM family to a spatio-temporal setting through a dynamic ICAR-AR(1) process, whereas [Gramatica et al. \(2021, 2024\)](#) focus on cross-sectional disease mapping and [Lee and Anderson](#)

(2023) augment the likelihood with imputed DZ-level counts rather than aggregating on the risk scale. The simulation studies reported in Sections 5.4 and 5.5 instead quantify the inferential cost of spatial misalignment itself, by comparing the proposed realignment model against naive deterministic interpolation and against an idealised direct DZ-level benchmark. The framework produces spatially smooth DZ-level risk surfaces with coherent uncertainty quantification using only GP-level prescription counts together with GP-DZ patient registration data.

In Chapter 4, I developed a spatio-temporal CAR-AR(1) model to quantify the health effects of air pollution on respiratory illness at the GP surgery level across Scotland. Although GP surgeries have known geographic coordinates, their registered patient populations live in surrounding Data Zones, whose catchment areas are overlapping and partially unknown. This limits the ability to produce consistent and interpretable maps of disease risk at the small-area level. Consequently, mapping disease risk at the DZ level is motivated by the two considerations. First, Data Zones are the official spatial unit for statistical reporting in Scotland, ensuring compatibility with socio-economic covariates used to assess health inequalities and facilitating intervention planning. Second, because patients are not necessarily registered at their nearest GP surgery, GP-level prescription counts may conflate disease risk across the residential areas of a surgery’s registered population. Realigning to the DZ level allows risk to be attributed to where people actually live rather than where they receive primary care.

The number of required DZ-level risk estimates ( $M = 746$  DZs per month in the motivating study) exceeds the number of observations per month ( $K = 141$  in 2019 and  $K = 139$  in 2020), so the latent DZ-level risk surface is unlikely to be fully identified by the data alone. Recovery of the risk surface therefore utilises two sources of external information: (i) spatially or spatio-temporally structured DZ-level covariates that inform the latent surface through the linear predictor; and (ii) a prior that assumes the surface is sufficiently smooth so that neighbouring Data Zones and time slices can borrow information from one another. The simulation studies in Section 5.4 and Section 5.5 quantify the extent to which these two sources contribute to recovery of the DZ-level risks.

The remainder of this chapter is organised as follows. Section 5.2 introduces the Glasgow City study region, describes the prescription and covariate data, and states the public health questions that motivate the analysis. Section 5.3 presents the Bayesian hierarchical model, including the change-of-support mechanism linking GP-level observations to DZ-level risks, the spatio-temporal random effects specification,

and prior and computational choices. Sections 5.4 and 5.5 evaluate the model through two simulation studies, the first evaluating a purely spatial setting and the second a spatio-temporal setting with three temporal evolution scenarios. Finally, Section 5.6 applies the model to the Glasgow City prescription data and presents the resulting small-area disease maps and inequality summaries.

## 5.2 Motivating study

### 5.2.1 Study data

The study region is Glasgow City, Scotland, partitioned into  $M = 746$  Data Zones. The study period covers the  $T = 24$  months from January 2019 to December 2020. Restricting the analysis to a single city allows the methodology to be validated in a tractable setting before a national-scale extension is discussed in Section 5.7. The 24-month study window spans both a pre-pandemic period and the onset of the Covid-19 pandemic, allowing examination of both baseline patterns and the response of the model to a sharp overall disruption in prescribing behaviour during the national lockdown.

Monthly GP-level prescription counts and corresponding expected counts are available for 141 (in 2019) and 139 (in 2020) GP surgeries with at least 75% of their registered patients residing in the study region. The 75% threshold is chosen by two criteria applied to the city-fraction distribution: (i) it should be high enough that the retained surgeries serve a Glasgow-dominant patient population; and (ii) the resulting set of surgeries should retain at least 95% of the Glasgow registered patient population. Thresholds in the range 75%–78% produce identical datasets (141 surgeries in 2019, 139 in 2020, with 95.7% Glasgow patient retention in both years). The outcome  $Y_t(\mathcal{S}_k)$  is the total number of dispensed items for chronic respiratory medications at surgery  $\mathbf{s}_k$  in month  $t$ , comprising bronchodilators (BNF Chapter 3.1) and inhaled corticosteroids (BNF Chapter 3.2). The corresponding expected count is obtained by indirect standardisation using age–sex strata and national prevalence rates,

$$e_t(\mathcal{S}_k) = \sum_{j=1}^J n_{kj}^{(t)} \times r_j,$$

where  $j = 1, \dots, J$  indexes age–sex strata,  $n_{kj}^{(t)}$  is the number of patients in stratum  $j$  registered at surgery  $\mathbf{s}_k$  in month  $t$ , and  $r_j$  is the national age–sex-specific pre-

scribing rate averaged over 2016–2020. Further details of the prescription data and standardisation procedure are given in Section 4.2.2.

Glasgow City Data Zones are geographically compact, with substantial socio-economic heterogeneity across Data Zones and relatively homogeneous populations within each Data Zone. The median Data Zone area is  $0.14 \text{ km}^2$  (5th and 95th percentile:  $0.04\text{--}0.58 \text{ km}^2$ ). Based on the Scottish Index of Multiple Deprivation (SIMD) 2020 classification, 227 Data Zones (30.4%) lie in the most deprived national decile and 112 (15.0%) lie in decile 2. By construction, each SIMD decile contains 10% of Scottish Data Zones, so the 30.4% in decile 1 and 15.0% in decile 2 observed in Glasgow indicate a high concentration of deprivation relative to Scotland as a whole (Figure 5.1). The least deprived deciles are correspondingly under-represented, with 5% (38 Data Zones) in decile 10 and 6% (47) in decile 9.

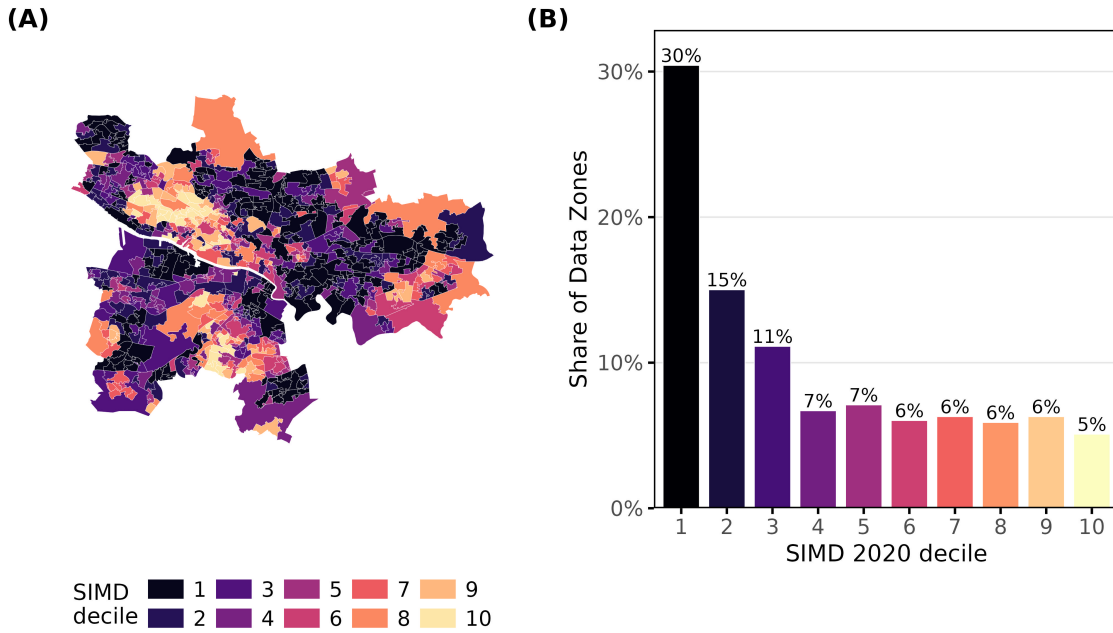


Figure 5.1: Socio-economic context of the Glasgow City study region. Panel (A) maps SIMD 2020 deprivation deciles across the  $M = 746$  Data Zones (decile 1 denotes the most deprived 10% of Scottish Data Zones); Panel (B) shows the distribution of Glasgow City Data Zones by SIMD decile.

At the city-wide level, the crude standardised prescription ratio (SPR), defined as the ratio of observed to expected prescriptions,  $\text{SPR}_t(\mathcal{S}_k) = Y_t(\mathcal{S}_k)/e_t(\mathcal{S}_k)$ , remained above one throughout the study period, with a median monthly value of approximately 1.18. This indicates that Glasgow’s observed respiratory prescribing burden during the study period was approximately 18% higher than the national age–sex standardised

average, consistent with the city’s elevated deprivation profile shown in Figure 5.1. The highest monthly SPR occurred in March 2020 (SPR  $\approx 1.66$ ), coinciding with the onset of the Covid-19 pandemic. Substantial heterogeneity is also evident between GP practices, where the median long-run GP-level SPR was approximately 1.22, with 5th and 95th percentiles of 0.59 and 2.11 respectively. Figure 5.2 summarises the temporal and cross-sectional variation in GP-level SPRs. Apart from the March 2020 spike, the monthly median SPR shows no discernible temporal trend over the study period, suggesting that the underlying city-wide burden of chronic respiratory prescriptions remained broadly stable between January 2019 and December 2020.

GP–DZ patient registration records are also available at annual resolution and provide the population weights that link each surgery’s registered patients to the Data Zones in which they reside. Letting  $p_t(\mathcal{A}_m \cap \mathcal{S}_k)$  denote the number of patients registered at surgery  $\mathcal{S}_k$  in month  $t$  who reside in Data Zone  $\mathcal{A}_m$ , the GP-to-DZ registration weight is

$$\eta_{km,t} = \frac{p_t(\mathcal{A}_m \cap \mathcal{S}_k)}{\sum_{\ell=1}^M p_t(\mathcal{A}_\ell \cap \mathcal{S}_k)}, \quad \text{where } \sum_{m=1}^M \eta_{km,t} = 1,$$

and the matrix  $\{\eta_{km,t}\}$  is the population-weighted linkage that drives the change-of-support mechanism developed in Section 5.3. Because registration data are published once per calendar year,  $\eta_{km,t}$  is constant within each year, so two distinct weight matrices apply to the motivating study (one for 2019, one for 2020).

### 5.2.2 Aims and motivating questions

The analysis in this chapter is designed to support inference on the small-area burden of respiratory illness treated in primary care in Glasgow City when outcomes are observed only at the GP surgery level. By producing spatially smooth maps, the analysis aims to provide in-depth epidemiological insights into the distribution of respiratory disease prevalence and the extent of health inequalities across Glasgow City. The study is guided by the following public health questions:

- (a) Which Data Zones have the highest and lowest burdens of respiratory disease over the study period?
- (b) How certain are the models’ estimates of DZ-level health risk?
- (c) What are the city-wide and DZ-specific temporal trends in the risk of chronic

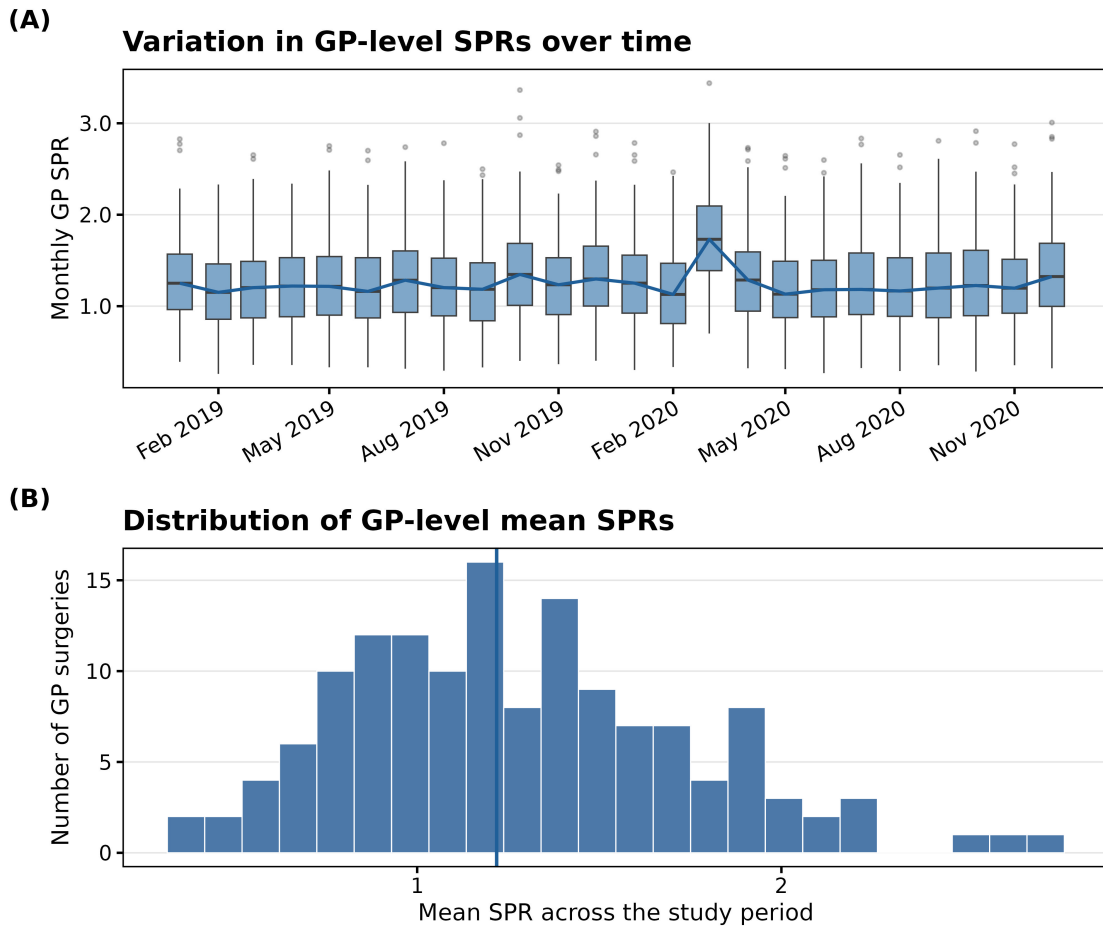


Figure 5.2: Summary of GP-level prescription SPRs in Glasgow City, January 2019 to December 2020. Panel (A) shows box plots of GP-level SPRs over time for the 141 surgeries retained in 2019 and 139 retained in 2020 by the 75% city-fraction rule, with the line tracking the monthly median. Panel (B) shows the distribution of time-averaged GP-level SPRs across the 142 distinct surgeries that satisfy the 75% rule in at least one of the two study years, with the vertical line marking the median. The union count (142) slightly exceeds each within-year count because a small number of surgeries opened, closed, or crossed the 75% threshold between the two years.

respiratory conditions treated in primary care, and which areas show the most pronounced changes over time?

- (d) To what extent do total and socio-economic health inequalities exist in disease burden, and how do these inequalities change over time?

Addressing questions (a) and (b) will help identify the communities with the highest burdens of respiratory disease and assess the reliability of these estimates, providing an evidence base for targeting interventions where they are most needed.



Question (c) focuses on understanding temporal patterns in disease prevalence, both at the national and local scale, and on identifying the factors associated with rising trends. Finally, question (d) concerns health inequalities, where total inequality reflects the overall variation in disease risk across all Data Zones, while socio-economic inequality captures systematic differences in disease burden between more and less deprived communities, with deprivation quantified using the SIMD deciles.

## 5.3 Methodology

The goal of this section is to describe a Bayesian hierarchical spatio-temporal realignment model that estimates disease risk at the Data Zone (DZ) level using only GP-level prescription counts and the observed GP-DZ patient registration data. The model is built in three layers. The data model specifies the likelihood of the observed GP-level counts conditional on latent DZ-level risks, with a population weighted aggregation mechanism linking the two spatial supports. The process model captures the residual spatio-temporal correlation in DZ-level risks using a dynamic CAR-AR(1) structure. Posterior inference is carried out using Markov chain Monte Carlo (MCMC) simulation implemented in the NIMBLE software ([de Valpine et al., 2017](#)). Throughout this section,  $k = 1, \dots, K$  indexes GP surgeries,  $m = 1, \dots, M$  indexes Data Zones, and  $t = 1, \dots, T$  indexes monthly time periods.

### 5.3.1 Data model and change of support

Let  $Y_t(\mathcal{S}_k)$  denote the observed prescription count for GP surgery  $\mathbf{s}_k$  at month  $t$ , and let  $e_t(\mathcal{S}_k)$  be the corresponding expected count of prescriptions. As in Chapter 4,  $e_t(\mathcal{S}_k)$  is obtained in two steps. Indirect standardisation of the registered patient population by age and sex first gives the expected number of eligible patients  $e_t^*(\mathcal{S}_k)$ , which is then rescaled by the national prescription rate  $\theta$  so that  $e_t(\mathcal{S}_k) = \theta e_t^*(\mathcal{S}_k)$  is on the scale of prescription counts. The latent relative risk  $\theta_t(\mathcal{S}_k)$  introduced below then measures departures of the observed prescribing from this expected count. The standardised prescription ratio  $\text{SPR}_t(\mathcal{S}_k)$  provides a demographically adjusted measure of disease burden managed in primary care relative to the national average.

GP-level prescription counts frequently exhibit over-dispersion relative to a Poisson model, meaning that the observed variance exceeds the mean. To accommodate this,

I adopt a negative binomial data likelihood model. Specifically,

$$Y_t(\mathcal{S}_k) \sim \text{NegBin}(e_t(\mathcal{S}_k) \theta_t(\mathcal{S}_k), \delta), \quad (5.1)$$

where  $\theta_t(\mathcal{S}_k) > 0$  is the latent relative risk at GP surgery  $\mathbf{s}_k$  in month  $t$ , and  $\delta > 0$  is an inverse-dispersion parameter shared across all surgeries and time periods. Under this parameterisation, the mean is  $\mathbb{E}[Y_t(\mathcal{S}_k)] = e_t(\mathcal{S}_k) \theta_t(\mathcal{S}_k)$  and the variance is  $\text{Var}[Y_t(\mathcal{S}_k)] = e_t(\mathcal{S}_k) \theta_t(\mathcal{S}_k) \{1 + e_t(\mathcal{S}_k) \theta_t(\mathcal{S}_k) / \delta\}$ , so that larger values of  $\delta$  correspond to less over-dispersion relative to the Poisson model (equal mean and variance) that arises in the limit as  $\delta \rightarrow \infty$ .

The central modelling challenge is that inference is desired at the DZ level, but outcomes are only observed at the GP level. The two spatial supports are linked through the change-of-support assumption that each GP-level relative risk is a population weighted average of the underlying DZ-level risks. Let  $\theta_t(\mathcal{A}_m)$  denote the latent relative risk in Data Zone  $\mathcal{A}_m$  at month  $t$ . The GP-level risk is then expressed as

$$\theta_t(\mathcal{S}_k) = \sum_{m=1}^M \eta_{km,t} \theta_t(\mathcal{A}_m), \quad (5.2)$$

where

$$\eta_{km,t} = \frac{p_t(\mathcal{A}_m \cap \mathcal{S}_k)}{\sum_{\ell=1}^M p_t(\mathcal{A}_\ell \cap \mathcal{S}_k)} \quad (5.3)$$

is the proportion of patients registered at surgery  $\mathbf{s}_k$  who reside in Data Zone  $\mathcal{A}_m$  in month  $t$ . Here  $p_t(\mathcal{A}_m \cap \mathcal{S}_k)$  denotes the number of patients living in  $\mathcal{A}_m$  and registered with surgery  $\mathbf{s}_k$ , obtained from annual GP-DZ patient registration records. Because the registration data are published once per calendar year, the weights are constant within each year, i.e.  $\eta_{km,t} = \eta_{km,t'}$  for all months  $(t, t')$  in the same year. In the motivating study this yields two distinct weight matrices, one for 2019 and one for 2020. By the definition in Equation (5.3), the weights satisfy  $\sum_{m=1}^M \eta_{km,t} = 1$  for each surgery  $\mathbf{s}_k$  once normalised over the Data Zones considered in the analysis. When the study region is restricted to a subset of all Data Zones (e.g., Glasgow City), only patients registered in that subset contribute to the denominator, and the weights are re-normalised so that  $\theta_t(\mathcal{S}_k)$  remains a convex combination of the DZ-level risks within the study region. The full details of how weights are prepared are given in Section 5.3.4.

The DZ-level log relative risk is modelled as a linear combination of fixed covariate

effects and a residual spatio-temporal random effect:

$$\log(\theta_t(\mathcal{A}_m)) = \mathbf{x}_t^\top(\mathcal{A}_m) \boldsymbol{\beta} + \sigma_\phi \cdot \phi_t(\mathcal{A}_m), \quad (5.4)$$

where  $\mathbf{x}_t(\mathcal{A}_m)$  is a  $p \times 1$  vector of DZ-level covariates including an intercept,  $\boldsymbol{\beta}$  is the corresponding  $p \times 1$  vector of regression coefficients,  $\phi_t(\mathcal{A}_m)$  is a standardised spatio-temporal random effect that captures residual variation not explained by the covariates, and  $\sigma_\phi > 0$  controls the marginal standard deviation of the random effects on the log-risk scale. The joint specification of Equations (5.1) to (5.4) constitutes the data likelihood model. Its key feature is that inference for the DZ-scale quantities  $\{\theta_t(\mathcal{A}_m)\}$  and  $\{\phi_t(\mathcal{A}_m)\}$  is achieved by propagating information from the GP-level observations through the population weighted aggregation in Equation (5.2).

### 5.3.2 Spatio-temporal random effects

The linear predictor Equation (5.4) includes a spatio-temporal effect  $\phi_t(\mathcal{A}_m)$  to capture residual variation in DZ-level risks not explained by the observed covariates, arising from unmeasured spatial risk factors, measurement error in covariates, or other spatio-temporal clustering and autocorrelation in disease risk. Letting  $\boldsymbol{\phi}_t = (\phi_t(\mathcal{A}_1), \dots, \phi_t(\mathcal{A}_M))^\top$ , I adopt a dynamic CAR-AR(1) specification (Rushworth et al., 2014) that combines spatial smoothing at each time point with a temporally correlated first-order autoregressive temporal evolution:

$$\begin{aligned} \boldsymbol{\phi}_1 &\sim N(\mathbf{0}, \mathbf{Q}^-), \\ \boldsymbol{\phi}_t | \boldsymbol{\phi}_{t-1} &\sim N(\alpha \boldsymbol{\phi}_{t-1}, \mathbf{Q}^-), \quad t = 2, \dots, T, \end{aligned} \quad (5.5)$$

where  $\mathbf{Q} = \text{diag}(\mathbf{W}\mathbf{1}) - \mathbf{W}$  is the  $M \times M$  ICAR precision matrix (Section 2.6.5) derived from the binary border-sharing adjacency matrix  $\mathbf{W}$ ,  $\mathbf{Q}^-$  denotes its generalised inverse, and  $\alpha \in (0, 1)$  is the temporal autocorrelation parameter. The conditional mean  $\alpha \boldsymbol{\phi}_{t-1}$  captures temporal persistence, with values of  $\alpha$  close to one indicating that the previous month's surface dominates the conditional mean of the current month, while values close to zero indicate that successive monthly surfaces are approximately conditionally independent. The month-to-month change in the risk surface is controlled jointly by  $\alpha$  and the innovation variance  $\sigma_\phi^2$ , so that a smoothly evolving risk surface requires both  $\alpha$  close to one and a small innovation variance.

As discussed in Chapter 2, the ICAR precision matrix  $\mathbf{Q}$  is rank-deficient with a one-dimensional null space spanned by  $\mathbf{1}$  so, without further restriction, the random

effects are not individually identifiable from the data. To resolve this, a sum-to-zero constraint is imposed on the initial surface only, i.e.,

$$\sum_{m=1}^M \phi_1(\mathcal{A}_m) = 0,$$

which restricts the global mean of the random effects at  $t = 1$  to zero and ensures that  $\beta_0$  is interpretable as the overall mean baseline log-risk at the start of the study period. Critically, no such constraint is imposed for  $t = 2, \dots, T$ . This means the global spatial mean  $\bar{\phi}_t = M^{-1} \sum_{m=1}^M \phi_t(\mathcal{A}_m)$  is free to drift over time, and it is precisely this drift that allows  $\phi_t$  to track changes in the overall level of disease burden across the study period. The AR(1) structure ensures that any such temporal drift can potentially evolve more smoothly or more coarsely, as informed by the data, with each surface  $\phi_t$  centred on  $\alpha\phi_{t-1}$ .

The parameterisation in Equation (5.4), where  $\sigma_\phi$  acts as a coefficient rather than being incorporated into the precision matrix, is a non-centred parameterisation. This formulation reduces posterior dependence between  $\sigma_\phi$  and the individual random effects  $\phi_t(\mathcal{A}_m)$ , which typically improves MCMC mixing. A further computational advantage is that the variance of the raw random effects does not depend on  $\sigma_\phi$ , so the proposal scale adapted by the M-H sampler for the random effect updates does not need to track changes in  $\sigma_\phi$  during sampling.

It should be noted that under this single constraint approach, the values  $\phi_t(\mathcal{A}_m)$  for  $t > 1$  contain both a global temporal component (the drift in  $\bar{\phi}_t$ ) and a local spatial deviation component. The estimated DZ-level risks  $\theta_t(\mathcal{A}_m)$  remain fully interpretable, but the spatial and temporal components of the random effects are not separately identified. When the number of time points (months) is only one ( $T = 1$ ), Equation (5.5) reduces to the standard spatial ICAR model with the single sum-to-zero constraint, which forms the basis for Simulation Study 1 in Section 5.4.

### 5.3.3 Prior specification and posterior inference

Prior distributions for the model parameters are specified here and used throughout the simulation studies and the motivating study. The same priors apply to both the spatial ( $T = 1$ ) and the spatio-temporal specifications, with the temporal-dependence

prior applied only when  $T > 1$ . The priors are defined as

$$\begin{aligned}\beta_0 &\sim N(0, 1^2), \\ \beta_1 &\sim N(0, 1^2) \quad \text{when a covariate is included,} \\ \sigma_\phi &\sim \text{Half-Normal}(0, 0.25^2), \\ \delta &\sim \text{Gamma}(2, 0.05), \\ \text{logit}(\alpha) &\sim N(1.5, 0.5^2) \quad \text{when } T > 1.\end{aligned}$$

These priors are weakly informative. On the relative-risk scale the  $N(0, 1^2)$  prior on  $\beta_0$  implies an approximate 95% prior range  $[0.14, 7.10]$ , which already excludes DZ-level relative risks larger than those reported in standard Scottish disease-mapping studies (typically well within  $[1/3, 3]$ ), so the prior is noticeably narrower than a truly non-informative choice. The same  $N(0, 1^2)$  prior on  $\beta_1$  is weakly informative on a standardised covariate, placing 95% of prior mass on per-SD relative-risk multipliers in  $[0.14, 7.10]$ . The half-normal prior  $\sigma_\phi \sim \text{Half-Normal}(0, 0.25^2)$  places 95% of its mass below approximately 0.5 on the log-risk scale, restricting the marginal spatial standard deviation of the log-risk to values that are empirically plausible for DZ-level disease mapping in Scotland. The logit-normal prior on  $\alpha$  concentrates posterior mass on moderate-to-strong temporal autocorrelation (approximate 95% prior CI  $[0.63, 0.99]$  on the probability scale). The Gamma prior on  $\delta$  has prior mean 40 and standard deviation 28, which accommodates moderate overdispersion without forcing the likelihood close to the Poisson limit.

Posterior inference is carried out via MCMC simulation in the R-NIMBLE software (de Valpine et al., 2017). The simulation studies run two chains per replicate to keep the per-replicate cost manageable across 50 replicates per scenario, while the motivating study runs three chains to give a more robust convergence assessment for a single, expensive fit. Point estimation for the DZ-level relative risks  $\theta_t(\mathcal{A}_m)$  uses the posterior median, and uncertainty is summarised by the 95% credible interval (CI), defined by the 2.5th and 97.5th posterior percentiles. Following Gelman and Rubin (1992), the potential scale reduction factor  $\hat{R}$  is used as a convergence diagnostic:  $\hat{R} < 1.05$  across scalar parameters and  $\hat{R} < 1.10$  across DZ-level random effects is treated as providing no evidence against non-convergence. The spatial random effects  $\phi_t$  are updated using a custom component-wise Metropolis–Hastings (CW-MH) sampler described in Appendix A. The scalar parameters are updated using automated factor slice samplers (Tibbits et al., 2014) or univariate slice samplers (Neal, 2003), with study-specific groupings described in the relevant model-specification

subsection.

### 5.3.4 Computational considerations

The GP–DZ registration weight matrix  $\eta = \{\eta_{km,t}\}$  is high-dimensional and inherently sparse, with many very small non-zero entries corresponding to pairs of GP surgeries and Data Zones with negligible patient overlap. In the NIMBLE implementation, retaining all non-zero entries requires the model to evaluate the aggregation sum  $\theta_t(\mathcal{S}_k) = \sum_{m=1}^M \eta_{km} \theta_t(\mathcal{A}_m)$  in Equation (5.2) over a large number of terms at every MCMC iteration, the vast majority of which contribute negligibly to the result. This increases the computational burden, raises memory requirements, and slows likelihood evaluations. To reduce this computational burden, I apply a sparsification threshold to the weight matrix. For each surgery  $\mathbf{s}_k$ , only Data Zones with  $\eta_{km} \geq \varepsilon$  are retained, and the remaining weights are re-normalised as

$$\eta_{km}^* = \frac{\eta_{km} \cdot \mathbb{I}(\eta_{km} \geq \varepsilon)}{\sum_{\ell=1}^M \eta_{k\ell} \cdot \mathbb{I}(\eta_{k\ell} \geq \varepsilon)},$$

so that  $\sum_{m=1}^M \eta_{km}^* = 1$  for each surgery  $\mathbf{s}_k$ . In the simulation studies the threshold is set to  $\varepsilon = 0.01$ , so that any Data Zone contributing less than 1% of a surgery’s registered patients is excluded from the aggregation, while the motivating study uses  $\varepsilon = 0.005$ . The simulation studies use the less aggressive  $\varepsilon = 0.01$  because they operate on the larger set of  $K = 260$  GP surgeries (the full set with any registered Glasgow patient), where the pre-sparsification linkage matrix is denser and the per-replicate computational cost becomes the binding constraint across 50 replicates per scenario. The motivating study uses  $K \approx 140$  surgeries only (after the 75% city-fraction filter), so the same linkage-pruning threshold discards a larger proportion of each surgery’s registration mass. With sparsification, each MCMC iteration would produce a sparser aggregation that makes the model computationally more efficient to fit.

Figure 5.3 summarises the empirical effect of sparsification on the registration matrix, evaluated on all  $K = 260$  surgeries with any registered Glasgow patient using the 2020 registration snapshot and  $\varepsilon = 0.01$ . Panel (A) shows that before sparsification a substantial number of surgeries have non-zero registrations spread across more than a hundred Data Zones, whereas after sparsification virtually every surgery is linked to fewer than 50. Panel (B) shows the distribution across surgeries of the share of each surgery’s total registration weight that is retained. The median retained share is 83% and 73% of surgeries retain at least 70% of their registration

weight, with a peak at 100% contributed by surgeries whose few pre-sparsification links all exceed the threshold.

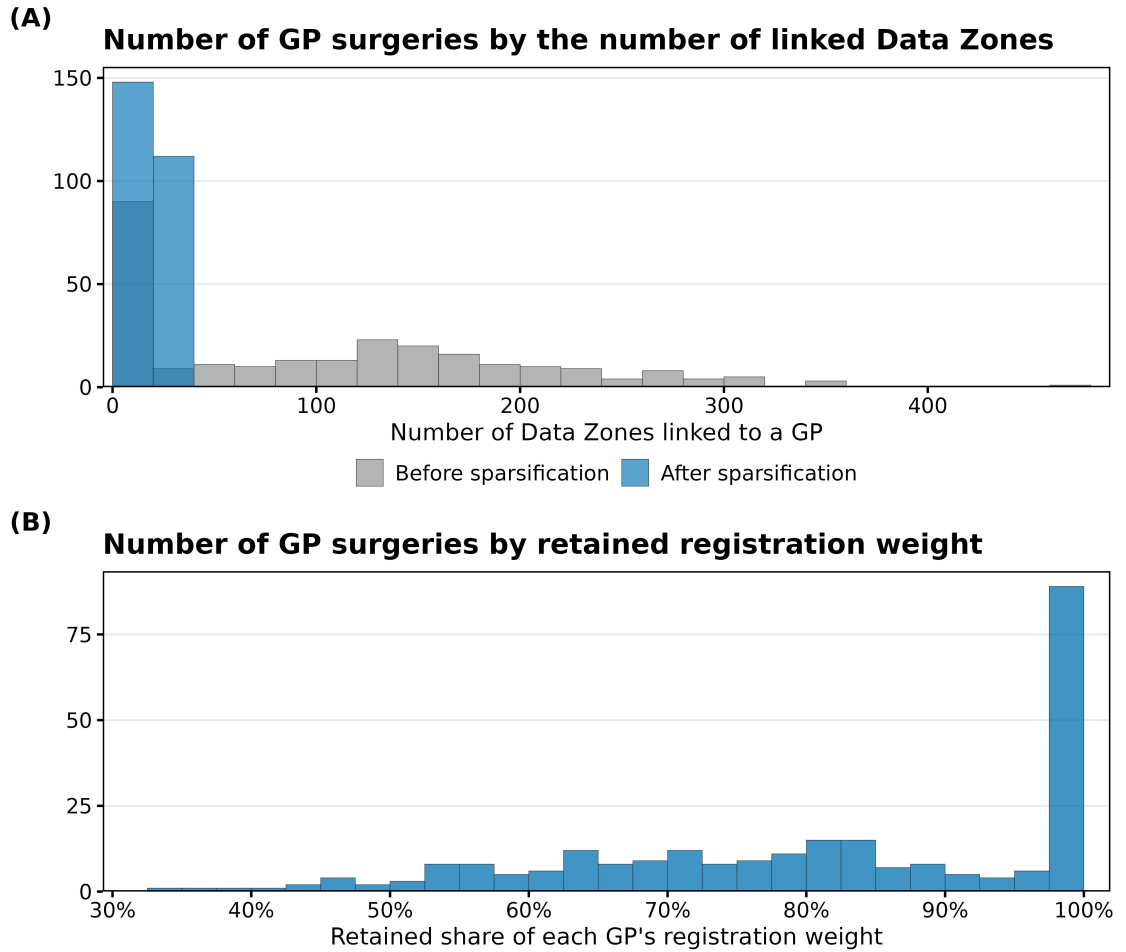


Figure 5.3: Registration-weight structure between GP surgeries and Data Zones, evaluated on the  $K = 260$  surgeries with any registered Glasgow patient using the 2020 registration snapshot and the sparsification threshold  $\varepsilon = 0.01$  used in the simulation studies. Panel (A) shows the distribution of the number of Data Zones with a non-zero registration weight for each surgery before (grey) and after (blue) applying the  $\eta_{km} \geq 0.01$  threshold. Panel (B) shows the distribution across surgeries of the share of each surgery's total registration weight that is retained after sparsification.

The proposed model is assessed through two complementary simulation studies before it is applied to the motivating data. Simulation study 1 (Section 5.4) isolates the purely spatial problem ( $T = 1$ ) and quantifies how accurately the DZ-level risk surface can be recovered from misaligned GP-level counts, and how this depends on the strength of spatial autocorrelation and on the presence of an informative covariate. Simulation study 2 (Section 5.5) then adds the temporal dimension, testing whether



the spatio-temporal ICAR–AR(1) prior recovers the risk surface when information is borrowed across both space and time. In both studies the realignment model is expected to outperform naive deterministic interpolation and to approach, but not match, an idealised benchmark fitted directly to DZ-level data, with the gap widening as the data become less informative (weaker autocorrelation, rarer outcomes, or no covariate).

## 5.4 Simulation study 1: spatial realignment

The first simulation study evaluates the proposed realignment model in a purely spatial setting, corresponding to the special case  $T = 1$  of the spatio-temporal framework described in Section 5.3, and hence the “ $t$ ” subscript is dropped. The aim is to establish a clear baseline understanding of whether and how much the model-based realignment improves upon naive deterministic interpolation – under different scenarios – and how close it comes to the idealised upper bound of fitting directly to DZ-level data (which would not be possible in practice as the DZ-level data are not known). All simulations are conducted on the  $M = 746$  Data Zones in Glasgow City, using the empirical GP–DZ registration matrix in 2020 to preserve a realistic misalignment structure.

### 5.4.1 Model specification

This subsection concerns the purely spatial ( $T = 1$ ) case of the framework in Section 5.3. The linear predictor in Setting 1 contains a single intercept  $\beta_0$  and no covariates, so that  $\log(\theta(\mathcal{A}_m)) = \beta_0 + \sigma_\phi \cdot \phi(\mathcal{A}_m)$ . Setting 2, introduced later in this section, adds a spatially structured DZ-level covariate  $x(\mathcal{A}_m)$  to evaluate how covariate inclusion changes realignment performance, with  $\log(\theta(\mathcal{A}_m)) = \beta_0 + \beta_1 \cdot x(\mathcal{A}_m) + \sigma_\phi \cdot \phi(\mathcal{A}_m)$ . The spatial random effects  $\boldsymbol{\phi} = (\phi(\mathcal{A}_1), \dots, \phi(\mathcal{A}_M))^\top$  follow an ICAR prior (Section 2.6.5) with precision matrix  $\mathbf{Q} = \text{diag}(\mathbf{W}\mathbf{1}) - \mathbf{W}$ , subject to the sum-to-zero constraint  $\sum_{m=1}^M \phi(\mathcal{A}_m) = 0$ .

Priors for  $\beta_0$ ,  $\beta_1$ ,  $\sigma_\phi$ , and  $\delta$  are as specified in Section 5.3.3. Posterior inference uses two parallel MCMC chains, each run for 50,000 iterations with the first 20,000 discarded as burn-in. The remaining 30,000 iterations are thinned every 10th draw, yielding 3,000 samples per chain (6,000 total). The spatial random effects  $\boldsymbol{\phi}$  are updated using the CW-MH sampler of Appendix A. The scalar pair  $(\beta_0, \sigma_\phi)$  is updated jointly using an automated factor slice sampler (Tibbits et al., 2014), which adapts



proposal directions to the local posterior correlation between the two parameters, while  $\delta$  is updated using a univariate slice sampler (Neal, 2003).

### 5.4.2 Competitor models

I compare the proposed realignment model (denoted *Realign*) against two alternatives that represent, respectively, a deterministic baseline that is straightforward to compute from the available data, and an idealised best-case modelling option only available in simulated data settings, which requires that DZ-level disease counts are observed.

**Naive population-weighted model (Naive).** This approach is the baseline comparator, applying a deterministic, population-weighted interpolation of GP-level standardised prescription ratios (SPRs) to the DZ scale. Specifically, the GP-level SPR is first computed as  $\text{SPR}(\mathcal{S}_k) = Y(\mathcal{S}_k)/e(\mathcal{S}_k)$ , and the DZ-level risk estimate is then obtained as

$$\widehat{\text{SPR}}(\mathcal{A}_m) = \sum_{k=1}^K \pi_{mk} \text{SPR}(\mathcal{S}_k),$$

where  $\pi_{mk} = p(\mathcal{A}_m \cap \mathcal{S}_k)/p(\mathcal{A}_m)$  is the proportion of patients in Data Zone  $\mathcal{A}_m$  registered at surgery  $\mathcal{S}_k$ , satisfying  $\sum_{k=1}^K \pi_{mk} = 1$  for each  $\mathcal{A}_m$ . Here  $\pi_{mk}$  denotes the DZ-to-GP direction of the registration weights, in contrast to  $\eta_{km,t}$  in the realignment model which operates in the GP-to-DZ direction. This approach implicitly assumes that the risk associated with each GP surgery is spatially uniform across all Data Zones it serves, thereby ignoring any within-catchment spatial heterogeneity in disease risk. Because it operates entirely on observed counts and does not involve a likelihood, uncertainty quantification under the Naive approach is not immediately available. It may be possible to quantify uncertainty using resampling techniques such as a non-parametric bootstrap. However, I consider this to be beyond the scope of this study, where I will focus on demonstrating the improvement of the realignment model over the Naive approach in point estimation accuracy.

**Direct DZ-level model (Benchmark).** This model assumes that the DZ-level counts  $Y(\mathcal{A}_m)$  are directly observed, bypassing the misalignment problem entirely. It fits a negative binomial log-linear model with ICAR spatial random effects directly to the

DZ-level data,

$$Y(\mathcal{A}_m) \sim \text{NegBin}(e(\mathcal{A}_m) \theta(\mathcal{A}_m), \delta),$$

$$\log(\theta(\mathcal{A}_m)) = \mathbf{x}^\top(\mathcal{A}_m) \boldsymbol{\beta} + \sigma_\phi \cdot \phi(\mathcal{A}_m),$$

with  $\boldsymbol{\phi} \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}^-)$ . Although this model cannot be applied to the motivating study, where DZ-level outcomes are not observed, it provides a gold standard against which the relative risk mapping accuracy lost from spatial misalignment can be quantified. The gap between *Benchmark* and *Realign* directly measures the inferential cost of observing outcomes at the GP level rather than the DZ level.

### 5.4.3 Data generation and simulation scenarios

Across all scenarios, data are generated in two stages: latent DZ-level risks and counts are first simulated, and these are then aggregated to the GP level using the empirical GP–DZ registration structure to create the spatially misaligned observations that are passed to the models.

**Stage 1: Latent risk generation.** True DZ-level relative risks  $\{\theta(\mathcal{A}_m)\}$  are generated on the log scale (detailed next). Expected counts  $\{e(\mathcal{A}_m)\}$  are drawn independently for each Data Zone from a uniform distribution whose range depends on the prevalence scenario (specified in the scenario descriptions below), and latent DZ-level prescription counts are then sampled as

$$Y(\mathcal{A}_m) \sim \text{Pois}(e(\mathcal{A}_m) \theta(\mathcal{A}_m)), \quad \text{for } m = 1, \dots, M.$$

**Stage 2: Aggregation to the GP level.** DZ-level counts are allocated to GP surgeries using the empirical registration proportions from the 2020 GP–DZ registration data, so that the simulated GP-level data retain the full misalignment structure before any sparsification is applied in the fitted model. For each Data Zone  $\mathcal{A}_m$ , the observed cases  $Y(\mathcal{A}_m)$  are distributed across surgeries via a multinomial draw,

$$\{Y(\mathcal{A}_m \cap \mathcal{S}_1), \dots, Y(\mathcal{A}_m \cap \mathcal{S}_K)\} \sim \text{Multi}(Y(\mathcal{A}_m) \mid \pi_{m1}, \dots, \pi_{mK}),$$

where  $\pi_{mk} = p(\mathcal{A}_m \cap \mathcal{S}_k)/p(\mathcal{A}_m)$  is the proportion of patients in  $\mathcal{A}_m$  registered at surgery  $\mathbf{s}_k$ . The observed GP-level count and expected count are then obtained by

summing over all Data Zones,

$$Y(\mathcal{S}_k) = \sum_{m=1}^M Y(\mathcal{A}_m \cap \mathcal{S}_k), \quad e(\mathcal{S}_k) = \sum_{m=1}^M \pi_{mk} e(\mathcal{A}_m),$$

where the expected count does not need a multinomial sampling step because it is a continuous quantity. This two-stage procedure ensures that the observed GP-level data are consistent with the underlying DZ-level process, and that the misalignment structure mirrors the motivating study.

### Setting 1: Spatial realignment without covariates

In this setting the log-relative risk is driven entirely by a latent spatial field, so that  $\log(\theta(\mathcal{A}_m)) = \phi(\mathcal{A}_m)$ . The intercept is fixed to  $\beta_0 = 0$  in the data-generating process so that the simulated relative risks are centred around the city-wide baseline level of one. Any non-zero baseline would be absorbed into the ICAR-constrained random effects and would not change the recovery behaviour assessed below. The spatial random effects  $\boldsymbol{\phi} = (\phi(\mathcal{A}_1), \dots, \phi(\mathcal{A}_M))^\top$  are drawn from a multivariate Gaussian distribution with a Matérn covariance function,

$$\text{Cov}(\phi(\mathcal{A}_i), \phi(\mathcal{A}_j)) = \sigma_\phi^2 \frac{(\kappa d_{ij})^\nu K_\nu(\kappa d_{ij})}{2^{\nu-1} \Gamma(\nu)},$$

where  $d_{ij}$  is the Euclidean distance between the centroids of  $\mathcal{A}_i$  and  $\mathcal{A}_j$ ,  $K_\nu(\cdot)$  is the modified Bessel function of order  $\nu$ , and  $\kappa > 0$  is a range parameter. The smoothness parameter is set to  $\nu = 1.5$ , which produces realisations that are once mean-square differentiable and therefore locally smooth rather than rough (Diggle and Ribeiro, 2007). The generated field is centred to have mean zero across all Data Zones. The range parameter  $\kappa$  and the marginal standard deviation  $\sigma_\phi$  together control the spatial structure of the latent risk surface, and are varied across scenarios as described below. Eight scenarios are considered in Setting 1, corresponding to all combinations of the following.

- **Disease prevalence.** Rare disease, with  $e(\mathcal{A}_m) \sim \text{Unif}(10, 30)$ , or common disease, with  $e(\mathcal{A}_m) \sim \text{Unif}(50, 150)$ . The precision with which DZ-level risks can be recovered depends on the volume of disease data available, so this comparison assesses such impact.
- **Spatial variability.** The marginal standard deviation  $\sigma_\phi \in \{0.2, 0.5\}$  controls

the overall variability of the log-risk surface, and is varied to cover the range of spatial heterogeneity. Here  $\sigma_\phi = 0.2$  corresponds to approximate 95% RR range  $[0.67, 1.49]$ , while  $\sigma_\phi = 0.5$  has respective  $[0.38, 2.66]$ .

- **Spatial correlation range.** The range parameter  $\kappa$  governs how quickly spatial correlations decay with distance. Here,  $\kappa = -\log(r)/1000$  for  $r \in \{0.5, 0.9\}$ , where  $r$  is the exponential decay rate per kilometre. A value of  $r = 0.5$  produces faster decay and shorter-range, reflecting more localised spatial variation. A value of  $r = 0.9$  produces a slowly decaying correlation structure with long-range spatial dependence, reflecting more diffuse, large-scale spatial variation.

### Setting 2: Spatial realignment with a structured covariate

This setting investigates whether including an informative DZ-level covariate in the realignment model enables a greater improvement in the recovery of the latent risk surface under misalignment relative to the Naïve approach. A spatially structured covariate  $\mathbf{x} = (x(\mathcal{A}_1), \dots, x(\mathcal{A}_M))^T$  is generated with strong spatial dependence via the CAR correlation structure specified by Leroux ([Leroux et al., 2000](#)),

$$\mathbf{x} \sim N(\mathbf{0}, \mathbf{Q}_x^{-1}), \quad \mathbf{Q}_x = \rho_x (\text{diag}(\mathbf{W}\mathbf{1}) - \mathbf{W}) + (1 - \rho_x)\mathbf{I},$$

with spatial dependence parameter  $\rho_x = 0.95$ . The covariate is standardised to have mean zero and unit variance across Data Zones. The log-relative risk surface is then constructed as a linear combination of the covariate and a residual Matérn spatial field,

$$\log(\theta) = \sigma \left( \sqrt{p} \cdot \mathbf{x} + \sqrt{1-p} \cdot \boldsymbol{\phi} \right) = \beta \mathbf{x} + \boldsymbol{\phi}^*, \quad (5.6)$$

where  $\sigma^2$  is the total variance of the log-risk surface,  $p \in (0, 1)$  is the proportion of this variance attributable to the covariate,  $\beta = \sigma\sqrt{p}$  is the implied regression coefficient, and  $\boldsymbol{\phi}^* = \sigma\sqrt{1-p} \cdot \boldsymbol{\phi}$  is the residual spatial random effect. This construction allows the relative contribution of the fixed and random components to be varied across scenarios while keeping the total variance of the log-risk surface fixed at  $\sigma^2$  (to avoid confounding conclusions). Twelve scenarios are considered in Setting 2, corresponding to all combinations of:

- **Disease prevalence.** Rare or common disease, as defined in Setting 1.
- **Total log-risk variability.**  $\sigma \in \{0.2, 0.5\}$ , representing surfaces with lower and higher overall variation in disease risk.

- **Variance partitioning:**  $p \in \{0.25, 0.5, 0.75\}$ , representing settings where the covariate explains a low, moderate, or high proportion of the total spatial variation in log risk.

For each scenario in both settings, 50 replicate datasets are generated to ensure stable estimates of all performance metrics reported in Section 5.4.5.

#### 5.4.4 Performance metrics

Model performance is assessed by comparing posterior summaries of the DZ-level relative risks  $\{\theta(\mathcal{A}_m)\}$  against the known true values across all 50 replicate datasets and all  $M = 746$  Data Zones. Six metrics are used, covering point estimation accuracy, systematic bias, and uncertainty quantification. For the Naive model, only point estimate metrics are reported, since it is a purely deterministic interpolation (Section 5.4.2) that does not produce a posterior distribution. Let  $\bar{\hat{\theta}}$  and  $\bar{\theta}$  denote means of  $\{\hat{\theta}(\mathcal{A}_m)\}$  and  $\{\theta(\mathcal{A}_m)\}$  over the  $M$  Data Zones, respectively.

**Pearson correlation (Corr)** measures the linear association between the posterior median estimates  $\{\hat{\theta}(\mathcal{A}_m)\}$  and the true risks  $\{\theta(\mathcal{A}_m)\}$ , with values closer to one indicating better agreement.

$$\text{Corr} = \frac{\sum_{m=1}^M (\hat{\theta}(\mathcal{A}_m) - \bar{\hat{\theta}})(\theta(\mathcal{A}_m) - \bar{\theta})}{\sqrt{\sum_{m=1}^M (\hat{\theta}(\mathcal{A}_m) - \bar{\hat{\theta}})^2 \sum_{m=1}^M (\theta(\mathcal{A}_m) - \bar{\theta})^2}}.$$

**Root mean squared error (RMSE)** measures the average magnitude of estimation error in the posterior medians, with larger values indicating poorer accuracy.

$$\text{RMSE} = \sqrt{\frac{1}{M} \sum_{m=1}^M (\hat{\theta}(\mathcal{A}_m) - \theta(\mathcal{A}_m))^2}.$$

**Median absolute error (MAE)** is the median of the absolute differences  $|\hat{\theta}(\mathcal{A}_m) - \theta(\mathcal{A}_m)|$  across all Data Zones. MAE is less sensitive than RMSE to a small number of extreme estimation errors and therefore provides a complementary summary of typical pointwise accuracy.

$$\text{MAE} = \text{median}_{m=1, \dots, M} |\hat{\theta}(\mathcal{A}_m) - \theta(\mathcal{A}_m)|.$$

**Bias** is the mean absolute error  $\{\hat{\theta}(\mathcal{A}_m) - \theta(\mathcal{A}_m)\}$  across Data Zones. Values close to zero indicate that the model neither systematically over- nor under-estimates DZ-level risks on average.

$$\text{Bias} = \frac{1}{M} \sum_{m=1}^M (\hat{\theta}(\mathcal{A}_m) - \theta(\mathcal{A}_m)).$$

**Coverage probability (CP)** is the proportion of true DZ-level risks  $\{\theta(\mathcal{A}_m)\}$  that fall within the estimated 95% credible intervals. A well-calibrated Bayesian model should achieve CP close to 0.95, where substantial under-coverage indicates over-confident inference (insufficient posterior variance).

$$\text{CP} = \frac{1}{M} \sum_{m=1}^M \mathbb{I}(\theta(\mathcal{A}_m) \in [L_m, U_m]),$$

where  $L_m$  and  $U_m$  are the 2.5th and 97.5th posterior percentiles of  $\theta(\mathcal{A}_m)$ .

**Mean interval width (MIW)** is the average width of the 95% credible intervals across all Data Zones. When interpreted alongside CP, MIW summarises the precision of the uncertainty estimates. Where coverage is satisfactory, narrower intervals indicate more precise inference.

$$\text{MIW} = \frac{1}{M} \sum_{m=1}^M (U_m - L_m).$$

In the spatio-temporal simulation study (Section 5.5), all metrics are computed over the  $M \times T$  space–time index pairs, with  $M$  replaced by  $M \times T$  in the expressions above. All six metrics are averaged over the 50 replicate datasets. Individual parameter estimates such as  $\hat{\sigma}$  and  $\hat{\beta}$  are not reported, because the parameterisation of the fitted models (which operates on log-risks with an ICAR spatial prior) does not correspond directly to the parameterisation used to generate the data (which uses a Matérn covariance on the log-risk scale). The focus throughout is on the quality of the recovered DZ-level risk surface, which is the primary inferential target of this study.

### 5.4.5 Results of simulation study 1

#### Setting 1 (no covariate)

Table 5.1 reports the mean performance metrics averaged over 50 replicate datasets, and Figure 5.4 displays the between-replicate variation in RMSE to show the level of within-scenario variability in the results. All three models produce near-zero mean bias across the eight scenarios (range  $-0.0284$  to  $0.0017$ ), and the Rare versus Common prevalence distinction has negligible effect on any metric. The Realign model exhibits the largest negative bias under high-variance, short-range surfaces ( $\sigma_\phi = 0.5$ ,  $r = 0.5$ ), reaching  $-0.0266$  (Common) and  $-0.0284$  (Rare). This arises from the combination of the model generally underestimating the full variance of log relative risks and the concavity of the exponential back-transformation, which compresses the right tail of the risk distribution when smoothing is applied on the log scale. Figure 5.5 illustrates this mechanism for a single replicate.

Under three point estimation metrics (Corr, RMSE, and MAE) the general ranking is Benchmark  $>$  Realign  $>$  Naive, with Realign achieving 2%–27% lower RMSE than Naive across seven of the eight scenarios. The smallest improvement (2.4%) occurs under the lowest-variance surface ( $\sigma_\phi = 0.2$ ,  $r = 0.5$ , Rare), while the largest improvements (22%–27%) occur under high-variance Common scenarios. Figure 5.4 confirms this ranking at replicate level, with Benchmark showing the narrowest spread and Realign widening beyond Naive only at  $\sigma_\phi = 0.5$ ,  $r = 0.5$ .

The coverage probabilities of the Realign model range from 0.9588 to 0.9963, uniformly above the nominal 95% level (Table 5.1). This under-confidence reflects the high posterior uncertainty inherent in estimating risk at a finer spatial resolution ( $M = 746$  Data Zones) than the data support ( $K = 260$  GP surgeries). The Benchmark model achieves similar coverage (0.9799–0.9965) with consistently narrower intervals.

#### Setting 2 (with spatially structured covariate)

Table 5.2 and Figure 5.6 report the Setting 2 results. The inclusion of a DZ-level covariate substantially widens the gap between Realign and Naive. When GP-level rates are obtained by aggregating over the DZs served by each surgery, the spatial structure of the covariate is smoothed out. The Naive model works only with GP-level rates and therefore cannot recover this lost structure, whereas Realign uses the covariate directly in its linear predictor at the DZ level. As  $p$  (the proportion of variation explained by the covariate) increases, the covariate also absorbs a greater share of the log-risk variation, reducing the residual random-field amplitude and the

back-transformation bias identified in Setting 1. Under  $\sigma = 0.5$ , Common prevalence, the Realign mean bias falls from  $-0.0106$  at  $p = 0.25$  to  $-0.0070$  at  $p = 0.75$ .

The correlation achieved by the Naive model drops from  $0.8400$ – $0.8700$  at  $p = 0.25$  to  $0.6374$ – $0.6818$  at  $p = 0.75$ , while the corresponding results for the Realign model remain above  $0.93$  across all twelve scenarios, with a minimum of  $0.9333$  at  $\sigma_\phi = 0.2$ ,  $p = 0.25$ , Rare. The RMSE reduction of the Realign model relative to Naive grows from  $33\%$ – $52\%$  at  $p = 0.25$  to  $53\%$ – $60\%$  at  $p = 0.75$ , far exceeding the  $2\%$ – $27\%$  observed in Setting 1. Figure 5.6 shows that Realign’s performance is stable across  $p$ , whereas Naive’s RMSE rises with  $p$  even though the total log-risk variance is fixed at  $\sigma_\phi^2$ . The increase arises because the Leroux covariate  $x(\mathcal{A}_m)$  and the Matérn residual field  $\phi(\mathcal{A}_m)$  have different smoothness at the DZ scale. The Matérn field with  $\nu = 1.5$  is visibly smooth, so neighbouring Data Zones within a GP catchment take similar values. The Leroux field is more heterogeneous, with more visible roughness between adjacent Data Zones even under strong spatial dependence. As  $p$  increases, this roughness is more prominent and adversely affects estimation.

Coverage probabilities of the Realign model become less conservative as  $p$  increases, from  $0.9815$ – $0.9929$  at  $p = 0.25$  to  $0.9149$ – $0.9649$  at  $p = 0.75$  (Table 5.2). The corresponding MIW decreases from  $0.7610$  to  $0.5624$  (under  $\sigma_\phi = 0.5$ , Common), reflecting plausible precision gains. At  $\sigma_\phi = 0.5$ ,  $p = 0.75$ , CP falls slightly below the nominal level ( $0.9149$  Common,  $0.9396$  Rare), suggesting that the credible intervals marginally underestimate residual variability when the covariate dominates the log-risk surface.

## 5.5 Simulation study 2: spatio-temporal realignment

Simulation study 1 showed that the proposed realignment model outperforms deterministic interpolation in a purely spatial setting, with the degree of improvement depending on the spatial autocorrelation range, the marginal variability of the latent risk surface, and the availability of an informative covariate. The next study extends this evaluation to a full spatio-temporal setting, where the DZ-level risk surface evolves over  $T = 12$  monthly time points. The study evaluates whether the AR(1) temporal structure in the proposed model tracks different forms of temporal evolution in the latent risk surface. The simulation setup uses the same  $K = 260$  GP surgeries as simulation study 1, so that the spatial misalignment structure is held fixed across studies, and introduces three scenarios for the temporal evolution of  $\phi_t$  described below.



Table 5.1: Mean performance metrics across 50 replicates for Simulation Study 1, Setting 1 (no covariates). Metrics are averaged over all  $M = 746$  Data Zones. CP and MIW are not applicable for Naive as it does not produce posterior distributions. The rows compare the Naive model (deterministic population-weighted interpolation), Realign (the proposed change-of-support model), and Benchmark (an idealised model fitted directly to DZ-level data). Prevalence is either Common or Rare.

| $\sigma_\phi$ | $r$ | Prevalence | Model     | Corr   | RMSE   | MAE    | Bias    | CP     | MIW    |
|---------------|-----|------------|-----------|--------|--------|--------|---------|--------|--------|
| 0.2           | 0.5 | Common     | Naive     | 0.8456 | 0.1015 | 0.0653 | 0.0000  | –      | –      |
|               |     |            | Realign   | 0.8844 | 0.0865 | 0.0544 | –0.0069 | 0.9702 | 0.4236 |
|               |     |            | Benchmark | 0.9559 | 0.0520 | 0.0332 | –0.0021 | 0.9838 | 0.2616 |
|               |     | Rare       | Naive     | 0.7946 | 0.1071 | 0.0690 | –0.0009 | –      | –      |
|               |     |            | Realign   | 0.8062 | 0.1045 | 0.0666 | –0.0072 | 0.9588 | 0.4578 |
|               |     |            | Benchmark | 0.8978 | 0.0761 | 0.0488 | –0.0056 | 0.9830 | 0.3835 |
|               | 0.9 | Common     | Naive     | 0.9307 | 0.0271 | 0.0171 | 0.0012  | –      | –      |
|               |     |            | Realign   | 0.9412 | 0.0251 | 0.0158 | 0.0003  | 0.9963 | 0.1836 |
|               |     |            | Benchmark | 0.9607 | 0.0204 | 0.0131 | 0.0007  | 0.9965 | 0.1335 |
|               |     | Rare       | Naive     | 0.8155 | 0.0445 | 0.0293 | 0.0017  | –      | –      |
|               |     |            | Realign   | 0.8282 | 0.0391 | 0.0263 | 0.0009  | 0.9621 | 0.2294 |
|               |     |            | Benchmark | 0.8831 | 0.0334 | 0.0226 | 0.0005  | 0.9804 | 0.1978 |
| 0.5           | 0.5 | Common     | Naive     | 0.8498 | 0.2847 | 0.1677 | 0.0005  | –      | –      |
|               |     |            | Realign   | 0.9196 | 0.2091 | 0.1085 | –0.0266 | 0.9733 | 0.9972 |
|               |     |            | Benchmark | 0.9854 | 0.0830 | 0.0483 | –0.0034 | 0.9804 | 0.3878 |
|               |     | Rare       | Naive     | 0.8416 | 0.2824 | 0.1645 | –0.0006 | –      | –      |
|               |     |            | Realign   | 0.8964 | 0.2290 | 0.1243 | –0.0284 | 0.9729 | 1.0603 |
|               |     |            | Benchmark | 0.9609 | 0.1348 | 0.0758 | –0.0109 | 0.9799 | 0.6202 |
|               | 0.9 | Common     | Naive     | 0.9563 | 0.0595 | 0.0347 | 0.0003  | –      | –      |
|               |     |            | Realign   | 0.9680 | 0.0464 | 0.0271 | –0.0023 | 0.9956 | 0.3642 |
|               |     |            | Benchmark | 0.9823 | 0.0341 | 0.0211 | –0.0010 | 0.9944 | 0.2069 |
|               |     | Rare       | Naive     | 0.9214 | 0.0688 | 0.0425 | –0.0002 | –      | –      |
|               |     |            | Realign   | 0.9380 | 0.0626 | 0.0382 | –0.0056 | 0.9910 | 0.4418 |
|               |     |            | Benchmark | 0.9562 | 0.0520 | 0.0322 | –0.0035 | 0.9934 | 0.3245 |

### 5.5.1 Model specification

This subsection extends simulation study 1 to the full spatio-temporal setting ( $T = 12$ ). The ICAR-AR(1) specification for  $\phi_t$  is given in Equation (5.5) with the sum-to-zero constraint imposed at  $t = 1$  only. The linear predictor includes an intercept  $\beta_0$  and a single covariate coefficient  $\beta_1$ , so that  $\log(\theta_t(\mathcal{A}_m)) = \beta_0 + \beta_1 \cdot x(\mathcal{A}_m) + \sigma_\phi \cdot \phi_t(\mathcal{A}_m)$ .

Priors for  $\beta_0$ ,  $\beta_1$ ,  $\sigma_\phi$ ,  $\alpha$  and  $\delta$  are as specified in Section 5.3.3. MCMC settings are unchanged from simulation study 1. The automated factor slice sampler of Tibbits et al. (2014) is now applied jointly to  $(\beta_0, \sigma_\phi, \alpha)$  to adapt to the three-way posterior correlation introduced by the temporal autocorrelation parameter, while  $\beta_1$  and  $\delta$

Table 5.2: Mean performance metrics across 50 replicates for Simulation Study 1, Setting 2 (with spatially structured covariate). Metrics are averaged (mean) over all  $M = 746$  Data Zones. CP and MIW are not applicable for Naive as it does not produce posterior distributions. The rows compare the Naive model (deterministic population-weighted interpolation), Realign (the proposed change-of-support model), and Benchmark (an idealised model fitted directly to DZ-level data).

| $\sigma_\phi$ | $p$  | Prevalence | Model     | Corr   | RMSE   | MAE    | Bias    | CP     | MIW    |
|---------------|------|------------|-----------|--------|--------|--------|---------|--------|--------|
| 0.2           | 0.25 | Common     | Naive     | 0.8700 | 0.1035 | 0.0656 | 0.0002  | –      | –      |
|               |      |            | Realign   | 0.9657 | 0.0560 | 0.0341 | –0.0021 | 0.9916 | 0.3406 |
|               |      |            | Benchmark | 0.9871 | 0.0329 | 0.0203 | –0.0010 | 0.9946 | 0.2007 |
|               |      | Rare       | Naive     | 0.8400 | 0.1107 | 0.0697 | 0.0005  | –      | –      |
|               |      |            | Realign   | 0.9333 | 0.0741 | 0.0450 | –0.0039 | 0.9929 | 0.4401 |
|               |      |            | Benchmark | 0.9672 | 0.0521 | 0.0326 | –0.0027 | 0.9958 | 0.3228 |
|               | 0.50 | Common     | Naive     | 0.7844 | 0.1295 | 0.0826 | 0.0013  | –      | –      |
|               |      |            | Realign   | 0.9663 | 0.0593 | 0.0371 | –0.0015 | 0.9872 | 0.3189 |
|               |      |            | Benchmark | 0.9884 | 0.0312 | 0.0193 | 0.0003  | 0.9943 | 0.1925 |
|               |      | Rare       | Naive     | 0.7566 | 0.1312 | 0.0836 | 0.0009  | –      | –      |
|               |      |            | Realign   | 0.9494 | 0.0684 | 0.0435 | –0.0006 | 0.9790 | 0.3589 |
|               |      |            | Benchmark | 0.9756 | 0.0450 | 0.0275 | –0.0016 | 0.9951 | 0.2800 |
|               | 0.75 | Common     | Naive     | 0.6683 | 0.1533 | 0.0969 | –0.0003 | –      | –      |
|               |      |            | Realign   | 0.9736 | 0.0614 | 0.0379 | –0.0021 | 0.9649 | 0.2596 |
|               |      |            | Benchmark | 0.9911 | 0.0273 | 0.0165 | –0.0013 | 0.9927 | 0.1652 |
|               |      | Rare       | Naive     | 0.6577 | 0.1548 | 0.0994 | –0.0011 | –      | –      |
|               |      |            | Realign   | 0.9619 | 0.0722 | 0.0460 | –0.0026 | 0.9330 | 0.2920 |
|               |      |            | Benchmark | 0.9804 | 0.0418 | 0.0258 | –0.0028 | 0.9898 | 0.2409 |
| 0.5           | 0.25 | Common     | Naive     | 0.8572 | 0.3198 | 0.1630 | –0.0001 | –      | –      |
|               |      |            | Realign   | 0.9711 | 0.1551 | 0.0717 | –0.0106 | 0.9815 | 0.7610 |
|               |      |            | Benchmark | 0.9946 | 0.0609 | 0.0317 | –0.0026 | 0.9908 | 0.3113 |
|               |      | Rare       | Naive     | 0.8559 | 0.3096 | 0.1594 | 0.0006  | –      | –      |
|               |      |            | Realign   | 0.9640 | 0.1635 | 0.0781 | –0.0155 | 0.9906 | 0.8932 |
|               |      |            | Benchmark | 0.9867 | 0.0936 | 0.0484 | –0.0058 | 0.9933 | 0.4999 |
|               | 0.50 | Common     | Naive     | 0.7454 | 0.3825 | 0.2125 | –0.0005 | –      | –      |
|               |      |            | Realign   | 0.9661 | 0.1720 | 0.0854 | –0.0116 | 0.9575 | 0.6580 |
|               |      |            | Benchmark | 0.9953 | 0.0539 | 0.0282 | –0.0020 | 0.9920 | 0.2848 |
|               |      | Rare       | Naive     | 0.7586 | 0.3837 | 0.2102 | 0.0018  | –      | –      |
|               |      |            | Realign   | 0.9626 | 0.1763 | 0.0885 | –0.0075 | 0.9734 | 0.7719 |
|               |      |            | Benchmark | 0.9893 | 0.0843 | 0.0432 | –0.0036 | 0.9935 | 0.4580 |
|               | 0.75 | Common     | Naive     | 0.6818 | 0.4536 | 0.2476 | –0.0011 | –      | –      |
|               |      |            | Realign   | 0.9733 | 0.1805 | 0.0897 | –0.0070 | 0.9149 | 0.5624 |
|               |      |            | Benchmark | 0.9970 | 0.0470 | 0.0237 | –0.0028 | 0.9941 | 0.2563 |
|               |      | Rare       | Naive     | 0.6374 | 0.4558 | 0.2521 | –0.0022 | –      | –      |
|               |      |            | Realign   | 0.9690 | 0.1873 | 0.0925 | –0.0123 | 0.9396 | 0.6268 |
|               |      |            | Benchmark | 0.9924 | 0.0720 | 0.0364 | –0.0062 | 0.9936 | 0.3941 |

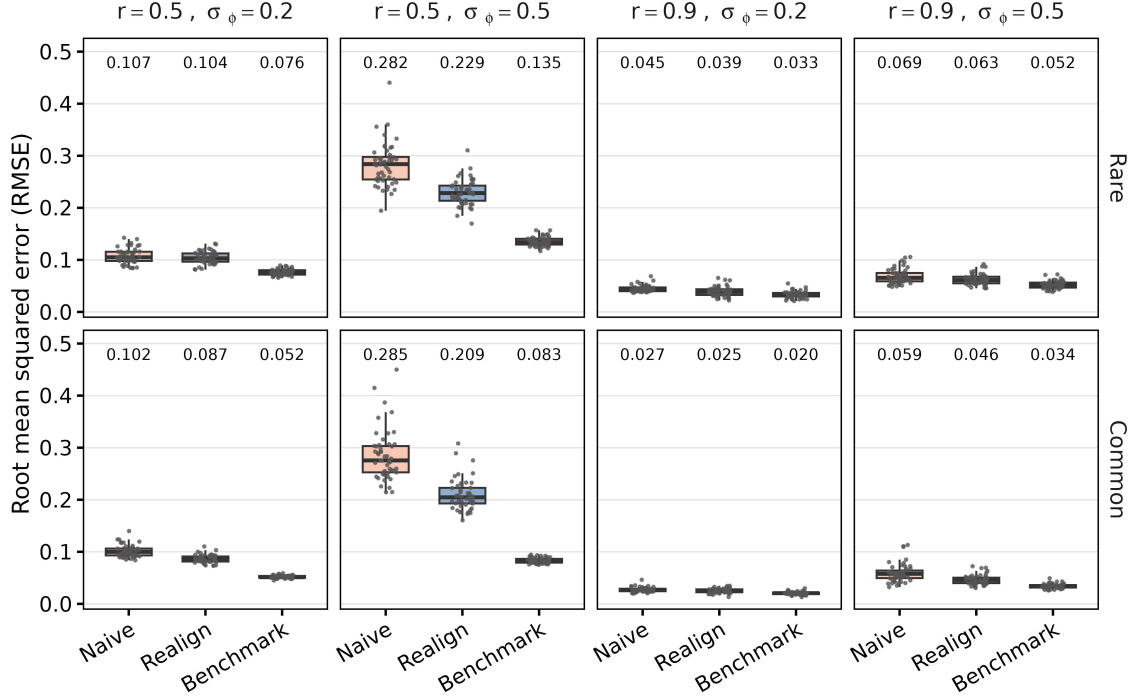


Figure 5.4: Between-replicate distribution of root mean squared error (RMSE) of posterior median DZ-level relative risks, Simulation Study 1 Setting 1 (no covariate). Panels are arranged by spatial correlation range ( $r$ ) and marginal variability ( $\sigma_\phi$ ), with rows separating Rare and Common disease prevalence.

are updated with univariate slice samplers.

### 5.5.2 Competitor models

The proposed spatio-temporal realignment model (*Realign*) is compared against three alternatives representing, respectively, two deterministic baselines and an idealised best-case scenario available only in simulation.

**Naive model for the current month only (Naive0).** Separately for each month  $t$ , the GP-level standardised prescription ratio is computed as  $\text{SPR}_t(\mathcal{S}_k) = Y_t(\mathcal{S}_k)/e_t(\mathcal{S}_k)$  and the DZ-level risk estimate is obtained by population weighted interpolation,

$$\hat{\theta}_t(\mathcal{A}_m) = \sum_{k=1}^K \pi_{mk,t} \text{SPR}_t(\mathcal{S}_k),$$

where  $\pi_{mk,t} = p_t(\mathcal{A}_m \cap \mathcal{S}_k)/p_t(\mathcal{A}_m)$  satisfies  $\sum_{k=1}^K \pi_{mk,t} = 1$  for each  $\mathcal{A}_m$ . This approach is purely deterministic and month-specific, making no use of information from

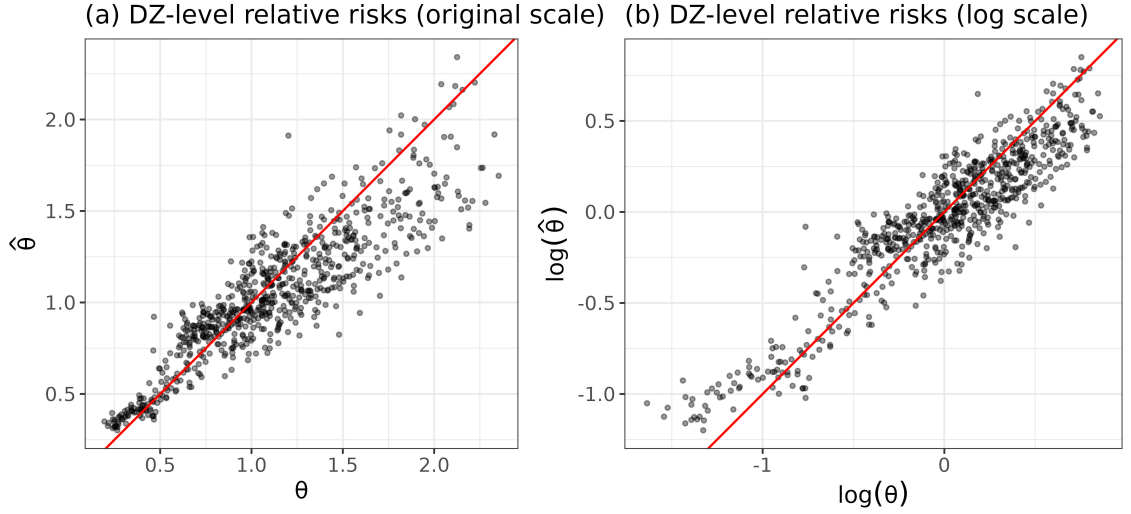


Figure 5.5: Scatter plot of true simulated versus estimated DZ-level relative risks under a rougher ( $r = 0.5$ ), higher-variance ( $\sigma_\phi = 0.5$ ) spatial surface. Panel (A) shows relative risks on the original scale and Panel (B) shows the same comparison on the log scale. The mean bias of the posterior median estimates across the 746 Data Zones for this replicate is  $-0.07$ .

neighbouring time points.

**Naive model with temporal smoothing (Naive1).** To partially stabilise month-to-month variability, Naive0 is applied at each time point and the resulting DZ-level time series is then smoothed using a symmetric three-point moving average,

$$\hat{\theta}_t^*(\mathcal{A}_m) = \frac{1}{3} \sum_{j=-1}^1 \hat{\theta}_{t+j}(\mathcal{A}_m) = \frac{1}{3} \sum_{j=-1}^1 \sum_{k=1}^K \pi_{mk,t+j} \text{SPR}_{t+j}(S_k), \quad t = 2, \dots, T-1,$$

with boundary months handled by averaging over the available neighbouring time points only, using  $\{t, t+1\}$  at  $t = 1$  and  $\{t-1, t\}$  at  $t = T$ . Like Naive0, this model does not produce posterior distributions and therefore provides no basis for uncertainty quantification.

**Direct DZ-level model (Benchmark).** As in simulation study 1, this model assumes DZ-level counts are directly observed at each time point, and therefore bypasses the spatial misalignment problem. Recall that although it cannot be fitted to the motivating GP-level prescription data, it provides a reference for the best achievable performance when outcomes are available on the target spatial support. I adopt the

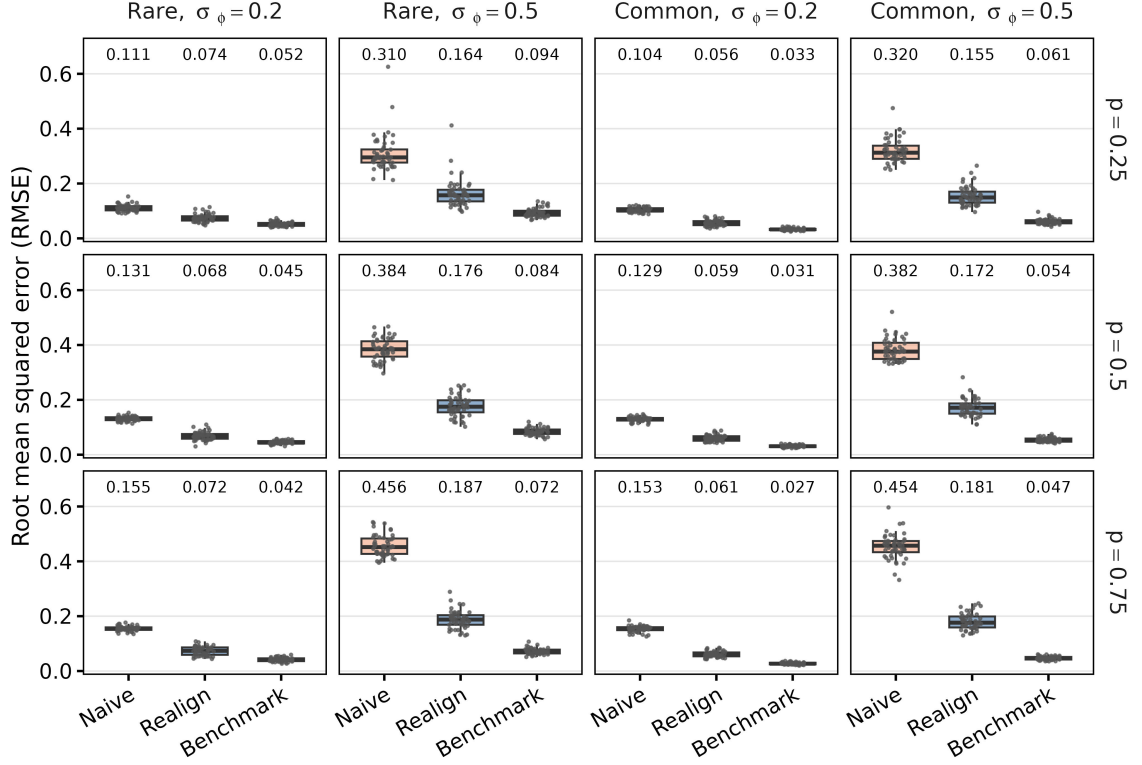


Figure 5.6: Between-replicate distribution of RMSE of posterior median DZ-level relative risks, Simulation Study 1 Setting 2 (with spatially structured covariate). Panels are arranged by covariate variance share  $p$  and marginal variability  $\sigma_\phi$ , with rows separating Rare and Common disease prevalence.

spatio-temporal disease mapping model of [Rushworth et al. \(2014\)](#), in which DZ-level counts follow a negative binomial log-linear model with a dynamic ICAR-AR(1) latent structure. Specifically,

$$Y_t(\mathcal{A}_m) \sim \text{NegBin}(e_t(\mathcal{A}_m)\theta_t(\mathcal{A}_m), \delta),$$

$$\log(\theta_t(\mathcal{A}_m)) = \mathbf{x}^\top(\mathcal{A}_m)\boldsymbol{\beta} + \sigma_\phi \cdot \phi_t(\mathcal{A}_m).$$

Here  $\boldsymbol{\phi}_t = (\phi_t(\mathcal{A}_1), \dots, \phi_t(\mathcal{A}_M))^\top$  is a spatial random effect vector at time  $t$ , with ICAR smoothing in space and AR(1) dependence in time (as specified in Equation (5.5)).

### 5.5.3 Data generation and simulation scenarios

The simulation is conducted on the  $M = 746$  Data Zones (DZs) in Glasgow City using the empirical GP-DZ registration matrix, following the same two-stage data generation procedure as for simulation study 1 (Section 5.4.3). The principal extension is

the introduction of temporal dynamics, in which the latent DZ-level log risk surface now evolves over  $T = 12$  months under one of three temporal scenarios described below. Expected counts are held constant across months,  $e(\mathcal{A}_m) \sim \text{Unif}(50, 150)$ , to ensure that any temporal variation in the model performance only reflects evolution in the latent risk surface.

As in simulation study 1, the DZ-level log relative risk is decomposed as

$$\log(\theta_t(\mathcal{A}_m)) = \sigma \left( \sqrt{p} \cdot x^*(\mathcal{A}_m) + \sqrt{1-p} \cdot \phi_t^*(\mathcal{A}_m) \right), \quad m = 1, \dots, M, \quad t = 1, \dots, T,$$

where  $\sigma_\phi = 0.3$  controls the overall variability of the log risk surface and  $p \in \{0.25, 0.75\}$  controls the proportion of that variability attributable to the covariate relative to the residual spatio-temporal random field.

Note that the approach to data generation for Simulation study 2 was partly informed by lessons learned from the results for Simulation study 1. I found that the performance of the Naïve approach worsened as the covariate contribution to the variability of the log risk increased, because GP-level aggregation averages over the DZ-level covariate structure and the Naïve approach has no way to recover the attenuated fine-scale variation. To provide a stable baseline against which improvement in accuracy achieved by the Realign model when an informative covariate is available can be addressed, I kept the expected spatial characteristics of the covariate and the unobserved random effect the same. Specifically, both the covariate and the baseline spatial surface are now generated from mean zero Gaussian processes with the same Matérn covariance structure, smoothness parameter  $\nu = 0.5$ , and range parameter  $\kappa = -\log(0.9)/1000$ . The smoothness  $\nu = 0.5$  corresponds to an exponential covariance function, producing realisations that are continuous but not mean-square differentiable (Cressie, 2015). The time-invariant covariate  $x^*(\mathcal{A}_m)$  is standardised to unit variance across Data Zones by

$$x^*(\mathcal{A}_m) = \frac{x(\mathcal{A}_m) - \bar{x}}{\text{sd}_m\{x(\mathcal{A}_m)\}},$$

where  $\bar{x} = \frac{1}{M} \sum_{m=1}^M x(\mathcal{A}_m)$  is the mean of the covariate across Data Zones.

The baseline surface  $\phi_1$  is mean-centred across Data Zones prior to temporal evolution. A single global standardisation is applied to the full spatio-temporal field after evolution,

$$\phi_t^*(\mathcal{A}_m) = \frac{\phi_t(\mathcal{A}_m) - \bar{\phi}}{\text{sd}_{t,m}\{\phi_t(\mathcal{A}_m)\}},$$

where  $\bar{\phi}$  and  $\text{sd}_{t,m}$  are computed over all  $M \times T$  values of the field. This ensures that  $p$  represents the variance partition averaged over the entire study period rather than at any single time slice, and prevents drift in the effective contribution of the residual field under scenarios where its marginal variance changes over time. The temporal evolution of  $\phi_t$  for  $t = 2, \dots, T$  is imposed under one of the following three scenarios:

**Scenario T1: Global linear trend.** This scenario represents a spatially uniform monotonic decrease in log risk over time, with the spatial pattern of the baseline surface otherwise unchanged:

$$\phi_t = \phi_1 - 0.5 \cdot \frac{t-1}{T-1}, \quad t = 2, \dots, T.$$

The coefficient 0.5 controls the magnitude of the decrease across the study period. Because the spatial structure of  $\phi_1$  is preserved at every time point, this scenario tests whether the model can track a global temporal trend in the overall level of disease risk without confounding it with changes in spatial structure.

**Scenario T2: AR(1) temporal persistence.** This scenario represents gradual month-to-month changes while preserving the broad spatial structure of the surface from one time point to the next. The field evolves as

$$\phi_t = 0.95 \cdot \phi_{t-1} + \sqrt{1 - 0.95^2} \cdot \mathbf{u}_t, \quad t = 2, \dots, T,$$

where  $\mathbf{u}_t \sim N(\mathbf{0}, \Sigma_u)$  is an independent spatial innovation with the same Matérn covariance structure and unit marginal variance as  $\phi_1$ . This construction – with an autoregressive coefficient of 0.95 – yields a stationary sequence of spatial surfaces with strong temporal autocorrelation.

**Scenario T3: Inequality widening.** This scenario represents a progressive divergence in disease risk across Data Zones, where areas with above-average baseline risk become increasingly elevated and areas with below-average risk become relatively lower over time,

$$\phi_t = \left\{ 1 + 0.5 \cdot \frac{t-1}{T-1} \right\} \cdot \phi_1, \quad t = 2, \dots, T.$$

The coefficient 0.5 controls the rate of divergence, so that by the final month the amplitude of the spatial field is 50% larger than at baseline. Because  $\phi_1$  is mean-

centred, this multiplicative scaling increases the marginal variability of  $\phi_t$  over time without shifting its cross-sectional mean, producing a mean preserving spread. This scenario tests whether the model can track widening spatial inequality in the disease burden.

For each combination of temporal scenario and variance partition  $p$ , 50 independent replicate datasets are generated, yielding 300 datasets in total across the six simulation scenarios. Performance metrics are the same six as in simulation study 1 (Section 5.4.4), computed by comparing posterior median estimates of  $\{\theta_t(\mathcal{A}_m)\}$  against the true generating surface across all  $M = 746$  Data Zones and all  $T = 12$  months.

### 5.5.4 Results of simulation study 2

Table 5.3 reports the mean performance metrics for the six simulation scenarios, and Figure 5.7 displays the between-replicate variation in RMSE. All four models produce near-zero mean bias (range  $-0.0051$  to  $0.0024$ ), indicating good model performance. The Realign performance is stable across all three temporal scenarios, with Corr values between  $0.8972$  and  $0.9635$ , and improves uniformly with the covariate variance share  $p$ . Under T2 (AR(1) persistence), Corr rises from  $0.9292$  at  $p = 0.25$  to  $0.9635$  at  $p = 0.75$ .

Realign achieves mean correlations between  $0.8972$  and  $0.9635$ , above both naive approaches ( $0.8051$ – $0.8615$  for Naive0 and  $0.8078$ – $0.8529$  for Naive1). Its RMSE reductions relative to Naive0 range from 27% ( $p = 0.25$ ) to 47% ( $p = 0.75$ ) across the six scenarios, exceeding the naive estimator in every cell. The improvement scales with  $p$ , and at  $p = 0.75$  the time-invariant covariate carries three quarters of the log-risk variance, which the realignment model recovers directly, whereas the naive estimator cannot reconstruct this DZ-scale structure from GP-level data. Naive1 is consistently worse than Naive0 under T2, because the three-month moving average smooths over genuine month-to-month innovations. The correlation from the Realign model ranges from  $0.8972$ – $0.9292$  at  $p = 0.25$  to  $0.9549$ – $0.9635$  at  $p = 0.75$ , while the naive approaches remain insensitive to  $p$ . Figure 5.7 confirms the model ranking at the replicate level, with the replicate-level RMSE distribution of Realign lower in location than that of Naive0 in all six cells.

The coverage probabilities of the Realign model range from  $0.7343$  to  $0.8558$ , falling below the nominal 95% level in all six scenarios (Table 5.3). The Benchmark model, which has direct access to DZ-level counts, attains near-nominal coverage



in every cell (CP between 0.9305 and 0.9496). The gap between the two reflects the inferential cost of misalignment. Under *Realign*, the effective observation count drops from  $M \times T = 8,952$  to  $K \times T \approx 3,120$ , and the resulting coverage probabilities are slightly lower than the true uncertainty warrants.

Table 5.3: Mean performance metrics across 50 replicates for Simulation Study 2. Metrics are averaged (mean) over all  $M = 746$  Data Zones and  $T = 12$  months. CP and MIW are not applicable for *Naive0* and *Naive1* as they do not produce posterior distributions. The temporal scenarios are T1 (global linear trend), T2 (AR(1) persistence) and T3 (widening inequality). The rows compare *Naive0* and *Naive1* (deterministic interpolation using the first-year and time-varying registration weights respectively), *Realign* (the proposed spatio-temporal change-of-support model), and *Benchmark* (an idealised model fitted directly to DZ-level data).

| $p$  | Scenario | Model     | Corr   | RMSE   | MAE    | Bias    | CP     | MIW    |
|------|----------|-----------|--------|--------|--------|---------|--------|--------|
| 0.25 | T1       | Naive0    | 0.8506 | 0.1715 | 0.1076 | 0.0002  | –      | –      |
|      |          | Naive1    | 0.8529 | 0.1710 | 0.1071 | 0.0001  | –      | –      |
|      |          | Realign   | 0.9252 | 0.1217 | 0.0727 | –0.0051 | 0.8108 | 0.3219 |
|      |          | Benchmark | 0.9878 | 0.0491 | 0.0270 | –0.0017 | 0.9425 | 0.2234 |
|      | T2       | Naive0    | 0.8615 | 0.1627 | 0.1009 | –0.0004 | –      | –      |
|      |          | Naive1    | 0.8471 | 0.1717 | 0.1074 | –0.0007 | –      | –      |
|      |          | Realign   | 0.9292 | 0.1176 | 0.0711 | –0.0029 | 0.8558 | 0.3520 |
|      |          | Benchmark | 0.9855 | 0.0530 | 0.0320 | –0.0021 | 0.9481 | 0.2349 |
|      | T3       | Naive0    | 0.8051 | 0.1829 | 0.1137 | 0.0001  | –      | –      |
|      |          | Naive1    | 0.8078 | 0.1824 | 0.1134 | 0.0001  | –      | –      |
|      |          | Realign   | 0.8972 | 0.1338 | 0.0803 | –0.0030 | 0.7343 | 0.2984 |
|      |          | Benchmark | 0.9869 | 0.0480 | 0.0278 | –0.0014 | 0.9305 | 0.2238 |
| 0.75 | T1       | Naive0    | 0.8324 | 0.1790 | 0.1104 | –0.0001 | –      | –      |
|      |          | Naive1    | 0.8349 | 0.1785 | 0.1100 | –0.0001 | –      | –      |
|      |          | Realign   | 0.9587 | 0.0956 | 0.0572 | 0.0024  | 0.7963 | 0.2494 |
|      |          | Benchmark | 0.9923 | 0.0377 | 0.0219 | –0.0007 | 0.9429 | 0.1643 |
|      | T2       | Naive0    | 0.8531 | 0.1792 | 0.1106 | –0.0001 | –      | –      |
|      |          | Naive1    | 0.8510 | 0.1816 | 0.1122 | –0.0001 | –      | –      |
|      |          | Realign   | 0.9635 | 0.0953 | 0.0560 | –0.0006 | 0.8298 | 0.2667 |
|      |          | Benchmark | 0.9923 | 0.0405 | 0.0245 | –0.0010 | 0.9496 | 0.1728 |
|      | T3       | Naive0    | 0.8262 | 0.1816 | 0.1131 | 0.0003  | –      | –      |
|      |          | Naive1    | 0.8285 | 0.1811 | 0.1127 | 0.0002  | –      | –      |
|      |          | Realign   | 0.9549 | 0.0995 | 0.0591 | 0.0000  | 0.7727 | 0.2457 |
|      |          | Benchmark | 0.9928 | 0.0368 | 0.0222 | –0.0008 | 0.9343 | 0.1615 |

The two simulation studies above assess the model primarily through the accuracy of the recovered Data Zone-level relative-risk surface. In addition, across a number of demonstration runs the regression coefficients (and, in Simulation study 2, the

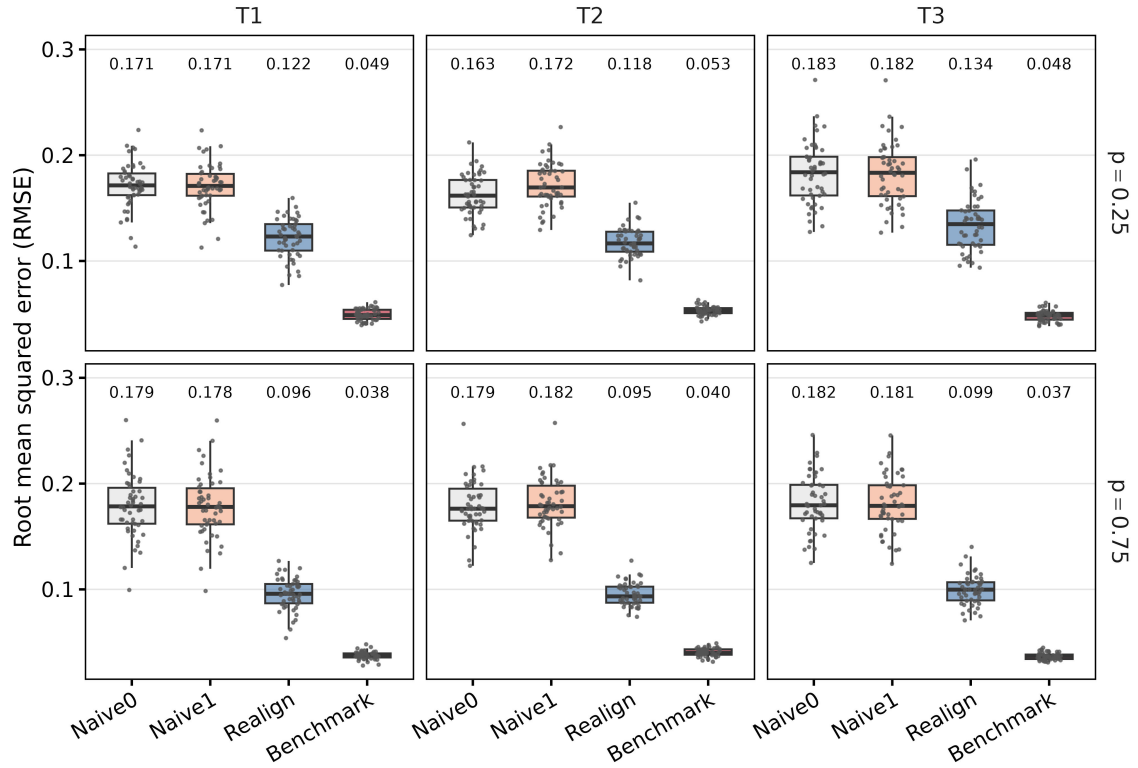


Figure 5.7: Between-replicate distribution of RMSE of posterior median DZ-level relative risks across 50 replicates per scenario, Simulation Study 2 (spatio-temporal). Panels are arranged by covariate variance share  $p$  (rows) and temporal scenario (columns: T1 global linear trend, T2 AR(1) persistence, T3 widening inequality).

temporal dependence parameter) were recovered well, with their posterior distributions concentrated around the values used to generate the data, indicating that these quantities are identifiable from the misaligned GP-level counts under the scenarios considered. A systematic parameter-recovery analysis spanning all simulation scenarios is left as a direction for future work.

## 5.6 Motivating study results

This section applies the realignment framework to the Glasgow City prescription data and organises the results around the four public health questions posed in Section 5.2.

### 5.6.1 Model specification

The ICAR-AR(1) realignment model from Section 5.3 is applied to the Glasgow City data with a scalar intercept  $\beta_0$  and a single DZ-level covariate, the income deprivation rate (hereafter *IncRate*). Each DZ's *IncRate* is the share of its working-age population classified as income deprived, sourced from the SIMD 2020 release and standardised to have mean zero and unit variance across the  $M = 746$  Glasgow Data Zones prior to fitting. The SIMD 2020 release is treated as a time-invariant deprivation proxy across the 2019–2020 study period. This indicator carries the highest weight (28%) among the seven SIMD 2020 domains, tied with the employment domain, and the two are near-collinear across Glasgow Data Zones (with Pearson correlation 0.97). Using *IncRate* alone therefore captures the dominant socio-economic gradient relevant to small-area respiratory prescribing, and income deprivation is an established determinant of the health burden ([Scottish Government, 2021](#)).

Priors are as specified in Section 5.3.3. Three MCMC chains are run for 50,000 iterations with 20,000 burn-in and thinning of 10, yielding 9,000 posterior samples in total. All scalar parameters exhibit  $\hat{R} \leq 1.01$  and minimum effective sample size above 500, while DZ-level relative risks show maximum  $\hat{R} = 1.03$ , providing no evidence against non-convergence. Point estimation throughout this section uses the posterior median  $\hat{\theta}(\mathcal{A}_m)$  (or the posterior median of the relevant derived summary), and uncertainty is summarised by the 95% credible interval (CI). Table 5.4 reports the scalar posterior summaries.

Table 5.4: Scalar posterior summaries for the ICAR-AR(1) model with *IncRate* covariate in the motivating study. Point estimates are posterior medians, and 95% credible intervals are the 2.5th and 97.5th posterior percentiles.

| Parameter       | Posterior median | 2.5%   | 97.5%  |
|-----------------|------------------|--------|--------|
| $\exp(\beta_1)$ | 1.6131           | 1.5382 | 1.6976 |
| $\sigma_\phi$   | 0.0971           | 0.0778 | 0.1140 |
| $\alpha$        | 0.9885           | 0.9524 | 0.9993 |
| $\delta$        | 35.49            | 29.03  | 40.90  |

The covariate effect is significantly positive, with a one-standard-deviation increase

in standardised IncRate multiplying the DZ-level relative risk by a posterior median factor of 1.6131 (95% CI [1.5382, 1.6976]). Expressed in natural units, one covariate standard deviation corresponds to a 12.2 percentage-point increase in the share of income-deprived working-age residents across Glasgow Data Zones, so an increase in IncRate by 10 percentage points is associated with an approximately 48% rise in the DZ-level relative risk. The gap in IncRate between the most-deprived and the least-deprived SIMD 2020 deciles is approximately 32 percentage points, which implies a covariate-driven relative-risk ratio of approximately 3.5. This accounts for the bulk of the observed fourfold time-averaged gradient reported in Section 5.6.5 and leaves a relatively small residual to be absorbed by the spatial random effect. The temporal persistence parameter  $\alpha$  concentrates near one (posterior median  $\hat{\alpha} = 0.9885$ , 95% CI [0.9524, 0.9993]), indicating that month-to-month changes in the residual surface are dominated by the previous month's structure. The posterior median spatial scale  $\hat{\sigma}_\phi = 0.0971$  (95% CI [0.0778, 0.1140]) is modest on the log-risk scale, consistent with the covariate absorbing a substantial share of the DZ-to-DZ variation in relative risk and leaving only the local residual for the spatial prior to explain. The posterior median dispersion parameter  $\hat{\delta} = 35.49$  places the level of overdispersion only moderately above the Poisson limit.

### 5.6.2 Model checking

To assess goodness-of-fit, posterior predictive checks were carried out on the GP-level support at which the data are observed. For each stored posterior draw, replicate prescription counts were generated from the fitted negative binomial likelihood, with the Data Zone-level risks aggregated to the GP catchments through the registration weights  $\eta$ . The resulting 95% posterior predictive intervals cover 96.5% of the 3,355 observed GP-month counts, close to the nominal level, and the posterior median fitted counts agree closely with the observations across the full range of the data (Pearson correlation 0.96; Figure 5.8 (a)). The randomised probability integral transform (PIT) histogram in Figure 5.8 (b) is centred and close to uniform, with a mild excess of central values indicating slightly conservative predictive intervals rather than any systematic misfit. Together with the convergence diagnostics reported above, these checks indicate that the fitted model provides an adequate description of the observed prescribing counts.

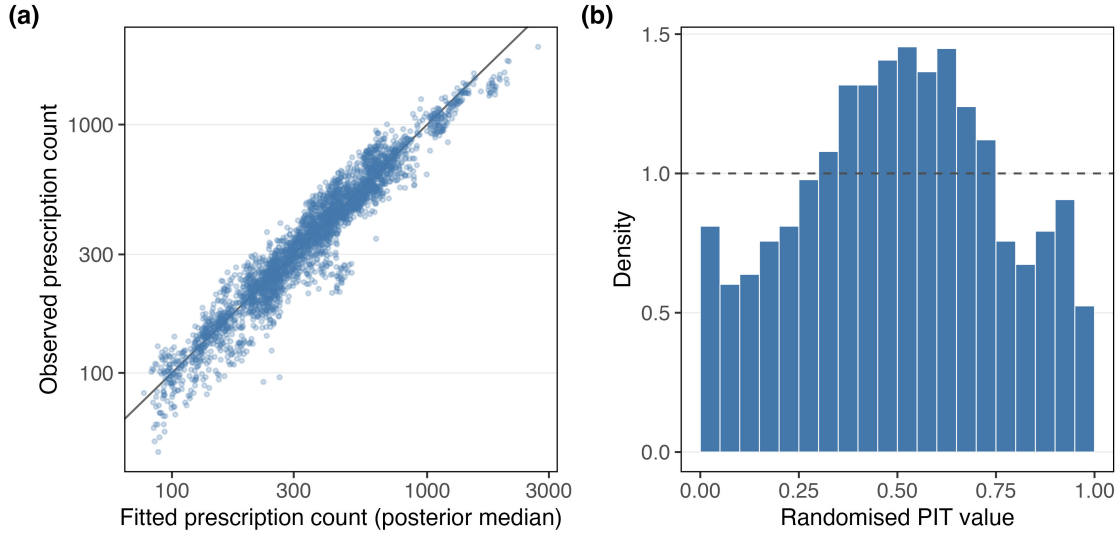


Figure 5.8: Posterior predictive checks for the ICAR-AR(1) model with `IncRate` covariate in the motivating study, showing (a) observed versus fitted (posterior median) GP-month prescription counts on logarithmic axes, with the 45° line in grey; (b) randomised PIT histogram, where the dashed line marks the density of the uniform distribution expected under a well-calibrated model.

### 5.6.3 Spatial distribution of relative risk

Figure 5.9 presents the estimated DZ-level relative risk surface and its uncertainty for the annual mean of 2019 (Panels A and B), the annual mean of 2020 (Panels C and D), and the single month March 2020 (Panels E and F). The annual maps are produced by first computing the mean of  $\theta_t(\mathcal{A}_m)$  across the 12 months of the corresponding year within each MCMC sample, and then summarising across samples by their posterior median. The March 2020 map is shown separately because that month contains the sharp pandemic-related prescribing spike identified in Section 5.6.4, and a separate annual average would hide this pattern.

The spatial pattern is broadly stable between 2019 and 2020, with low-risk Data Zones concentrating in the central and western parts of the city (covering the neighbourhoods commonly known as the West End, the City Centre, and much of Southside) and clusters of elevated risk in the north-west and north-east. These spatial patterns are consistent with areas identified as persistently deprived in Figure 5.1. The median posterior 95% credible interval width across all DZ-month combinations is 0.64, with an interquartile range of [0.41, 1.02]. The March 2020 risk surface (Panel E) preserves the socio-economic ordering seen in the annual maps but shifts the whole distribution upward. The city-wide median DZ-level relative increase

over the 2019 annual mean is 42% (IQR [34%, 50%]), and the increase is close to spatially uniform on the relative-risk scale (within-decile medians range from 39% to 48% across the 10 SIMD deciles). In absolute terms, however, the spike loads onto the already-deprived areas. The median DZ-level absolute increase is 0.69 in the most-deprived SIMD decile and only 0.18 in the least deprived. In addition, March 2020 shows visibly wider credible intervals across the city (Panel F), reflecting the additional posterior uncertainty induced by the prescribing spike.

The right column of Figure 5.9 displays the 95% credible interval widths. Two mechanisms contribute to the spatial variation in interval width. First, Data Zones with elevated posterior median relative risks tend to have wider credible intervals because the log-linear link induces a multiplicative transformation on the risk scale: the same posterior variance on the log scale translates to a wider interval at higher relative risks. Second, Data Zones linked to fewer GP surgeries, or contributing only peripherally to any GP catchment, likely reflect wider credible intervals through limited information coming from the likelihood.

#### 5.6.4 Temporal patterns and the Covid-19 prescribing spike

Figure 5.10 displays the city-wide posterior median relative risk over the 24-month study period, defined at each month  $t$  as the posterior median of the spatial mean  $\hat{\theta}_t = M^{-1} \sum_{m=1}^M \theta_t(\mathcal{A}_m)$ . The posterior median lies in the range 1.14–1.34 in every month except March 2020, when it peaks at 1.76 (95% CI [1.62, 1.89]). This March 2020 elevation corresponds to a 45% increase over the 14-month pre-spike average of 1.22, and is consistent with pandemic-related stocking up on respiratory medications reported elsewhere in the literature. The 95% credible band at the March 2020 point does not overlap the band of any adjacent month, confirming that the spike is a posterior feature rather than an artefact of sampling uncertainty. The posterior median relative risk returns to the pre-spike range in April 2020 and the rest of the year (1.16–1.34), so the spike is narrow in time. The relative risk is benchmarked against the national age–sex standardised prescribing rates averaged over 2016–2020 (Section 5.2.1), not against the Glasgow city-wide expected counts at the start of the study. Glasgow’s mean therefore sits above one throughout, consistent with its elevated deprivation profile compared to the rest of Scotland.



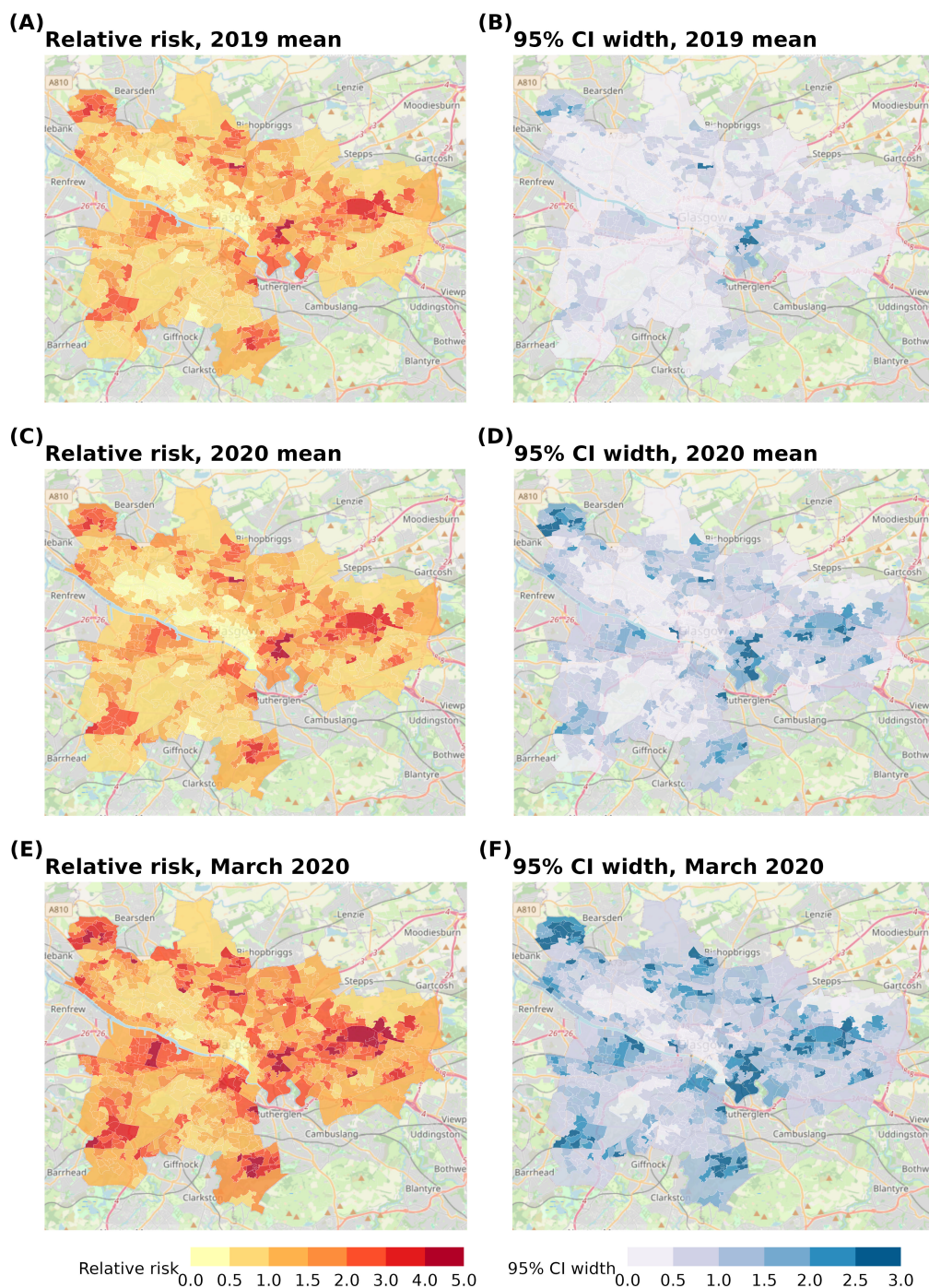


Figure 5.9: Estimated DZ-level relative risks and uncertainty under the ICAR-AR(1) model with IncRate covariate. Left column shows posterior median relative risks averaged over 2019 (A), averaged over 2020 (C), and for March 2020 (E). Right column shows the corresponding 95% credible interval widths (B, D, F). Colour scales are shared within each column.

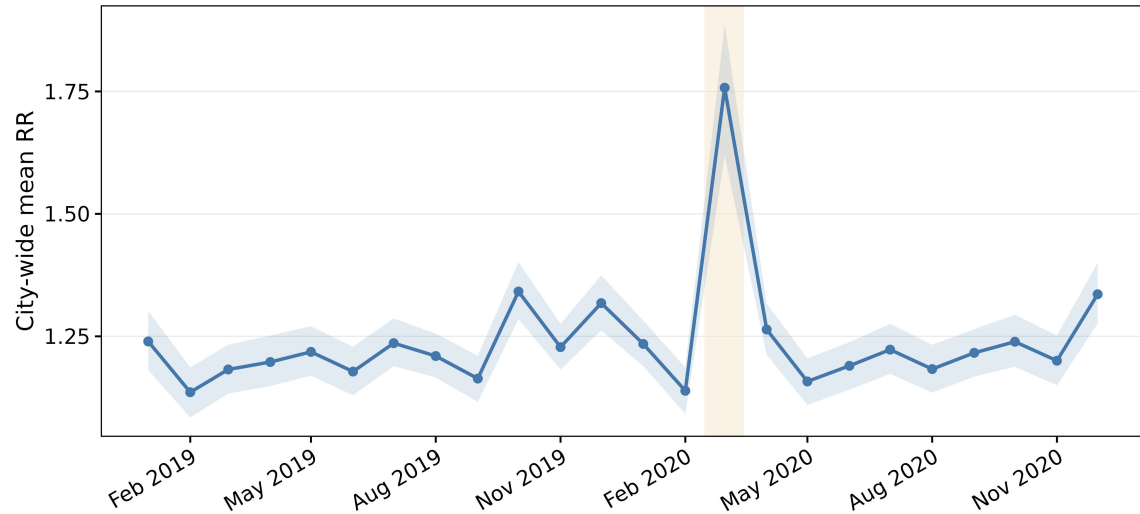


Figure 5.10: City-wide posterior median relative risk (relative to the national level over 2016–2020) over the 24-month study period under the ICAR-AR(1) model with IncRate covariate. Points show the posterior median of the monthly mean relative risk across all  $M = 746$  Data Zones, with the shaded band indicating the 95% credible interval. The highlighted window is a one-month band centred on March 2020, when the posterior median relative risk peaks at 1.76.

### 5.6.5 Health inequalities by deprivation

Figure 5.11 presents two complementary views of health inequality over the study period. Panel (A) shows the posterior median of the city-wide spatial standard deviation of DZ-level relative risks with 95% credible bands, while Panel (B) shows the time-evolving posterior median of the decile-averaged DZ-level relative risk stratified by SIMD 2020 deprivation decile. Panel (A) quantifies overall total inequality. The posterior median spatial standard deviation of relative risks fluctuates in a narrow band, with a substantial elevation in March 2020 consistent with the city-wide posterior median spike in Panel (A) of Figure 5.10. The socio-economic gradient in Panel (B) is the dominant and stable feature. Time-averaged posterior median relative risks increase monotonically with deprivation, from 0.50 in the least-deprived decile 10 ( $n = 38$  Data Zones), to 2.02 in the most-deprived decile 1 ( $n = 227$  Data Zones). This gradient is present in every month. The March 2020 spike lifts all decile-specific posterior medians without visible re-ordering.

Both panels show a gradual upward drift over the 24-month window. In Panel (A), the posterior median spatial standard deviation of DZ-level relative risks rises from a monthly range of  $[0.60, 0.65]$  across January–March 2019 to  $[0.78, 0.87]$  across October–December 2020, with a sharp peak of 1.07 in March 2020 mirroring the



city-wide spike. In Panel (B), the decile-1 / decile-10 ratio has monthly mean 4.00 during 2019 (range [3.77, 4.07]) and 4.21 during 2020 excluding March (range [4.05, 4.34]), a small widening of the deprivation gradient between the two calendar years. The March 2020 spike transiently compresses the ratio to 3.93, but the ratio returns to 4.13 in April 2020.

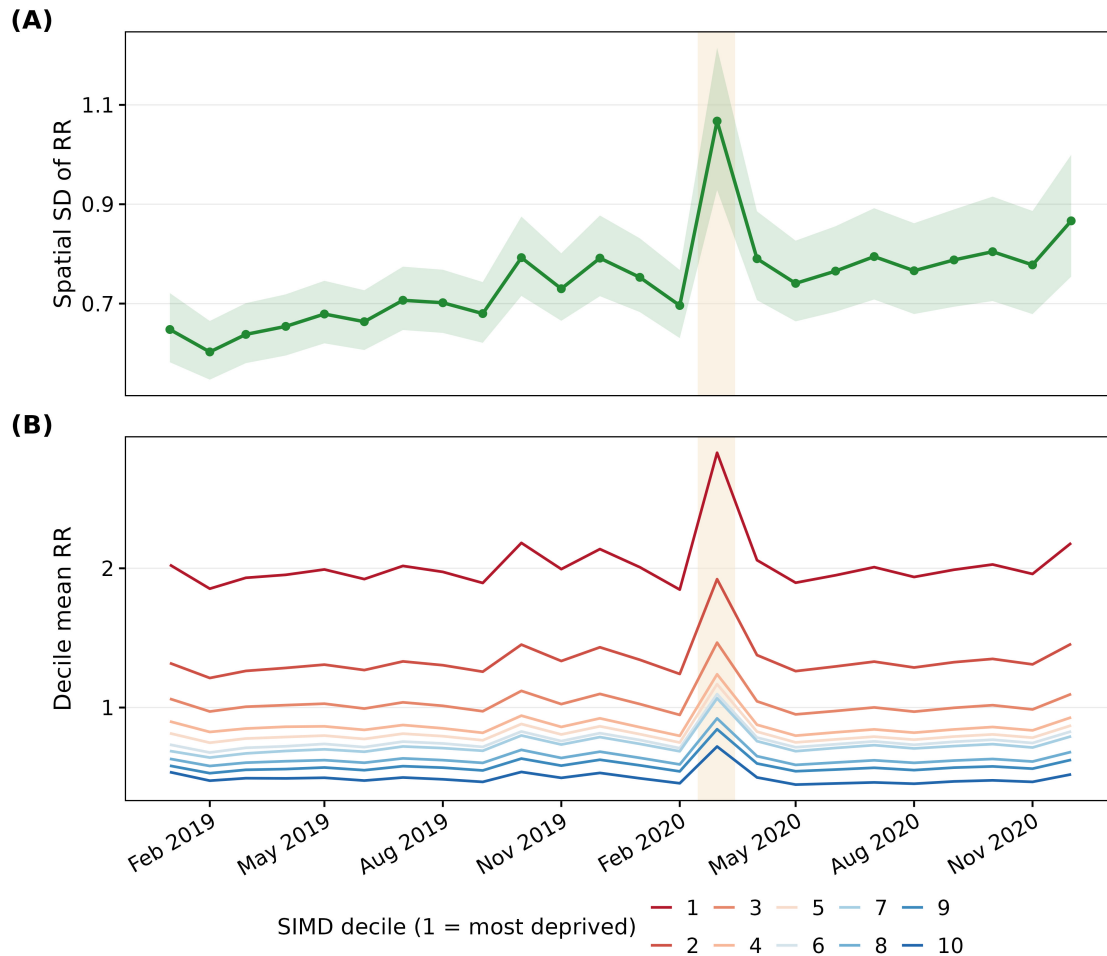


Figure 5.11: Health inequality in respiratory prescribing risk over the study period under the ICAR-AR(1) model with `IncRate` covariate. Panel (A) shows the posterior median of the city-wide spatial standard deviation of DZ-level relative risks with 95% credible bands. Panel (B) shows the posterior median of the decile-averaged DZ-level relative risk by SIMD 2020 deprivation decile, where decile 1 denotes the most deprived 10% of Scottish Data Zones. The highlighted window is centred on March 2020.

## 5.7 Discussion

This chapter has developed and evaluated a Bayesian spatio-temporal disease mapping framework for estimating small-area respiratory prescribing risk from spatially misaligned data with unknown and overlapping boundaries. In the context of the motivating study, the framework links GP-level prescription counts to latent Data Zone-level risks through a population-weighted change-of-support mechanism, and estimates the resulting risk surface using a CAR-AR(1) random effects structure fitted via MCMC in NIMBLE. The approach was first validated through two simulation studies, and then applied to monthly prescription data from Glasgow City over 2019–2020.

The simulation studies established that the realignment model consistently outperforms deterministic interpolation across all scenarios. In a purely spatial setting, the model reduced RMSE by 2%–27% relative to naive interpolation, with the largest gains in high-variance scenarios where model-based smoothing recovers the most additional signal. The inclusion of a DZ-level covariate amplified this advantage substantially, with RMSE reductions of 33%–52% at  $p = 0.25$  rising to 53%–60% at  $p = 0.75$  when the covariate explained a larger share of the log-risk variation. This Setting 2 evidence directly motivated the use of the `IncRate` covariate in the motivating study. In the spatio-temporal setting, the realignment model reduced RMSE by 27%–47% relative to naive interpolation across all three temporal scenarios, with the largest gains at  $p = 0.75$  where the time-invariant covariate captures most of the log-risk variance. The benchmark model attained near-nominal coverage (0.93–0.95), while the realignment model under-covered by roughly 0.05–0.20, reflecting the inferential cost of inferring DZ-level risks from misaligned GP-level counts.

The motivating study answered the four public health questions (a)–(d) posed in Section 5.2, most fully on (a), (c) and (d). On question (a), concerning the spatial distribution of prescribing burden, the posterior median risk surface identified north-western and eastern Glasgow as areas of elevated risk, with a clear spatial alignment to patterns of socio-economic deprivation (see Figure 5.9). On question (b), concerning estimation precision, the median 95% credible interval width across DZ-months was 0.64 with an interquartile range of [0.41, 1.02]. In addition, this study provided the risk surface with uncertainty quantification showing wider CIs for higher-risk areas. On question (c), concerning temporal patterns, the model detected a sharp city-wide elevation in prescribing risk in March 2020 (posterior median relative risk 1.76, versus a 1.14–1.34 range in other months), coinciding with the onset of the

Covid-19 pandemic. Outside the March 2020 spike, the shift between 2019 and 2020 is small (city-wide median change  $-3\%$ ). On question (d), concerning health inequalities, a persistent and visible deprivation gradient was evident throughout the study period. Time-averaged posterior median relative risks rose monotonically across the SIMD 2020 deciles, from 0.50 in the least-deprived decile 10 to 2.02 in the most-deprived decile 1 (a ratio of approximately 4.1). The posterior median for income deprivation is 1.6131 ([1.5382, 1.6976]) per one standard deviation increase in standardised IncRate.

The change-of-support approach developed here is in the same family as the multiple-membership (MM) framework of [Gramatica et al. \(2021, 2024\)](#) and the joint-realignment formulation of [Lee and Anderson \(2023\)](#). All three approaches link observations on one areal support to a latent risk surface on another through a population-weighted linkage. The present work differs from them in the two respects set out in Section 5.3: the aggregation in the change-of-support mechanism here operates on the relative-risk scale rather than on the log-linear predictor scale, and the framework extends the MM family to a spatio-temporal setting through a dynamic ICAR-AR(1) process. No systematic head-to-head performance comparison with these alternatives was undertaken; the present work therefore does not claim superiority over MM-CAR or data-augmentation approaches as a predictor of DZ-level risks. Its contribution is the construction of a risk-scale spatio-temporal extension of the MM family, together with simulation evidence quantifying the inferential cost of misalignment relative to naive interpolation and an idealised direct DZ-level benchmark.

Several limitations of this work should be noted. First, the realignment model estimates  $M = 746$  DZ-level risk parameters from  $K \approx 140$  GP-level observations in the motivating study, so inference relies heavily on the spatio-temporal smoothing imposed by the ICAR-AR(1) prior and on the informativeness of the DZ-level covariate. The resulting credible intervals have median width 0.64 (Section 5.6.3), and point estimates of small-area risk may remain sensitive to the choice of spatial prior. Second, the motivating study is restricted to Glasgow City. GP surgeries near the city boundary serve patients who reside both inside and outside the study region. Retaining only surgeries with at least 75% of their registered patients within Glasgow and rescaling the registration weights accordingly mitigates but does not eliminate the resulting edge effect, which appears in wider credible intervals at the periphery of the maps (Figure 5.9). Extending the framework to all of Scotland would remove this boundary issue but would substantially increase the computational burden, as the number of

DZ-level parameters would grow from  $746 \times 24$  to approximately  $7,000 \times 60$  while the sparse  $\eta$  matrix would become correspondingly larger. Third, the change-of-support mechanism models the GP-level relative risk as a population-weighted sum of DZ-level relative risks. This mechanism does not require that the spatial distribution of prescription-receiving patients matches the distribution of all registered patients, because differential risk across Data Zones is accounted for through the  $\theta_t(\mathcal{A}_m)$  terms in the sum. The assumption is that the registration weights  $\eta_{km,t}$  accurately reflect the proportional allocation of each surgery's registered population across Data Zones. If these weights are mis-measured, then the DZ-level estimates may be biased.

Future work could address these limitations in several directions. Scaling the framework to a national study region covering all of Scotland would require efficient sparse matrix operations and careful attention to the computational scaling of the  $\eta$  matrix and the DZ-level random effects. A systematic head-to-head comparison between the proposed risk-scale realignment, the log-scale MM-CAR formulation of [Gramatica et al. \(2021, 2024\)](#) and the data-augmentation approach of [Lee and Anderson \(2023\)](#) under a common simulation protocol would clarify the relative merits of each. Investigating alternative spatial priors, such as the Leroux CAR model, may help stabilise parameter identification in settings with very sparse GP-DZ linkage by introducing a data-driven balance between spatial smoothing and independent variation. Finally, adapting the realignment framework to other GP-level health outcomes, such as consultation rates, would test the generalisability of the approach beyond the respiratory prescribing context considered here.

# Chapter 6

## Conclusions

This chapter summarises the main contributions of the thesis, discusses their common limitations, and outlines directions for future research. The unifying methodological theme is statistical inference from spatially misaligned, aggregated, and partially observed data. Across all three studies, the spatial support at which data are observed does not coincide with the support at which inference is required. This mismatch shapes exposure assignment, spatial smoothing, uncertainty propagation, and the interpretation of respiratory ill health treated in primary care.

### 6.1 Summary of contributions

The three studies address successive stages from environmental exposure to respiratory ill health treated in primary care. Chapter 3 produces spatially complete monthly pollution predictions from sparse monitoring data, providing exposure inputs for the subsequent health analyses. Chapter 4 uses these predictions to estimate the health effect of air pollution on respiratory prescribing at the GP surgery level, while accounting for partially observed registered patient populations. Chapter 5 extends the change-of-support perspective in the reverse direction, estimating Data Zone-level respiratory prescribing risk from GP-level prescription counts. Together, the chapters form a progression from exposure prediction, to population-informed health-effect estimation, to small-area disease mapping under spatial misalignment.

Chapter 3 compared statistical and machine learning approaches for predicting monthly  $\text{NO}_2$ ,  $\text{PM}_{10}$  and  $\text{PM}_{2.5}$  concentrations across Scotland at  $1 \text{ km}^2$  resolution between 2016 and 2020. The candidate models ranged from linear regression to additive models, hierarchical spatio-temporal smoothing models, and random forests,

compared under a spatial cross-validation design. The linear models impose rigid predictor–response relationships, whereas random forests can capture more flexible non-linear and interaction effects. Random forests achieved the best predictive performance for  $\text{NO}_2$  and  $\text{PM}_{2.5}$ , but the linear model was preferred for  $\text{PM}_{10}$  on grounds of both accuracy and parsimony. The performance gap between random forests and linear models was small and the choice was not universal across pollutants. This is consistent with the explanation that the relevant pollutant–predictor relationships are fairly simple, so that the additional flexibility of random forests does not provide substantive gains beyond a well-specified linear model. Bayesian models with explicit spatio-temporal smoothing performed comparatively poorly, because the available covariates already captured most of the spatial signal and autocorrelation in the data. The resulting monthly prediction surfaces, with 95% prediction intervals, form a new data resource that serves as the exposure input for the rest of the thesis.

Building on these exposure prediction surfaces, Chapter 4 quantified the health effects that the three air pollutants ( $\text{NO}_2$ ,  $\text{PM}_{10}$  and  $\text{PM}_{2.5}$ ) have on respiratory ill health treated in primary care across Scotland. The methodological contribution was the development of population-informed neighbourhood matrices that define spatial closeness between GP catchment areas through registered population overlap rather than the geographic distance between surgery buildings. The matrices substantially reduced residual spatial autocorrelation while preventing over-smoothing, compared with a conventional 5-nearest-neighbours approach to defining spatial closeness similarly. Specifically, the 5-nearest-neighbours baseline induced over-smoothing evidenced by negative residual Moran's I values, with the General Spatial Autocorrelation Indicator (GSAI) reaching 15%–32% across medication and period combinations. The population-informed matrices instead brought the GSAI close to the nominal 5% level and reduced the Mean Absolute Moran's I (MAI) by 20.99%–48.63% across model variants. Pollution exposures were assigned to each GP based on the likely residential areas of its registered patients rather than at the surgery location which should reduce the exposure measurement error and hence improve the accuracy of the estimated health effects. The study found a consistent and statistically significant positive association between  $\text{PM}_{10}$  and preventer medication prescribing, with an estimated 0.5% increase in prescribing rates per one standard deviation increase in  $\text{PM}_{10}$  ( $2.77 \mu\text{g}/\text{m}^3$ ). The associations for  $\text{NO}_2$  and  $\text{PM}_{2.5}$  were weaker and not consistently significant. A sensitivity analysis that propagated the full predictive uncertainty of the  $\text{PM}_{10}$  exposure model through to the health model confirmed that the  $\text{PM}_{10}$  association remained significant.

Chapter 5 developed a Bayesian spatio-temporal disease mapping framework for estimating Data Zone-level respiratory prescribing risk from spatially misaligned GP-level prescription counts. The model links GP-level observations to a latent DZ-level relative-risk surface through a population-weighted change-of-support mechanism. It extends the multiple-membership framework of [Gramatica et al. \(2021\)](#) from a spatial setting to a spatio-temporal setting through a dynamic ICAR-AR(1) process. Two simulation studies provided systematic evidence on the inferential cost of spatial misalignment. This cost was quantified as the gap in DZ-level risk recovery between the proposed realignment model and an idealised DZ-level direct-fit benchmark, with the realignment model also benchmarked against naive deterministic interpolation. In the purely spatial setting, the realignment model reduced RMSE relative to naive interpolation by 2%–27% without a DZ-level covariate, rising to 33%–52% and 53%–60% when a structured covariate accounted for a quarter and three quarters of the log-risk variation. In the spatio-temporal setting, RMSE reductions of 27%–47% were obtained across all three temporal scenarios (linear trend, AR(1) persistence, and widening inequality), with the largest gains at  $p = 0.75$  where the time-invariant covariate captures most of the log-risk variance. The benchmark model attained near-nominal coverage (0.93–0.95), whereas the realignment model under-covered by roughly 0.05–0.20, reflecting the inferential cost of inferring DZ-level risks from misaligned GP-level counts.

The framework was applied to Glasgow City over 2019–2020 with a socio-economic covariate. It recovered a persistent monotonic deprivation gradient, showing that posterior median relative risks in the most-deprived SIMD 2020 decile were approximately four times those in the least-deprived decile. The framework also captured a narrow city-wide spike in March 2020 during the national lockdown following the Covid-19 pandemic that lifted relative risks across all deprivation deciles roughly in parallel without re-ordering the gradient. A chapter-specific limitation of Chapter 5 is that no formal head-to-head comparison was carried out with two existing related approaches: the log-scale multiple-membership formulation of [Gramatica et al. \(2021, 2024\)](#) and the data-augmentation approach of [Lee and Anderson \(2023\)](#). A direct comparison is constrained by the fact that the existing approaches are formulated for the purely spatial setting, whereas the framework presented here is space-time. The present framework differs from these alternatives in two respects: aggregation operates on the relative-risk scale rather than the log-linear predictor scale, and the multiple-membership family is extended to the spatio-temporal setting through a dynamic ICAR-AR(1) process. A systematic comparison under a common simulation



protocol would clarify when each spatial strategy is preferred, and how these differences in aggregation scale and dynamic structure translate into inferential and computational trade-offs. This question is specific to the disease mapping framework rather than a thesis-wide concern, and remains a target for future work.

## 6.2 Common methodological themes

Several common themes run across the three studies. Spatial support mismatch arises at every stage of the pipeline, though the form of the challenge differs across the chapters: Chapter 3 predicts from a sparse monitoring network to a finer target grid, Chapter 4 reconfigures exposure assignment and small-area smoothing across overlapping and partially unknown GP patient catchments, and Chapter 5 reverses the direction of inference and estimates DZ-level risks from GP-level observations. In each case, handling the mismatch between the spatial supports of the observations and the inference targets explicitly affects the resulting inference rather than only the preprocessing pipeline. The population-informed neighbourhood matrices of Chapter 4 reduced residual autocorrelation in the GP-level model, and the change-of-support realignment of Chapter 5 projected the small-area distribution of risk onto a finer resolution. The role of GP catchments illustrates why this matters: they are overlapping, partly unknown, and not bounded by fixed administrative units, which creates modelling complexity beyond what standard areal-unit smoothing can absorb, and motivated both the population-informed matrices and the change-of-support realignment. GP-level inference would be methodologically simpler, but it would not deliver risk estimates on the spatial unit at which public health surveillance, deprivation indices, and area-based policy decisions are defined. The GP-to-DZ realignment of Chapter 5 is therefore valuable because it places small-area inference on the support that public health users actually need. Uncertainty also takes on a different character at each stage. Predictive uncertainty in Chapter 3 propagates downstream into the health-effect estimates of Chapter 4, whereas in Chapter 5 the dominant uncertainty is structural, arising from the need to recover  $M = 746$  DZ-level risks from roughly  $K = 140$  GP-level observations together with weak linkage information and edge effects. This  $M \gg K$  imbalance, combined with the catchment complexity discussed above, defines the joint computational and inferential challenge that the thesis addresses through Bayesian hierarchical models with informative covariates and spatial priors.

The frameworks in Chapters 4 and 5 are complementary rather than competing,



because they answer different questions and target different spatial supports. The Chapter 4 framework is an ecological regression model whose aim is to estimate the association between air pollution and respiratory prescribing. It works on the GP-catchment support and is the appropriate choice when the quantity of interest is an exposure–response effect and covariates such as pollution are available on that support. The Chapter 5 framework, by contrast, is a change-of-support disease mapping model whose aim is to produce a smooth, interpretable map of relative risk on fixed administrative units. It should be preferred when the deliverable is small-area risk estimates for surveillance, resource allocation or the study of health inequalities on Data Zones. In this sense the Chapter 5 model is not a strict successor to the Chapter 4 model, as one estimates *why* risk varies, the other estimates *where* it is. When both an exposure–response estimate and an administrative-unit risk map are required, the two can be used in sequence, with the population-informed neighbourhood structure of Chapter 4 and the risk-scale realignment of Chapter 5 addressing the two halves of the problem.

### 6.3 Limitations

The three studies share several limitations, beginning with the spatial supports at which the data are recorded. The Scottish monitoring network underlying Chapter 3 is sparse and concentrated in urban areas, an inherent feature of the data context rather than a study choice. Predictions in unmonitored regions therefore rely on the assumption that predictor–response relationships estimated from the available sites apply across Scotland, and any geographic bias in prediction quality propagates downstream into the health analyses. The chapter also implicitly treats each 1 km<sup>2</sup> grid square as a point at its centre rather than integrating the underlying continuous concentration field over the cell (Gelfand et al., 2001). Monitoring locations are also chosen preferentially, since monitors are typically placed where higher pollution is expected, so the locations at which the spatial process is observed are not independent of the values being measured Diggle et al. (2010). A further data limitation is specific to Chapter 4, in which the population-informed neighbourhood matrices assume that, for any GP surgery, the patients receiving respiratory prescriptions are spatially distributed across Data Zones in the same proportions as the surgery’s full registered population. This assumption does not in general hold, because prescription rates differ across the Data Zones served by a single surgery.

Beyond the data themselves, the structure of the models imposes its own limi-

tations. The health-effect estimates in Chapter 4 may be conservative, particularly where prediction uncertainty is not fully considered. Random prediction error in the exposure surfaces attenuates pollution–prescribing associations toward the null (Carroll et al., 2006), and only  $\text{PM}_{10}$  has so far been re-estimated under full predictive uncertainty propagation. The spatio-temporal smoothing model in Chapter 4 also assumes a static neighbourhood structure  $\mathbf{W}$  and a fixed random-effects vector length over the 60-month window. In practice, GP surgeries open, close, and merge over a five-year period, while patient registration patterns shift in response to housing developments and population mobility, so a static  $\mathbf{W}$  cannot fully represent the evolving spatial dependence structure of GP-level prescribing. The disease mapping framework in Chapter 5 estimates 746 DZ-level risk parameters from approximately 140 GP-level observations at each time point, so inference depends heavily on the spatial smoothing imposed by the ICAR-AR(1) prior and on the informativeness of the DZ-level covariate. The median 95% credible interval width across DZ-months is 0.64. Point estimates of small-area risk may remain sensitive to the choice of spatial prior in DZs with sparse GP linkage.

A further set of limitations concerns the scope of the disease mapping framework in Chapter 5, which arises from two distinct sources. The first is computational: the CAR-AR(1) random-effect structure scales poorly with the size of the study region, since a Scotland-wide spatio-temporal analogue would involve roughly 7,000 Data Zones across a comparable monthly window, or about 420,000 latent random effects per chain, which is computationally infeasible under the current sampling implementation. The motivating study was therefore restricted to Glasgow City. The second concerns boundary GP surgeries within the city itself. These surgeries serve registered patients both inside and outside Glasgow, and the change-of-support framework as deployed here treats both groups as contributing to the same Glasgow-level prescribing risk. This is equivalent to assuming that the risk experienced by patients living outside Glasgow is the same as the risk inside the city. To limit the impact of this assumption, the motivating study selected only Glasgow-dominating GP surgeries, retaining those for which the majority (75% or more) of registered patients reside within Glasgow City. Nevertheless, residual edge effects remain visible as wider credible intervals near the city periphery.

## 6.4 Future work

Several directions for future work follow from these limitations. The most immediate is to integrate air pollution prediction uncertainty into health-effect estimation. The sensitivity analysis in Chapter 4 propagated the predictive distribution of  $PM_{10}$  through to the health model and confirmed that the association remained significant. Extending this approach to all three pollutants simultaneously, within a fully Bayesian framework that jointly estimates exposure and health effects, would provide a more coherent treatment of measurement error, and would reduce the attenuation bias that affects the current two-stage estimates.

Refining the air pollution prediction model itself offers a complementary direction. Chapter 3 currently assigns each  $1 \text{ km}^2$  grid square the prediction generated at its centre. A rigorous point-to-area realignment, in which each grid-square exposure is represented as a spatial integral of the underlying continuous concentration field with associated integration uncertainty, would tighten the link between the modelled spatial process and the prediction target but may be computationally challenging. The same chapter's monitor placement mechanism is also worth modelling explicitly. Jointly modelling monitoring locations and the concentration surface, in the spirit of Diggle et al. (2010), would account for the preferential placement mechanism rather than conditioning on it.

Within the health-effect analysis, one natural extension concerns the GP–Data Zone linkage structure in Chapter 4. The current model treats the rows of the registration matrix as fixed quantities derived from annually published patient counts, and assumes that the neighbourhood structure  $\mathbf{W}$  remains constant over the 60-month period. A latent Dirichlet specification for the rows of the registration matrix, informed by the observed weights, would propagate the uncertainty about how prescription-receiving patients are distributed across Data Zones through to the inferred associations.

For the disease mapping framework, two further extensions concern the spatial prior and the computational scaling of the model in Chapter 5. The ICAR-AR(1) prior used here imposes full spatial smoothing, which may over-smooth risk estimates in Data Zones with sparse GP linkages. The Leroux CAR prior offers a data-driven balance between spatial smoothing and independent variation through its dependence parameter  $\rho \in [0, 1]$ , and may yield more stable inference where the GP–DZ linkage is sparser than in Glasgow City. To enable application of the framework at the national scale, the full CAR-AR(1) random-effect structure would need to be replaced

with a reduced-rank spatio-temporal basis representation ([Lindgren et al., 2011](#); [Cameletti et al., 2013](#)) that captures large-scale variation in fewer parameters. Such an extension would eliminate the boundary effects inherent in the city-level analysis and would enable comparison of respiratory ill health treated in primary care, and the corresponding health inequalities, across Scottish Health Boards.

The methods developed in this thesis are not specific to respiratory prescribing data and could be applied to other GP-level health outcomes, such as consultation rates for mental health conditions or referral rates for secondary care. Extending the analysis to later post-pandemic years would also help separate the persistence of the deprivation gradient from the transient Covid-19 disruption observed in 2020. More broadly, this thesis has argued that spatial support mismatch is not a nuisance preprocessing issue but a modelling problem in its own right. The spatial unit at which data are observed is not always the spatial unit at which inference is required, and treating this distinction explicitly changes both how models are formulated and how their outputs are interpreted. The studies presented here have provided practical Bayesian solutions for each stage of the environmental health pipeline, and have suggested a coherent set of methodological directions for extending these solutions to richer data structures and broader application contexts.

# Appendix A

## Component-wise Metropolis–Hastings sampler for the ICAR-AR(1) model

This appendix describes the MCMC sampling strategy used in Section 5.6 for posterior inference in the spatio-temporal realignment model. In the motivating study, the model contains  $M \times T = 746 \times 24 = 17,904$  latent spatio-temporal parameters together with four scalar parameters  $(\beta_0, \sigma_\phi, \alpha, \delta)$ . The central component is a component-wise Metropolis–Hastings (CW-MH) sampler for the spatial innovation fields, adapted from the per-site updating strategy implemented in the CARBayesST R package (Lee et al., 2018; Lee, 2020). The present implementation modifies this strategy to accommodate the population-weighted GP–DZ deconvolution likelihood (Equation (5.2)) and the innovation-based decomposition of the AR(1) structure.

### A.1 Innovation decomposition and sampling targets

The AR(1) specification in Equation (5.5) is implemented through a set of innovation fields  $\boldsymbol{\varepsilon}_t = (\varepsilon_{1,t}, \dots, \varepsilon_{M,t})^\top$  for  $t = 1, \dots, T$ , defined by

$$\phi_1(\mathcal{A}_m) = \varepsilon_{m,1}, \quad \phi_t(\mathcal{A}_m) = \alpha \phi_{t-1}(\mathcal{A}_m) + \varepsilon_{m,t}, \quad t = 2, \dots, T. \quad (\text{A.1})$$

Each innovation vector follows the ICAR prior independently across time,

$$\boldsymbol{\varepsilon}_t \mid \mathbf{Q} \sim \text{N}(\mathbf{0}, \mathbf{Q}^-), \quad t = 1, \dots, T, \quad (\text{A.2})$$

where  $\mathbf{Q} = \text{diag}(\mathbf{W}\mathbf{1}) - \mathbf{W}$  is the ICAR precision matrix derived from the binary adjacency matrix  $\mathbf{W}$ . The sampler targets the first-period field  $\boldsymbol{\phi}_1$  directly at  $t = 1$  and

the innovation fields  $\varepsilon_2, \dots, \varepsilon_T$  at subsequent time points, each of dimension  $M$ . The remaining field values  $\phi_t(\mathcal{A}_m)$  for  $t \geq 2$  are deterministic functions of  $(\phi_1, \varepsilon_2, \dots, \varepsilon_t)$  via Equation (A.1). A sum-to-zero constraint  $\sum_{m=1}^M \phi_1(\mathcal{A}_m) = 0$  is imposed on the first-period field only, consistent with Section 5.3.2, while the innovations  $\varepsilon_t$  for  $t \geq 2$  are not mean-centred.

## A.2 Parameter blocking and sampler assignment

The model parameters are partitioned into three groups, each assigned a different sampling algorithm.

**First-period field and subsequent innovations  $\phi_1, \varepsilon_2, \dots, \varepsilon_T$ : component-wise Metropolis–Hastings.** The first-period field  $\phi_1$  and each innovation field  $\varepsilon_t$ ,  $t \geq 2$ , are updated in turn by sweeping over all  $M$  sites with univariate Gaussian proposals whose variance is inversely proportional to the local neighbourhood size  $d_m$ . The acceptance ratio combines a closed-form ICAR prior ratio with an incrementally computed likelihood ratio. Full details are given in Appendix A.3.

**Correlated scalars  $(\beta_0, \sigma_\phi, \alpha)$ : automated factor slice sampler.** These three parameters are updated jointly using NIMBLE’s automated factor slice sampler (AF\_slice), which performs univariate slice sampling (Neal, 2003) along adaptively chosen directions given by the eigenvectors of an estimated posterior covariance matrix (de Valpine et al., 2017). The logit transformation  $\text{logit}(\alpha) = \log(\alpha/(1 - \alpha))$  maps the constrained interval  $(0, 1)$  to the real line, enabling unconstrained slice sampling. The intercept  $\beta_0$  is included in the same block because it is moderately correlated with  $\sigma_\phi$  through the overall mean of the log-risk surface.

**Overdispersion  $\delta$ : univariate slice sampler.** The negative binomial dispersion parameter  $\delta$  is updated independently using a univariate slice sampler (Neal, 2003). This parameter is only weakly correlated with the spatial parameters in the posterior, so univariate updating is adequate.

### A.3 Component-wise Metropolis–Hastings algorithm

The CW-MH sampler processes one field per invocation:  $\phi_1$  at the first time point and  $\varepsilon_t$  for  $t = 2, \dots, T$  thereafter. Within one MCMC iteration, the update order is:

1. For each  $t = 1, \dots, T$ , invoke the CW-MH sampler once on the corresponding field, sweeping over all  $M$  sites with the per-site update defined below.
2. Apply the sum-to-zero constraint to  $\phi_1$  after its sweep.
3. Update values back to the model, recompute the deterministic nodes and log-probabilities, and refresh the shadow vectors used for incremental evaluation.
4. Update the correlated scalars  $(\beta_0, \sigma_\phi, \alpha)$  jointly via `AF_slice`, and the dispersion  $\delta$  by a univariate slice step.

The per-site update at time  $t$  and site  $m$  is described below. For notational brevity the symbol  $\varepsilon_{m,t}$  denotes both the innovation  $\varepsilon_{m,t}$  for  $t \geq 2$  and the field value  $\phi_1(\mathcal{A}_m)$  at the first time point. For the current value  $\varepsilon_{m,t}^{\text{cur}}$ , the update proceeds as follows.

#### Step 1: ICAR conditional mean

Let  $d_m = \sum_{j=1}^M w_{mj}$  denote the number of neighbours of Data Zone  $\mathcal{A}_m$ . The ICAR full conditional for site  $m$  at time  $t$  is

$$\varepsilon_{m,t} \mid \varepsilon_{-m,t} \sim \text{N}\left(\bar{\varepsilon}_{m,t}, \frac{1}{d_m}\right), \quad (\text{A.3})$$

where  $\bar{\varepsilon}_{m,t} = d_m^{-1} \sum_{j \sim m} \varepsilon_{j,t}^{\text{cur}}$  is the mean of the current innovation values at the neighbours of site  $m$ , and the notation  $j \sim m$  indicates that  $\mathcal{A}_j$  and  $\mathcal{A}_m$  share a boundary.

#### Step 2: proposal

Draw a candidate value from

$$\varepsilon_{m,t}^* \sim \text{N}\left(\varepsilon_{m,t}^{\text{cur}}, \frac{s^2}{d_m}\right), \quad (\text{A.4})$$

where  $s > 0$  is a global tuning parameter shared across all sites and time slices. The proposal variance  $s^2/d_m$  is inversely proportional to the number of neighbours, matching the conditional prior variance  $1/d_m$  up to the factor  $s^2$ . This explicit scaling

incorporates the ICAR conditional structure directly into the proposal, so that sites with many neighbours (tight conditional, small variance  $1/d_m$ ) receive proportionally smaller proposals. In contrast, NIMBLE's default sampler uses a proposal variance that does not depend on  $d_m$ , relying instead on per-site adaptive tuning to learn the appropriate scale empirically. Define the proposal offset  $d = \varepsilon_{m,t}^* - \varepsilon_{m,t}^{\text{cur}}$ .

### Step 3: log prior ratio

Since the proposal in Equation (A.4) is symmetric about  $\varepsilon_{m,t}^{\text{cur}}$ , the Hastings correction cancels and the log acceptance ratio reduces to  $\log r = \log r_{\text{prior}} + \log r_{\text{lik}}$ . The log prior ratio is the difference in ICAR full conditional log-densities evaluated at the proposed and current values. The log-density of the ICAR conditional (Equation (A.3)) at a value  $\varepsilon$  is  $-\frac{d_m}{2}(\varepsilon - \bar{\varepsilon}_{m,t})^2$  plus a normalising constant that does not depend on  $\varepsilon$ . Subtracting gives

$$\log r_{\text{prior}} = -\frac{d_m}{2} [(\varepsilon_{m,t}^* - \bar{\varepsilon}_{m,t})^2 - (\varepsilon_{m,t}^{\text{cur}} - \bar{\varepsilon}_{m,t})^2]. \quad (\text{A.5})$$

This computation requires only the  $d_m$  neighbour values, exploiting the Markov property of the ICAR prior.

### Step 4: AR(1) forward propagation

Expanding the AR(1) recursion in Equation (A.1), the field at time  $t'$  depends on  $\varepsilon_{m,t}$  through

$$\phi_{t'}(\mathcal{A}_m) = \alpha^{t'-t} \varepsilon_{m,t} + (\text{terms not involving } \varepsilon_{m,t}), \quad t' \geq t.$$

When  $\varepsilon_{m,t}$  changes by  $d$ , the field at all subsequent time slices changes by

$$\phi_{t'}^{\text{new}}(\mathcal{A}_m) - \phi_{t'}^{\text{cur}}(\mathcal{A}_m) = \alpha^{t'-t} \cdot d, \quad t' = t, t+1, \dots, T. \quad (\text{A.6})$$

Since the linear predictor (Equation (5.4)) reduces to  $\log \theta_{t'}(\mathcal{A}_m) = \beta_0 + \sigma_\phi \cdot \phi_{t'}(\mathcal{A}_m)$  in the intercept-only motivating study model, the updated DZ-level relative risk follows as

$$\theta_{t'}^{\text{new}}(\mathcal{A}_m) = \theta_{t'}^{\text{cur}}(\mathcal{A}_m) \cdot \exp(\sigma_\phi \cdot \alpha^{t'-t} \cdot d). \quad (\text{A.7})$$

This closed-form expression avoids recomputing the full AR(1) recursion at each Metropolis step, reducing the per-proposal cost from  $O(T-t)$  sequential operations to a single exponentiation per affected time slice.



### Step 5: log likelihood ratio

Because the GP-level risk  $\theta_{t'}(\mathcal{S}_k) = \sum_{m'=1}^M \eta_{km',t'} \theta_{t'}(\mathcal{A}_{m'})$  (Equation (5.2)) is a weighted sum over all DZ-level risks, and only DZ  $\mathcal{A}_m$  changes, the updated GP-level risk for any surgery  $\mathcal{S}_k$  linked to  $\mathcal{A}_m$  is

$$\theta_{t'}^{\text{new}}(\mathcal{S}_k) = \theta_{t'}^{\text{cur}}(\mathcal{S}_k) + \eta_{km,t'} [\theta_{t'}^{\text{new}}(\mathcal{A}_m) - \theta_{t'}^{\text{cur}}(\mathcal{A}_m)]. \quad (\text{A.8})$$

This incremental update is  $O(1)$  per observation, compared with  $O(M)$  for a full recomputation of the weighted sum.

Under the negative binomial parameterisation in Equation (5.1), the log-likelihood for a single observation  $(k, t')$  with count  $Y_{t'}(\mathcal{S}_k)$  and expected mean  $\mu_{k,t'} = e_{t'}(\mathcal{S}_k) \cdot \theta_{t'}(\mathcal{S}_k)$  is

$$\ell_{k,t'} = \log \binom{Y_{t'}(\mathcal{S}_k) + \delta - 1}{Y_{t'}(\mathcal{S}_k)} + \delta \log \frac{\delta}{\delta + \mu_{k,t'}} + Y_{t'}(\mathcal{S}_k) \log \frac{\mu_{k,t'}}{\delta + \mu_{k,t'}}.$$

The combinatorial prefactor does not depend on  $\mu_{k,t'}$  and cancels in the ratio. The remaining terms give

$$\begin{aligned} \ell_{k,t'}^{\text{new}} - \ell_{k,t'}^{\text{cur}} &= \delta \log \frac{\delta + \mu_{k,t'}^{\text{cur}}}{\delta + \mu_{k,t'}^{\text{new}}} + Y_{t'}(\mathcal{S}_k) \left[ \log \frac{\mu_{k,t'}^{\text{new}}}{\mu_{k,t'}^{\text{cur}}} + \log \frac{\delta + \mu_{k,t'}^{\text{cur}}}{\delta + \mu_{k,t'}^{\text{new}}} \right] \\ &= (\delta + Y_{t'}(\mathcal{S}_k)) \log \frac{\delta + \mu_{k,t'}^{\text{cur}}}{\delta + \mu_{k,t'}^{\text{new}}} + Y_{t'}(\mathcal{S}_k) \log \frac{\mu_{k,t'}^{\text{new}}}{\mu_{k,t'}^{\text{cur}}}. \end{aligned} \quad (\text{A.9})$$

Summing over all observations linked to DZ  $\mathcal{A}_m$  at times  $t' \geq t$ , the total log likelihood ratio is

$$\log r_{\text{lik}} = \sum_{\substack{(k,t'): \eta_{km,t'} > 0, \\ t' \geq t}} \left[ (\delta + Y_{t'}(\mathcal{S}_k)) \log \frac{\delta + \mu_{k,t'}^{\text{cur}}}{\delta + \mu_{k,t'}^{\text{new}}} + Y_{t'}(\mathcal{S}_k) \log \frac{\mu_{k,t'}^{\text{new}}}{\mu_{k,t'}^{\text{cur}}} \right]. \quad (\text{A.10})$$

### Step 6: accept or reject

Accept the proposal with probability  $\min(1, \exp(\log r_{\text{prior}} + \log r_{\text{lik}}))$ . Upon acceptance, set  $\varepsilon_{m,t} \leftarrow \varepsilon_{m,t}^*$  and update the shadow vectors storing  $\theta_{t'}(\mathcal{A}_m)$  for  $t' = t, \dots, T$  and  $\theta_{t'}(\mathcal{S}_k)$  for all affected observations. All quantities are maintained in pre-allocated vectors and updated incrementally upon acceptance, avoiding re-computation through the compiled model graph.

After completing the sweep over all  $M$  sites for a given time slice, the sampler

writes the updated field values back to the model and triggers a full re-computation of the deterministic nodes and log-probabilities. The shadow vectors are then refreshed from the model to correct any numerical drift accumulated during incremental updates.

## A.4 Comparison with the default CAR\_normal sampler

NIMBLE’s default CAR\_normal sampler assigns a separate random walk Metropolis move to each site  $m$ , with each sampler maintaining its own adaptive proposal scale. The proposal  $\varepsilon_{m,t}^* \sim N(\varepsilon_{m,t}^{\text{cur}}, \text{scale}_m^2)$  does not incorporate the local neighbourhood structure, where the proposal variance is adapted purely from the site’s own acceptance history, without reference to the conditional prior variance  $1/d_m$ . When the adjacency graph has heterogeneous degree, as in the Glasgow Data Zone graph where  $d_m$  ranges from 1 to 17, the per-site adaptation must independently learn the appropriate scale for each node. More importantly, each per-site update evaluates the acceptance ratio by calling NIMBLE’s compiled dependency graph traversal (`model$calculate`), which recomputes all deterministic descendants of the updated node. In the spatio-temporal model, updating  $\varepsilon_{m,t}$  triggers re-computation of  $\phi_{m,t'}$ ,  $\theta_{t'}(\mathcal{A}_m)$  and  $\theta_{t'}(\mathcal{S}_k)$  for all  $t' \geq t$  and all linked GP observations, so that a full sweep of  $M = 746$  sites repeatedly traverses a substantial fraction of the model graph. The CW-MH sampler instead updates each site from pre-allocated shadow vectors in closed form (Equations (A.7) and (A.8)), bypassing the graph.

The cost of each per-site evaluation interacts with the AR(1) temporal structure. When  $\varepsilon_{m,t}$  changes by  $d$ , the effect on  $\theta_{t'}(\mathcal{A}_m)$  is scaled by  $\sigma_\phi \alpha^{t'-t}$  across the  $T - t + 1$  subsequent time slices (Equation (A.7)). Under the posterior value  $\hat{\alpha} \approx 0.99$ , the largest part of this propagation persists across most of the temporal dimension, amplifying the cost disparity between the two samplers and making mixing failures at individual sites visible across many time slices.

## A.5 Proposal adaptation and identifiability

The global tuning parameter  $s$  is adapted every 200 iterations to target a sweep-level acceptance rate of 30–50%. If the observed rate over the previous window exceeds 50%,  $s$  is multiplied by 1.1; if it falls below 30%,  $s$  is divided by the same factor; otherwise  $s$  is left unchanged. The adjustment rule is applied throughout the chain

rather than switched off at a fixed iteration, and it is not damped over time by construction, so the scheme is persistently adaptive. In practice the acceptance rate stabilises within the burn-in period and  $s$  changes negligibly thereafter; inference in Section 5.6 is based on post-burn-in draws during this regime, and no formal adaptive-MCMC convergence guarantees (Roberts and Rosenthal, 2007) are invoked.

A sum-to-zero constraint is imposed on the first-period field  $\phi_1$  after each sweep of that block, by subtracting the sample mean  $M^{-1} \sum_{m=1}^M \phi_1(\mathcal{A}_m)$  from all sites. This enforces the identifiability condition described in Section 5.3.2. The innovations  $\varepsilon_t$  for  $t \geq 2$  are not mean-centred, consistent with the model specification where the sum-to-zero constraint applies only at the initial time point.

# Bibliography

- Air Quality Expert Group (2004). Nitrogen dioxide in the United Kingdom. Report, Department for Environment, Food and Rural Affairs.
- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723.
- Alshammari, L., Large, D. J., Boyd, D. S., Sowter, A., Anderson, R., Andersen, R., and Marsh, S. (2018). Long-term peatland condition assessment via surface motion monitoring using the ISBAS DInSAR technique over the Flow Country, Scotland. *Remote Sensing*, 10(7):1103.
- Anderson, H. R. (2009). Air pollution and mortality: A history. *Atmospheric Environment*, 43(1):142–152.
- Bakar, K. S. and Sahu, S. K. (2015). spTimer: Spatio-temporal Bayesian modeling using r. *Journal of Statistical Software*, 63(15):1–32.
- Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2014). *Hierarchical Modeling and Analysis for Spatial Data*. CRC Press.
- Beckwith, M., Bates, E., Gillah, A., and Carslaw, N. (2019). NO<sub>2</sub> hotspots: Are we measuring in the right places? *Atmospheric Environment: X*, 2:100025.
- Bernardinelli, L., Clayton, D., Pascutto, C., Montomoli, C., Ghislandi, M., and Songini, M. (1995). Bayesian analysis of space-time variation in disease risk. *Statistics in Medicine*, 14(21-22):2433–2443.
- Berrocal, V. J., Gelfand, A. E., and Holland, D. M. (2010). A spatio-temporal downscaler for output from numerical models. *Journal of Agricultural, Biological, and Environmental Statistics*, 15(2):176.

- Berrocal, V. J., Guan, Y., Muyskens, A., Wang, H., Reich, B. J., Mulholland, J. A., and Chang, H. H. (2020). A comparison of statistical and machine learning methods for creating national daily maps of ambient PM<sub>2.5</sub> concentration. *Atmospheric Environment*, 222:117130.
- Besag, J., York, J., and Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, 43:1–20.
- Blangiardo, M., Cameletti, M., Baio, G., and Rue, H. (2013). Spatial and spatio-temporal models with R-INLA. *Spatial and Spatio-temporal Epidemiology*, 4:33–49.
- Blangiardo, M., Finazzi, F., and Cameletti, M. (2016). Two-stage Bayesian model to evaluate the effect of air pollution on chronic respiratory diseases using drug prescriptions. *Spatial and Spatio-temporal Epidemiology*, 18:1–12.
- Bradley, J. R., Wikle, C. K., and Holan, S. H. (2016). Bayesian spatial change of support for count-valued survey data with application to the American Community Survey. *Journal of the American Statistical Association*, 111(514):472–487.
- Brauer, M., Freedman, G., Frostad, J., Van Donkelaar, A., Martin, R. V., Dentener, F., van Dingenen, R., Estep, K., Amini, H., Apte, J. S., et al. (2016). Ambient air pollution exposure estimation for the global burden of disease 2013. *Environmental Science & Technology*, 50(1):79–88.
- Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24:123–140.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45:5–32.
- Breiman, L., Friedman, J., Olshen, R. A., and Stone, C. J. (1984). *Classification and Regression Trees*. Chapman and Hall/CRC, 1 edition.
- Brooks, S., Gelman, A., Jones, G., and Meng, X.-L. (2011). *Handbook of Markov Chain Monte Carlo*. CRC Press.
- Brunekreef, B. and Holgate, S. T. (2002). Air pollution and health. *The Lancet*, 360(9341):1233–1242.
- Brunsdon, C., Fotheringham, S., and Charlton, M. (1998). Geographically weighted regression—modelling spatial non-stationarity. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 47(3):431–443.

- Byun, D. and Schere, K. L. (2006). Review of the governing equations, computational algorithms, and other components of the Models-3 Community Multiscale Air Quality (CMAQ) modeling system. *Applied Mechanics Reviews*, 59(2):51–77.
- Caillaud, D., Leynaert, B., Keirsbulck, M., and Nadif, R. (2018). Indoor mould exposure, asthma and rhinitis: findings from systematic reviews and recent longitudinal studies. *European Respiratory Review*, 27(148).
- Cameletti, M., Ignaccolo, R., and Bande, S. (2011). Comparing spatio-temporal models for particulate matter in Piemonte. *Environmetrics*, 22(8):985–996.
- Cameletti, M., Lindgren, F., Simpson, D., and Rue, H. (2013). Spatio-temporal modeling of particulate matter concentration through the SPDE approach. *AStA Advances in Statistical Analysis*, 97:109–131.
- Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective*. Chapman and Hall/CRC, 2 edition.
- Chapman, K., Mannino, D., Soriano, J., Vermeire, P., Buist, A., Thun, M., Connell, C., Jemal, A., Lee, T., Miravitlles, M., et al. (2006). Epidemiology and costs of chronic obstructive pulmonary disease. *European Respiratory Journal*, 27(1):188–207.
- Chauhan, A. J., Inskip, H. M., Linaker, C. H., Smith, S., Schreiber, J., Johnston, S. L., and Holgate, S. T. (2003). Personal exposure to nitrogen dioxide (NO<sub>2</sub>) and the severity of virus-induced asthma in children. *The Lancet*, 361(9373):1939–1944.
- Chen, J., de Hoogh, K., Gulliver, J., Hoffmann, B., Hertel, O., Ketzel, M., Bauwelinck, M., Van Donkelaar, A., Hvidtfeldt, U. A., Katsouyanni, K., et al. (2019). A comparison of linear regression, regularization, and machine learning algorithms to develop Europe-wide spatial models of fine particles and nitrogen dioxide. *Environment International*, 130:104934.
- Chen, X., Zhou, C.-W., Fu, Y.-Y., Li, Y.-Z., Chen, L., Zhang, Q.-W., and Chen, Y.-F. (2023). Global, regional, and national burden of chronic respiratory diseases and associated risk factors, 1990–2019: Results from the Global Burden of Disease Study 2019. *Frontiers in Medicine*, 10:1066804.
- Chief Medical Officer (2022). Chief Medical Officer’s annual report 2022: Air pollution. Report, Department of Health and Social Care.

- Chowdhury, N. U., Guntur, V. P., Newcomb, D. C., and Wechsler, M. E. (2021). Sex and gender in asthma. *European Respiratory Review*, 30(162).
- Clayton, D. and Kaldor, J. (1987). Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics*, pages 671–681.
- Cocchi, D., Greco, F., and Trivisano, C. (2007). Hierarchical space-time modelling of PM<sub>10</sub> pollution. *Atmospheric Environment*, 41(3):532–542.
- Cohen, A. J., Brauer, M., Burnett, R., Anderson, H. R., Frostad, J., Estep, K., Balakrishnan, K., Brunekreef, B., Dandona, L., Dandona, R., et al. (2017). Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: an analysis of data from the Global Burden of Diseases Study 2015. *The Lancet*, 389(10082):1907–1918.
- COMEAP (2023). Summary of COMEAP recommendations for quantification of health effects associated with air pollutants: 2023. Technical report, Committee on the Medical Effects of Air Pollutants.
- Cowley, J., Kiely, J., and Collins, D. (2016). Unravelling the Glasgow Effect: The relationship between accumulative bio-psychosocial stress, stress reactivity and Scotland’s health problems. *Preventive Medicine Reports*, 4:370–375.
- Cox, M. G. (1972). The numerical evaluation of B-Splines. *IMA Journal of Applied Mathematics*, 10(2):134–149.
- Cressie, N. (2015). *Statistics for Spatial Data*. John Wiley & Sons.
- D’Amato, M., Molino, A., Calabrese, G., Cecchi, L., Annesi-Maesano, I., and D’Amato, G. (2018). The impact of cold on the respiratory tract and its consequences to respiratory health. *Clinical and Translational Allergy*, 8:1–8.
- De Boor, C. (1972). On calculating with B-Splines. *Journal of Approximation Theory*, 6(1):50–62.
- De Boor, C. (1978). *A Practical Guide to Splines*, volume 27. Springer-Verlag New York.
- de Valpine, P., Turek, D., Paciorek, C., Anderson-Bergman, C., Temple Lang, D., and Bodik, R. (2017). Programming with models: writing statistical algorithms for general model structures with NIMBLE. *Journal of Computational and Graphical Statistics*, 26:403–417.

- Delmas, M.-C., Bénézet, L., Ribet, C., Iwatsubo, Y., Zins, M., Nadif, R., Roche, N., and Leynaert, B. (2021). Underdiagnosis of obstructive lung disease: findings from the French CONSTANCES cohort. *BMC Pulmonary Medicine*, 21:1–10.
- Department for Environment, Food and Rural Affairs (2011). The air quality strategy for England, Scotland, Wales and Northern Ireland. Technical report, Department for Environment, Food and Rural Affairs.
- Department for Environment, Food and Rural Affairs (2019). Clean air strategy 2019. Technical report, Department for Environment, Food and Rural Affairs.
- Department for Environment, Food and Rural Affairs (2020). Air pollution in the UK 2019. Technical report, Department for Environment, Food and Rural Affairs, London, United Kingdom.
- Department for Environment, Food and Rural Affairs (2023). Air quality statistics in the UK, 1987 to 2022 - Particulate matter ( $\text{pm}_{10}/\text{pm}_{2.5}$ ). Technical report, Department for Environment, Food and Rural Affairs, London, United Kingdom. Updated 27 April 2023.
- Department for Environment, Food and Rural Affairs (2025). Air quality statistics in the UK, 1987 to 2024. Technical report, Department for Environment, Food and Rural Affairs.
- Dierckx, P. (1995). *Curve and Surface Fitting with Splines*. Oxford University Press.
- Diggle, P. J., Menezes, R., and Su, T.-L. (2010). Geostatistical inference under preferential sampling. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 59(2):191–232.
- Diggle, P. J. and Ribeiro, P. J. (2007). *Model-based Geostatistics*. Springer.
- Dobson, A. J. and Barnett, A. G. (2018). *An Introduction to Generalized Linear Models*. Chapman and Hall/CRC, 4 edition.
- Eilers, P. H. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, 11(2):89–121.
- European Parliament and Council (2024). Directive (EU) 2024/2881 on ambient air quality and cleaner air for Europe (recast). Accessed: 2026-03-08.



- Flowerdew, R. and Green, M. (1992). Developments in areal interpolation methods and GIS. *The Annals of Regional Science*, 26:67–78.
- Fox, J. and Monette, G. (1992). Generalized collinearity diagnostics. *Journal of the American Statistical Association*, 87(417):178–183.
- GBD 2013 Risk Factors Collaborators (2015). Global, regional, and national comparative risk assessment of 79 behavioural, environmental and occupational, and metabolic risks or clusters of risks in 188 countries, 1990–2013: A systematic analysis for the Global Burden of Disease Study 2013. *The Lancet*, 386(10010):2287–2323.
- Gelfand, A. E., Zhu, L., and Carlin, B. P. (2001). On the change of support problem for spatio-temporal data. *Biostatistics*, 2(1):31–45.
- Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis*, 1(3).
- Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis*. Chapman and Hall/CRC, third edition.
- Gelman, A., Hwang, J., and Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, 24(6):997–1016.
- Gelman, A., Roberts, G. O., and Gilks, W. R. (1996). Efficient Metropolis jumping rules. *Bayesian Statistics 5*, 5:599–608.
- Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–472.
- Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6(6):721–741.
- Georganos, S., Grippa, T., Niang Gadiaga, A., Linard, C., Lennert, M., Vanhuyse, S., Mboga, N., Wolff, E., and Kalogirou, S. (2021). Geographical random forests: a spatial extension of the random forest algorithm to address spatial heterogeneity in remote sensing and population modelling. *Geocarto International*, 36(2):121–136.
- Gilbert, N. L., Goldberg, M. S., Beckerman, B., Brook, J. R., and Jerrett, M. (2005). Assessing spatial variability of ambient nitrogen dioxide in Montréal, Canada, with

- a land-use regression model. *Journal of the Air & Waste Management Association*, 55(8):1059–1063.
- Gilkes, A., Hull, S., Durbaba, S., Schofield, P., Ashworth, M., Mathur, R., and White, P. (2017). Ethnic differences in smoking intensity and COPD risk: an observational study in primary care. *NPJ Primary Care Respiratory Medicine*, 27(1):50.
- Gramatica, M., Congdon, P., and Liverani, S. (2021). Bayesian modelling for spatially misaligned health areal data: A multiple membership approach. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 70(3):645–666.
- Gramatica, M., Liverani, S., and Congdon, P. (2024). Structure induced by a multiple membership transformation on the conditional autoregressive model. *Bayesian Analysis*, 19(3):725–749.
- Guo, J. and Feng, Z. (2019). Study on spatial temporal distribution characteristics of air quality index in Beijing and its correlation with local meteorological conditions. *Discrete Dynamics in Nature and Society*, 2019.
- Hansel, N. N., McCormack, M. C., and Kim, V. (2016). The effects of air pollution and temperature on COPD. *COPD: Journal of Chronic Obstructive Pulmonary Disease*, 13(3):372–379.
- Hart, J. E., Yanosky, J. D., Puett, R. C., Ryan, L., Dockery, D. W., Smith, T. J., Garshick, E., and Laden, F. (2009). Spatial modeling of PM<sub>10</sub> and NO<sub>2</sub> in the continental United States, 1985–2000. *Environmental Health Perspectives*, 117(11):1690–1696.
- Hastie, T. and Tibshirani, R. (1986). Generalized additive models. *Statistical Science*, 1(3):297–310.
- Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models*. Chapman and Hall/CRC, first edition.
- Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109.
- Hayes Jr, D., Collins, P. B., Khosravi, M., Lin, R.-L., and Lee, L.-Y. (2012). Bronchoconstriction triggered by breathing hot humid air in patients with asthma: Role of cholinergic reflex. *American Journal of Respiratory and Critical Care Medicine*, 185(11):1190–1196.

- Hengl, T., Heuvelink, G. B., Kempen, B., Leenaars, J. G., Walsh, M. G., Shepherd, K. D., Sila, A., MacMillan, R. A., Mendes de Jesus, J., Tamene, L., et al. (2015). Mapping soil properties of Africa at 250 m resolution: Random forests significantly improve current predictions. *PLOS One*, 10(6):e0125814.
- Hou, K. and Xu, X. (2022). Evaluation of the influence between local meteorology and air quality in Beijing using generalized additive models. *Atmosphere*, 13(1):24.
- Hu, K., Rahman, A., Bhrugubanda, H., and Sivaraman, V. (2017a). Hazeest: Machine learning based metropolitan air pollution estimation from fixed and mobile sensors. *IEEE Sensors Journal*, 17(11):3517–3525.
- Hu, X., Belle, J. H., Meng, X., Wildani, A., Waller, L. A., Strickland, M. J., and Liu, Y. (2017b). Estimating PM<sub>2.5</sub> concentrations in the conterminous United States using the random forest approach. *Environmental Science & Technology*, 51(12):6936–6944.
- Huang, G., Lee, D., and Scott, E. M. (2018). Multivariate space-time modelling of multiple air pollutants and their health effects accounting for exposure uncertainty. *Statistics in Medicine*, 37(7):1134–1148.
- International Agency for Research on Cancer (2013). Outdoor air pollution a leading environmental cause of cancer deaths. Technical Report Press Release No. 221, World Health Organization.
- James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An Introduction to Statistical Learning: with Applications in R*. Springer.
- Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 186(1007):453–461.
- Kampa, M. and Castanas, E. (2008). Human health effects of air pollution. *Environmental Pollution*, 151(2):362–367.
- Khomenko, S., Cirach, M., Pereira-Barboza, E., Mueller, N., Barrera-Gómez, J., Rojas-Rueda, D., de Hoogh, K., Hoek, G., and Nieuwenhuijsen, M. (2021). Premature mortality due to air pollution in European cities: a health impact assessment. *The Lancet Planetary Health*, 5(3):e121–e134.

- Knorr-Held, L. (2000). Bayesian modelling of inseparable space-time variation in disease risk. *Statistics in Medicine*, 19(17-18):2555–2567.
- Kosmidis, C., Hashad, R., Mathioudakis, A. G., McCahery, T., Richardson, M. D., and Vestbo, J. (2023). Impact of self-reported environmental mould exposure on COPD outcomes. *Pulmonology*, 29(5):375–384.
- Larkin, A., Geddes, J. A., Martin, R. V., Xiao, Q., Liu, Y., Marshall, J. D., Brauer, M., and Hystad, P. (2017). Global land use regression model for nitrogen dioxide air pollution. *Environmental Science & Technology*, 51(12):6957–6964.
- Lawson, A. B. (2018). *Bayesian Disease Mapping: Hierarchical Modeling in Spatial Epidemiology*. Chapman and Hall/CRC, third edition.
- Lawson, A. B. and Lee, D. (2017). Bayesian disease mapping for public health. In *Handbook of Statistics*, volume 36, pages 443–481. Elsevier.
- Lee, A., Chiu, Y.-H. M., Rosa, M. J., Jara, C., Wright, R. O., Coull, B. A., and Wright, R. J. (2016). Prenatal and postnatal stress and asthma in children: Temporal- and sex-specific associations. *Journal of Allergy and Clinical Immunology*, 138(3):740–747.
- Lee, D. (2018). A locally adaptive process-convolution model for estimating the health impact of air pollution. *The Annals of Applied Statistics*, 12(4):2540–2558.
- Lee, D. (2020). A tutorial on spatio-temporal disease risk modelling in R using Markov chain Monte Carlo simulation and the CARBayesST package. *Spatial and Spatio-temporal Epidemiology*, 34:100353.
- Lee, D. (2023). Identifying boundaries in spatially continuous risk surfaces from spatially aggregated disease count data. *The Annals of Applied Statistics*, 17(4).
- Lee, D. (2025). Computationally efficient spatio-temporal disease mapping for big data. *Spatial Statistics*, 67:100901.
- Lee, D. and Anderson, C. (2023). Delivering spatially comparable inference on the risks of multiple severities of respiratory disease from spatially misaligned disease count data. *Biometrics*, 79(3):2691–2704.
- Lee, D., Ferguson, C., and Mitchell, R. (2009). Air pollution and health in Scotland: A multicity study. *Biostatistics*, 10(3):409–423.

- Lee, D., Mukhopadhyay, S., Rushworth, A., and Sahu, S. K. (2017). A rigorous statistical framework for spatio-temporal pollution prediction and estimation of its long-term impact on health. *Biostatistics*, 18(2):370–385.
- Lee, D., Rushworth, A., and Napier, G. (2018). Spatio-temporal areal unit modeling in R with conditional autoregressive priors using the CARBayesST package. *Journal of Statistical Software*, 84(9):1–39.
- Leroux, B. G., Lei, X., and Breslow, N. (2000). Estimation of disease rates in small areas: A new mixed model for spatial dependence. In *Statistical Models in Epidemiology, the Environment, and Clinical Trials*, pages 179–191. Springer.
- Li, L., Wu, J., Wilhelm, M., and Ritz, B. (2012). Use of generalized additive models and cokriging of spatial residuals to improve land-use regression estimates of nitrogen oxides in Southern California. *Atmospheric Environment*, 55:220–228.
- Li, M., Li, Q., Yang, G., Kolosov, V. P., Perelman, J. M., and Zhou, X. D. (2011). Cold temperature induces mucin hypersecretion from normal human bronchial epithelial cells in vitro through a transient receptor potential melastatin 8 (TRPM8)-mediated mechanism. *Journal of Allergy and Clinical Immunology*, 128(3):626–634.
- Lin, S., Luo, M., Walker, R. J., Liu, X., Hwang, S.-A., and Chinery, R. (2009). Extreme high temperatures and hospital admissions for respiratory and cardiovascular diseases. *Epidemiology*, 20(5):738–746.
- Lindén, A. and Mäntyniemi, S. (2011). Using the negative binomial distribution to model overdispersion in ecological count data. *Ecology*, 92(7):1414–1421.
- Lindgren, F. and Rue, H. (2015). Bayesian spatial modelling with R-INLA. *Journal of Statistical Software*, 63:1–25.
- Lindgren, F., Rue, H., and Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(4):423–498.
- Liu, B.-C., Binaykia, A., Chang, P.-C., Tiwari, M. K., and Tsao, C.-C. (2017). Urban air quality forecasting based on multi-dimensional collaborative Support Vector Regression (SVR): A case study of Beijing-Tianjin-Shijiazhuang. *PLOS One*, 12(7):e0179763.

- Liu, Y., Cao, G., Zhao, N., Mulligan, K., and Ye, X. (2018). Improve ground-level PM<sub>2.5</sub> concentration mapping using a random forests-based geostatistical approach. *Environmental Pollution*, 235:272–282.
- Liu, Y., Paciorek, C. J., and Koutrakis, P. (2009). Estimating regional spatial and temporal variability of PM<sub>2.5</sub> concentrations using satellite data, meteorology, and land use information. *Environmental Health Perspectives*, 117(6):886–892.
- Liu, Y., Wang, P., Li, Y., Wen, L., and Deng, X. (2022). Air quality prediction models based on meteorological factors and real-time data of industrial waste gas. *Scientific Reports*, 12(1):1–15.
- Luo, Y., Yan, J., and McClure, S. (2021). Distribution of the environmental and socioeconomic risk factors on COVID-19 death rate across continental USA: a spatial nonlinear analysis. *Environmental Science and Pollution Research*, 28:6587–6599.
- MacNab, Y. C. (2022). Bayesian disease mapping: Past, present, and future. *Spatial Statistics*, 50:100593. Special Issue: The Impact of Spatial Statistics.
- Manisalidis, I., Stavropoulou, E., Stavropoulos, A., and Bezirtzoglou, E. (2020). Environmental and health impacts of air pollution: A review. *Frontiers in Public Health*, 8:14.
- Mardia, K. V. and Goodall, C. R. (1993). Spatial-temporal analysis of multivariate environmental monitoring data. *Multivariate Environmental Statistics*, 6(76):347–385.
- McKernan, M. (2004). *Northern Ireland Yearbook 2005: A Comprehensive Reference Guide to the Political, Economic and Social Life of Northern Ireland*. The Stationery Office.
- Mead, J. (1980). Dysanapsis in normal lungs assessed by the relationship between maximal flow, static recoil, and vital capacity. *American Review of Respiratory Disease*, 121(2):339–342.
- Meinshausen, N. and Ridgeway, G. (2006). Quantile regression forests. *Journal of Machine Learning Research*, 7(6).
- Meng, X., Hand, J. L., Schichtel, B. A., and Liu, Y. (2018). Space-time trends of PM<sub>2.5</sub> constituents in the conterminous United States estimated by a machine learning approach, 2005–2015. *Environment International*, 121:1137–1147.

- Met Office, Hollis, D., McCarthy, M., Kendon, M., and Legg, T. (2022). HadUK-Grid gridded climate observations on a 1km grid over the UK, v1.1.0.0 (1836-2021).
- Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092.
- Meyer, H., Reudenbach, C., Hengl, T., Katurji, M., and Nauss, T. (2018). Improving performance of spatio-temporal machine learning models using forward feature selection and target-oriented validation. *Environmental Modelling & Software*, 101:1–9.
- Moran, P. A. P. (1950). Notes on continuous stochastic phenomena. *Biometrika*, 37(1/2):17–23.
- Mu, Z., Chen, P.-L., Geng, F.-H., Ren, L., Gu, W.-C., Ma, J.-Y., Peng, L., and Li, Q.-Y. (2017). Synergistic effects of temperature and humidity on the symptoms of COPD patients. *International Journal of Biometeorology*, 61:1919–1925.
- Mukhopadhyay, S. and Sahu, S. K. (2017). A Bayesian spatiotemporal model to estimate long-term exposure to outdoor air pollution at coarser administrative geographies in England and Wales. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 181(2):465–486.
- Neal, R. M. (2003). Slice sampling. *The Annals of Statistics*, 31(3):705–767.
- Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135(3):370–384.
- NICE (2025a). Clinical Knowledge Summaries: Asthma.
- NICE (2025b). Clinical Knowledge Summaries: Chronic Obstructive Pulmonary Disease.
- Novotny, E. V., Bechle, M. J., Millet, D. B., and Marshall, J. D. (2011). National satellite-based land-use regression: NO<sub>2</sub> in the United States. *Environmental Science & Technology*, 45(10):4407–4414.
- Office for Health Improvement and Disparities (2022). Air pollution: applying All Our Health. Technical report, UK Government.

- Public Health Scotland (2022). GP workforce and practice list sizes 2012–2022. Statistical report, Public Health Scotland.
- Public Health Scotland (2024). Outdoor air pollution and health. Accessed: 2026-03-08.
- Pugsley, K. L., Stedman, J. R., Brookes, D. M., Kent, A. J., Morris, R. J., Whiting, S. L., Wareham, J. V., and Goodhand, A. (2022). Technical report on UK supplementary modelling assessment under the Air Quality Standards Regulations 2010 for 2020. Technical report, Ricardo Energy & Environment, Harwell Oxford, United Kingdom.
- Raherison, C. and Girodet, P.-O. (2009). Epidemiology of COPD. *European Respiratory Review*, 18(114):213–221.
- Rajagopalan, S., Al-Kindi, S. G., and Brook, R. D. (2018). Air pollution and cardiovascular disease: JACC state-of-the-art review. *Journal of the American College of Cardiology*, 72(17):2054–2070.
- Reid, C. E., Jerrett, M., Petersen, M. L., Pfister, G. G., Morefield, P. E., Tager, I. B., Raffuse, S. M., and Balme, J. R. (2015). Spatiotemporal prediction of fine particulate matter during the 2008 northern California wildfires using machine learning. *Environmental Science & Technology*, 49(6):3887–3896.
- Reinsch, C. H. (1967). Smoothing by spline functions. *Numerische Mathematik*, 10(3):177–183.
- Ren, X., Mi, Z., and Georgopoulos, P. G. (2020). Comparison of Machine Learning and Land Use Regression for fine scale spatiotemporal estimation of ambient air pollution: Modeling ozone concentrations across the contiguous United States. *Environment International*, 142:105827.
- Rigby, R. A. and Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(3):507–554.
- Robert, C. P. and Casella, G. (2005). *Monte Carlo Statistical Methods (Springer Texts in Statistics)*. Springer Verlag, New York, United States.
- Roberts, G. O. and Rosenthal, J. S. (2007). Coupling and ergodicity of adaptive Markov chain Monte Carlo algorithms. *Journal of Applied Probability*, 44(2):458–475.



- Ross, Z., English, P. B., Scalf, R., Gunier, R., Smorodinsky, S., Wall, S., and Jerrett, M. (2006). Nitrogen dioxide prediction in Southern California using land use regression modeling: potential for environmental health analyses. *Journal of Exposure Science & Environmental Epidemiology*, 16(2):106–114.
- Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*. CRC Press.
- Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2):319–392.
- Rural and Environment Science and Analytical Services Division (2022). Scottish Government Urban Rural Classification 2020. Report, Scottish Government.
- Rushworth, A., Lee, D., and Mitchell, R. (2014). A spatio-temporal model for estimating the long-term effects of air pollution on respiratory hospital admissions in Greater London. *Spatial and Spatio-temporal Epidemiology*, 10:29–38.
- Sahu, S. (2022). *Bayesian Modeling of Spatio-Temporal Data With R*. CRC Press.
- Sahu, S. K. (2012). Hierarchical Bayesian models for space-time air pollution data. In *Handbook of Statistics*, volume 30, pages 477–495. Elsevier.
- Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464.
- Scottish Government (2010). The air quality standards (Scotland) regulations 2010. As amended by The Air Quality Standards (Scotland) Amendment Regulations 2016.
- Scottish Government (2021). Cleaner Air for Scotland 2: Towards a Better Place for Everyone. Technical report, Scottish Government.
- Scottish Government (2023). The Scottish Health Survey: 2022 Edition. Report, Scottish Government.
- Shaddick, G., Thomas, M. L., Green, A., Brauer, M., Donkelaar, A., Burnett, R., Chang, H. H., Cohen, A., Dingenen, R. V., Dora, C., et al. (2018). Data integration model for air quality: A hierarchical approach to the global estimation of exposures to ambient air pollution. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 67(1):231–253.

- Sin, D. D., Doiron, D., Agusti, A., Anzueto, A., Barnes, P. J., Celli, B. R., Criner, G. J., Halpin, D., Han, M. K., Martinez, F. J., et al. (2023). Air pollution and COPD: GOLD 2023 committee report. *European Respiratory Journal*, 61(5).
- Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 64(4):583–639.
- Stasinopoulos, M. D., Rigby, R. A., Heller, G. Z., Voudouris, V., and De Bastiani, F. (2017). *Flexible Regression and Smoothing: Using GAMLSS in R*. CRC Press.
- Stern, H. S. and Cressie, N. (2000). Posterior predictive model checks for disease mapping models. *Statistics in Medicine*, 19(17-18):2377–2397.
- Tibbits, M. M., Groendyke, C., Haran, M., and Liechty, J. C. (2014). Automated factor slice sampling. *Journal of Computational and Graphical Statistics*, 23(2):543–563.
- Tobler, W. R. (1970). A computer movie simulating urban growth in the Detroit region. *Economic Geography*, 46:234–240.
- Tran, H. N., Siu, S., Iribarren, C., Udaltsova, N., and Klatsky, A. L. (2011). Ethnicity and risk of hospitalization for asthma and chronic obstructive pulmonary disease. *Annals of Epidemiology*, 21(8):615–622.
- Ugarte, M. D., Etxeberria, J., Goicoa, T., and Ardanaz, E. (2012). Gender-specific spatio-temporal patterns of colorectal cancer incidence in Navarre, Spain (1990–2005). *Cancer Epidemiology*, 36(3):254–262.
- UK Government (2022). Air pollution: applying All Our Health. Report, Office for Health Improvement and Disparities.
- UK Parliament (1993). Clean air act 1993.
- UK Parliament (2021). Environment act 2021.
- Van den Bossche, J., De Baets, B., Verwaeren, J., Botteldooren, D., and Theunis, J. (2018). Development and evaluation of land use regression models for black carbon based on bicycle and pedestrian measurements in the urban environment. *Environmental Modelling & Software*, 99:58–69.

- Van Donkelaar, A., Martin, R. V., Brauer, M., Hsu, N. C., Kahn, R. A., Levy, R. C., Lyapustin, A., Sayer, A. M., and Winker, D. M. (2016). Global estimates of fine particulate matter using a combined geophysical-statistical method with information from satellites, models, and monitors. *Environmental Science & Technology*, 50(7):3762–3772.
- Van Niekerk, J., Krainski, E., Rustand, D., and Rue, H. (2023). A new avenue for Bayesian inference with INLA. *Computational Statistics & Data Analysis*, 181:107692.
- Van Niekerk, J. and Rue, H. (2024). Low-rank variational Bayes correction to the Laplace method. *Journal of Machine Learning Research*, 25(62):1–25.
- Ver Hoef, J. M. and Boveng, P. L. (2007). Quasi-Poisson vs. negative binomial regression: How should we model overdispersed count data? *Ecology*, 88(11):2766–2772.
- Walsh, D., Bendel, N., Jones, R., and Hanlon, P. (2010). It’s not just deprivation: Why do equally deprived UK cities experience different health outcomes? *Public Health*, 124(9):487–495.
- Wang, M., Sampson, P. D., Hu, J., Kleeman, M., Keller, J. P., Olives, C., Szpiro, A. A., Vedal, S., and Kaufman, J. D. (2016). Combining land-use regression and chemical transport modeling in a spatiotemporal geostatistical model for ozone and PM<sub>2.5</sub>. *Environmental Science & Technology*, 50(10):5111–5118.
- Wang, Y. (2011). *Smoothing Splines: Methods and Applications*. CRC Press.
- Watanabe, S. and Opper, M. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11(116):3571–3594.
- Whyte, B., Young, M., and Timpson, K. (2021). Health in a changing city: Glasgow 2021. A study of changes in health, demographic, socioeconomic and environmental factors in Glasgow over the last 20 years. Technical report, Glasgow Centre for Population Health.
- Wikle, C. K. (2003). Hierarchical models in environmental science. *International Statistical Review*, 71(2):181–199.

- Wikle, C. K., Berliner, L. M., and Cressie, N. (1998). Hierarchical Bayesian space-time models. *Environmental and Ecological Statistics*, 5:117–154.
- Wood, S. N. (2017). *Generalized Additive Models: An Introduction With R*. Chapman and Hall/CRC, second edition.
- Wood, S. N., Pya, N., and Säfken, B. (2016). Smoothing parameter and model selection for general smooth models. *Journal of the American Statistical Association*, 111(516):1548–1563.
- World Health Organization (2021). *WHO Global Air Quality Guidelines: Particulate Matter PM<sub>2.5</sub> and PM<sub>10</sub>, Ozone, Nitrogen Dioxide, Sulfur Dioxide and Carbon Monoxide*. World Health Organization.
- World Health Organization (2024). Ambient (outdoor) air pollution.
- World Health Organization (2025). Air pollution data portal. Accessed: 7 March 2026.
- Wright, M. N. and Ziegler, A. (2017). Ranger: A fast implementation of Random Forests for high dimensional data in C++ and R. *Journal of Statistical Software*, 77(1):1–17.
- Wu, Q. and Lin, H. (2019). A novel optimal-hybrid model for daily air quality index prediction considering air pollutant factors. *Science of the Total Environment*, 683:808–821.
- Zhan, Y., Luo, Y., Deng, X., Chen, H., Grieneisen, M. L., Shen, X., Zhu, L., and Zhang, M. (2017). Spatiotemporal prediction of continuous daily PM<sub>2.5</sub> concentrations across China using a spatially explicit machine learning algorithm. *Atmospheric Environment*, 155:129–139.
- Zhan, Y., Luo, Y., Deng, X., Grieneisen, M. L., Zhang, M., and Di, B. (2018). Spatiotemporal prediction of daily ambient ozone levels across China using random forest for human exposure assessment. *Environmental Pollution*, 233:464–473.
- Zhou, X., Sampath, V., and Nadeau, K. C. (2024). Effect of air pollution on asthma. *Annals of Allergy, Asthma & Immunology*, 132(4):426–432.
- Zhu, Q., Lee, D., and Stoner, O. (2024). A comparison of statistical and machine learning models for spatio-temporal prediction of ambient air pollutant concentrations in Scotland. *Environmental and Ecological Statistics*, 31(4):1085–1108.

- Zhu, Q., Lee, D., and Stoner, O. (2026). Quantifying the effects of air pollution on respiratory ill health treated in primary care when the locations of the populations at risk are partially unknown. *Statistical Methods in Medical Research*, 35(6):1215–1229.