MA50264: Inter-disciplinary research projects

Air pollution as a driver of the epidemiological dynamics of influenza and other respiratory infections

Chiara Boetti [cb2605@bath.ac.uk]

Final report — A.Y. 2022-2023

University of Bath
EPSRC Centre for Doctoral Training in Statistical Applied Mathematics (SAMBa)

1 Introduction

1.1 Overview and Motivations

According to the World Health Organization (2018), influenza is a respiratory illness that affects millions of people globally every year, leading to significant morbidity, mortality, and economic burden. Recent studies suggest that environmental factors, such as air pollution, can increase the susceptibility, transmissibility, and severity of the disease (Chen et al. 2017, Singer et al. 2020, Domingo & Rovira 2020). Therefore, investigating the relationship between influenza cases and pollution levels is crucial for improving our understanding of the disease and designing effective public health interventions.

The aim of this research is to explore the correlation between influenza cases and PM10 levels in Sweden by conducting a spatial statistical analysis. Despite being regarded as a country with relatively low levels of air pollution compared to other nations, Sweden still experiences high pollution levels in various urban areas, particularly during winter months when wood-burning stoves are commonly used for heating (IQAir 2023). Furthermore, due to transportation and industrial activities, pollutants such as nitrogen oxides and particulate matter can exceed recommended levels in some regions, posing a health hazard to certain areas and populations. On the other hand, influenza is a notable public health issue in Sweden, as epidemics of the disease are experienced every winter and result in a high incidence of sickness, hospitalizations, and even death. According to the Swedish Public Health Agency (2023), there were over $20,600$ laboratory-confirmed influenza cases during the 2017-2018 influenza season. Hence, by investigating the spatial distribution of influenza cases and pollution levels, we can determine high-risk populations and regions, as well as examine the disease's transmissibility in various areas.

The report is organised as follows. After having described our data, we will focus on the methodology, where we discuss the approaches undertaken to formulate our models. We will then apply these methods to our data, and analyse both the pollution and the influenza estimates. Although our work could allow us to understand the future spread of influenza based on the air pollution dynamics, our results can not be critically used. Indeed, further modifications would be necessary in order to provide reliable feedbacks on the role of pollution as a driver of influenza dynamics. The final two sections will regard the connections to other research strands and the directions for future research.

1.2 Data

We collected open-source data on pollution, weather, and influenza-like illness (ILI) in Sweden during the 2017-2018 season. The data comprise:

(a) Pollutant measurements from 44 monitoring stations across the country, provided by the European Environment Agency (2019). Here, we focus on the daily mean concentrations of PM10 ($\mu g/m^3$).

(b) Daily weather data, including air temperature ($^\circ C$), precipitation amount (mm), relative humidity (%), wind speed (m/s) and air pressure(Pa), obtained from the Swedish Meteorological and Hydrological Institute (2023).

(c) Weekly reports on the count of positive laboratory-confirmed cases of influenza A and B, from the Swedish Public Health Agency (2023), for all 21 counties. Influenza data are seasonally available, i.e., from mid-November (week 44) to early May (week 18), and weeks 51-52 and the last six weeks are summarised in four fortnightly reports.

We also make use of the political boundaries of the 21 Swedish regions (Hijmans & University of California, Berkeley 2015), and the population density at approximately 1 km (WorldPop & Bondarenko 2020). Being a spatial analysis, the locations of data are crucial. Notice that weather and pollution records are local, with longitude and latitude coordinates, whereas influenza cases are area-level aggregated summaries. Two types of spatial information are displayed in Figure 1.

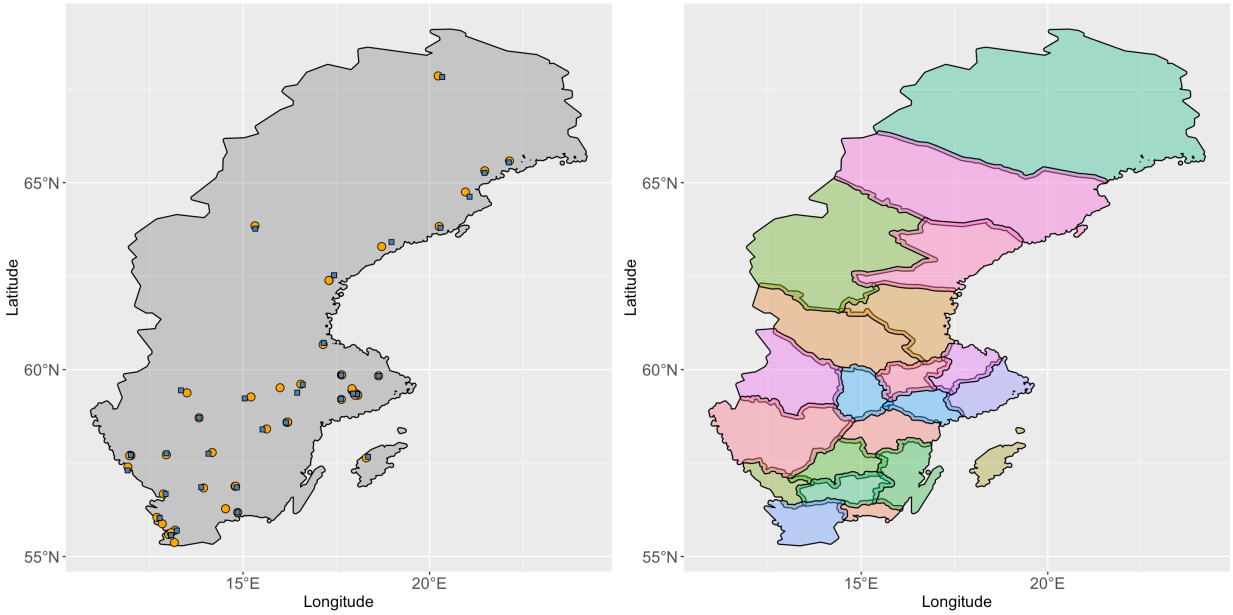

Figure 1: Left: pollution (orange circles) and weather (blue squares) stations in Sweden. Right: regions in Sweden.

2 Methodology and Model Formulation

In this section, the focus will be on the methodology used to analyse PM10 concentration and influenza during the 2017-2018 season. First, we will employ a hierarchical spatio-temporal model to define and implement a random field for modeling the pollution, as described by Cameletti et al. (2013). After having estimated the PM10 levels at every point in Sweden, the next step will be to incorporate them into the influenza model, following the approach taken by Wilson & Wakefield (2018). Within the second framework, we will fix the time index and analyse the incidence of influenza on a weekly basis. This approach has been adopted over incorporating the temporal dynamics in the model due to the computational complexity of the latter.

2.1 The Spatio-Temporal Model for PM10

The concentration of PM10 over Sweden can be reasonably described by means of a random field indexed in time, t, and in space, $\mathbf{s}$, (Cocchi et al. 2007, Cameletti et al. 2011, Sahu 2012):

$$Z(t,\mathbf{s}) = \left\{ z(t,\mathbf{s}) : (t,\mathbf{s}) \in D \subset \mathbb{R} \times \mathbb{R}^2 \right\}. \tag{1}$$

Our data correspond to discrete realizations of (1) measured at monitoring points sparse across the country, and can be exploited to infer the air pollution levels over the whole Sweden. Just as Cameletti et al. (2013), we assume (1) to be a Gaussian random field (GRF) and we define a hierarchical structure to model the PM10 concentration. In particular, the GRF $Z(t,\mathbf{s})$ is characterised by an intercept α_0 and various covariate information, specifically the altitude and the meteorological variables outlined in Section 1.2, with the respective coefficients $\alpha_1, \ldots, \alpha_p$. Besides these predictors, the model has a measurement error term, $\epsilon(t,\mathbf{s}) \sim N(0, \sigma^2_\epsilon)$, with temporal and spatial independent and identically distributed realizations. Finally, the state process $\xi(t,\mathbf{s})$ is incorporated into the model as the true, unobserved level of pollution. This component is a GRF and follows first-order autoregressive dynamics with coefficient a such that $|a| < 1$. Mathematically, the pollution model is defined as follows:

$$\begin{cases} Z(t,\mathbf{s}) = \alpha_0 + \sum_{j=1}^{p} \alpha_j X_j(t,\mathbf{s}) + \xi(t,\mathbf{s}) + \epsilon(t,\mathbf{s}) \\ \xi(t,\mathbf{s}) = a\xi(t-1,\mathbf{s}) + \omega(t,\mathbf{s}) \end{cases}. \tag{2}$$

The innovations $\omega(t,\mathbf{s})$ are time-independent, and have a zero-mean Gaussian distribution with covariance function

$$\mathrm{Cov}(\omega(t,\mathbf{s}), \omega(t',\mathbf{s}')) = \begin{cases} 0 & \text{if } t \neq t' \\ \sigma^2_\omega \mathcal{C}(d) & \text{if } t = t' \end{cases}.$$

Here, σ_ω is the general spatial variance parameter, $d = ||\mathbf{s} - \mathbf{s}'||_2$ is the distance between two points, and $\mathcal{C}(d)$ is the Matérn covariance function, i.e.,

$$\mathcal{C}(d) = \frac{2^{1-\nu}}{\Gamma(\nu)} (\kappa d)^\nu K_\nu (\kappa d). \tag{3}$$

Such a GRF $\omega(t,\mathbf{s})$ is called Matérn field, and its theoretical properties can be found in Lindgren et al. (2011). Of course, parameters in (3) determine the spatial features of $\omega(t,\mathbf{s})$. For instance, ν controls the smoothness of the spatial process, whereas $\kappa = \frac{\sqrt{8\nu}}{\rho}$ is the scaling parameter and relies on the range ρ, namely the distance at which spatial correlations become small. On the other hand, Γ is the gamma function, and K_ν is the modified Bessel function of the second kind.

2.2 Stochastic Partial Differential Equation (SPDE) Approach

Let $\mathbf{s}_i$ be the two-dimensional spatial coordinate of the i-th pollution station, $i = 1, \ldots, 44$. We collect all the PM10 measurements on day t into the vector $\mathbf{z}_t = (z(t,\mathbf{s}_1), \ldots, z(t,\mathbf{s}_{44}))$, and we rewrite (2) as

$$\begin{cases} \mathbf{z}_t = \alpha_0 + \sum_{j=1}^{p} \alpha_j \mathbf{x}_{t,j} + \boldsymbol{\xi}_t + \boldsymbol{\epsilon}_t & \text{with } \boldsymbol{\epsilon}_t \sim N(\mathbf{0}, \sigma^2_\epsilon \mathbf{I}) \\ \boldsymbol{\xi}_t = a\boldsymbol{\xi}_{t-1} + \boldsymbol{\omega}_t & \text{with } \boldsymbol{\omega}_t \sim N(\mathbf{0}, \sigma^2_\omega \boldsymbol{\Sigma}) \end{cases}.$$

Fitting this kind of model and carrying out inference are computationally demanding tasks. In fact, computational costs involved in linear algebra operations are quite high, especially when dealing with large datasets in space and time (Banerjee, Carlin & Gelfand 2014). To overcome this issue, we use the Stochastic Partial Differential Equation (SPDE) approach by Lindgren et al. (2011). Specifically, we can replace the continuous GRF model with a discretely indexed Gaussian Markov Random Field (GMRF) solution to a specific type of SPDE. As its name suggests, a GMRF is a GRF where the conditional dependencies between observations form a graph with a Markovian structure. The sparsity pattern of its precision matrix $\mathbf{Q}$ reflects that structure. Indeed, the zero entries indicate conditional independencies, as opposed to the conditional dependencies associated to the non-zero values.

The objective of the SPDE approach is to find a GMRF that best represents the Matérn field $\boldsymbol{\omega}_t$. This can be done in few steps. Firstly, the domain of the GRF is partitioned into a set of non-intersecting triangles. Such a triangulation is often called mesh and one example for Sweden is showed in Figure 2. Then, for every fixed time t, the GRF is expressed as an approximate solution to a certain SPDE defined on the triangulation. Such an approximate solution can be written in terms of basis functions, namely

$$\omega(t, \mathbf{s}) \approx \tilde{\omega}(t, \mathbf{s}) = \sum_{l=1}^{n} w_{t,l} \psi_l(\mathbf{s}),$$

where n is the total number of mesh vertices, $\{\psi_l(\mathbf{s})\}_{l=1}^n$ are the basis functions, and $\{w_l\}_{l=1}^n$ are Gaussian distributed weights. This finite representation establishes the explicit link between the GRF $\omega(t, \mathbf{s})$ and the GMRF $\mathbf{w}_t = (w_{t,1}, \ldots, w_{t,n})$. In particular, $\mathbf{w}_t \sim N(\mathbf{0}, \mathbf{Q})$, and its sparse precision matrix $\mathbf{Q}$ is such that $\tilde{\omega}(t, \mathbf{s})$ best approximates the Matérn field. By representing the GRF in this way, we can exploit the special structure of $\mathbf{Q}$ and efficiently perform Bayesian inference on the process and its parameters, as well as predictions at new locations. Rigorous derivations of the SPDE approach are provided by Lindgren et al. (2011).

Figure 2: Triangulation of Sweden with pollution stations (orange circles).

Taking everything into account, for each time point t, the GRF $\boldsymbol{\omega}_t$ in (2) is represented through the GMRF $\mathbf{w}_t \sim N(\mathbf{0}, \mathbf{Q})$, which is independent on time t. Consequently, model (2) becomes

$$\begin{cases} \mathbf{z}_t = \alpha_0 + \sum_{j=1}^{p} \alpha_j \mathbf{x}_{t,j} + \mathbf{B}\boldsymbol{\xi}_t + \boldsymbol{\epsilon}_t & \text{with } \boldsymbol{\epsilon}_t \sim N(\mathbf{0}, \sigma_\epsilon^2 \mathbf{I}) \\ \boldsymbol{\xi}_t = a\boldsymbol{\xi}_{t-1} + \mathbf{A}\mathbf{w}_t \end{cases}, \tag{4}$$

where $\mathbf{B}$ is the matrix that selects the value of the GMRF $\boldsymbol{\xi}_t$ for each observation vector $\mathbf{z}_t$, and $\mathbf{A}$ is the respective matrix for the innovations approximation.

How can we actually fit the pollution model (4)?
We need to estimate both the spatial hyperparameters $\boldsymbol{\delta} = (\sigma_\omega^2, a, \kappa, \sigma_\epsilon^2)$, the regression effects $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \ldots, \alpha_p)$, and the latent effect $\boldsymbol{\xi}$. Since (4) belongs to the class of latent Gaussian models, we can use the Integrated Nested Laplace Approximation (INLA) algorithm for Bayesian inference proposed by Rue et al. (2009). This method directly approximates the posterior marginals of the model parameters. Unlike traditional simulation-based techniques like the Markov chain Monte Carlo (MCMC) method, INLA is computationally efficient for hierarchical problems. Furthermore, the model can be easily implemented through the `R` package *R-INLA* (Version 22.05.07).

2.3 The Pointless Spatial Model for Influenza

In the previous parts, we have seen how to define and fit a spatial-temporal model for modeling PM10 levels across the whole country. We now focus on the influenza model. Recall that, differently from above, influenza counts are regional-based data. Here, we assume people living in a certain Swedish region can be tested in that region only. In other words, counts of positive influenza tests are aggregated over partition sets whose irregular boundaries are given by the administrative boundaries. Moreover, we are modelling influenza outcomes week by week, meaning that time is fixed, thus the index t can be omitted.

We would like to make inference at a point-level spatial resolution, even if data is available at another resolution. This scenario is common in many epidemiological and social studies, and it is related to the change of support problem (COSP). Over the years, many studies have been carried out, for instance by Fuentes & Raftery (2005), Diggle et al. (2013), Moraga et al. (2017). In the following, we will refer to the work of Wilson & Wakefield (2018), as they use the SPDE approach to relate a continuous surface to non-normal areal-level data, like our influenza outcomes.

Let R_i denote the i-th region in Sweden and let N_i be the number of people living in R_i, $i = 1, \ldots, 21$. We define $Y_{ij}|p_{ij} \sim Bern(p_{ij})$ as the binary RV of person j living in county R_i and being tested positive to influenza, $j = 1, \ldots, N_i$. Then, $Y_i = \sum_{j=1}^{N_i} Y_{ij}$ is the total number of ill people in county R_i. We can reasonably assume that having influenza is a rare event, hence $Y_i|\mu_i \sim Pois(\mu_i)$, and that $Y_i|\mu_i$ are independent RVs.

We are interested in estimating the risk of influenza μ_i in every region R_i, $i = 1, \ldots, 21$. We have information about the total number of positive tests, Y_i, and not about each person Y_{ij} living at specific points in R_i. Thus, similarly to what we did in Section 2.1, we add a spatial component υ and define

$$\mu_i = \sum_{j=1}^{N_i} e^{\beta_0 + \beta_1 g(Z(\mathbf{s}_{ij})) + \upsilon(\mathbf{s}_{ij})}, \qquad \forall i = 1, \ldots, 21. \tag{5}$$

Here, $\mathbf{s}_{ij} \in \mathbb{R}^2$ is the 2-dimensional coordinate of person j in county R_i, $Z(\mathbf{s}_{ij})$ is the pollution level at that location, and g is a particular function we need to investigate for linking pollution

and influenza properly. For example, g could be mean or the maximum of PM10 concentration in point $\mathbf{s}_{ij}$ during the chosen week. The zero-mean GRF υ has a Matérn covariance function (3) with variance parameter λ, namely $\text{Cov}(\upsilon(\mathbf{s}), \upsilon(\mathbf{s}')) = \lambda^2 \mathcal{C}(||\mathbf{s} - \mathbf{s}'||_2)$. Being difficult to estimate, we set the smoothness parameter $\nu = 1$, whilst we let the scaling parameter κ to be learned. Of course, now the GRF υ does not depend on the time t.

Once again, we follow the SPDE approach by Lindgren et al. (2011). As described in Section 2.2, we approximate the GRF over a new triangulation of the domain by means of a mixture of basis functions:

$$\upsilon(\mathbf{s}) \approx \tilde{\upsilon}(\mathbf{s}) = \sum_{k=1}^{m} u_k \phi_k(\mathbf{s}).$$

Given the GMRF $\mathbf{u} = (u_1, \ldots, u_m) \sim N(\mathbf{0}, \mathbf{Q}_{\boldsymbol{\theta}})$, its precision matrix $\mathbf{Q}_{\boldsymbol{\theta}}$ depends on the spatial hyperparameters $\boldsymbol{\theta} = (\lambda, \kappa)$, and it is such that the resulting distribution for $\tilde{\upsilon}(\mathbf{s})$ best represents the distribution of the solution of the SPDE. As a result, we can implement (5) by approximating the spatial components over the mesh.

Additionally, we incorporate the population information to the model. Let $d(\mathbf{s}_{ik})$ be the population density in the k-th mesh point of region R_i, $\mathbf{s}_{ik}$. This is the relative region population density, namely $d(\mathbf{s}_{ik}) \geq 0$ and $\sum_{k=1}^{m_i} d(\mathbf{s}_{ik}) = 1$, with m_i being the number of mesh vertices in R_i. Taking everything into account, the influenza model is approximated by

$$\mu_i \approx N_i e^{\beta_0} \sum_{k=1}^{m_i} d(\mathbf{s}_{ik}) e^{\beta_1 g(Z(\mathbf{s}_{ik})) + u_{ik}}, \qquad \forall i = 1, \ldots, 21. \tag{6}$$

Differently from the pollution case, we cannot use the INLA algorithm to fit (6). Indeed, because of the model structure, we are actually dealing with a nonlinear random effect model. As an alternative, we resort to empirical Bayes (EB) with Laplace approximations. We implement this approach with the `R` package `TMB`, which stands for Template Model Builder, by Kristensen et al. (2016). More precisely, first we specify the likelihood and the random effects of the model. The `TMB` package then computes the Laplace approximation of the marginal likelihood, integrate out the random effects automatically, and maximises the final output. `TMB` is particularly useful for models involving random effects or complex nonlinear functions, like (6), as computations are performed quickly. See Kristensen et al. (2016) for more details.

Given the vector of the spatial hyperparameters $\boldsymbol{\theta}$ and the GMFR $\mathbf{u} = (u_1, \ldots, u_m)$, the EB estimates are

$$\begin{aligned}
\left(\hat{\boldsymbol{\theta}}, \hat{\beta}_0, \hat{\beta}_1\right) &= \underset{\boldsymbol{\theta}, \beta_0, \beta_1}{\text{argmax}}\ p(\boldsymbol{\theta}, \beta_0, \beta_1 | \mathbf{y}) \\
&= \underset{\boldsymbol{\theta}, \beta_0, \beta_1}{\text{argmax}} \int_{\mathbf{u}} p(\boldsymbol{\theta}, \beta_0, \beta_1, \mathbf{u} | \mathbf{y}) \text{d}\mathbf{u} \\
&= \underset{\boldsymbol{\theta}, \beta_0, \beta_1}{\text{argmax}} \int_{\mathbf{u}} f(\mathbf{y} | \beta_0, \beta_1, \mathbf{u}) p(\mathbf{u} | \boldsymbol{\theta}) p(\boldsymbol{\theta}) p(\beta_0) p(\beta_1) \text{d}\mathbf{u}.
\end{aligned}$$

Recall that $Y_i | \mu_i \sim Pois(\mu_i)$ are independent RVs, therefore $f(\mathbf{y} | \beta_0, \beta_1, \mathbf{u})$ is the product of the Poisson density functions with parameter μ_i defined as (6). The marginal $p(\mathbf{u} | \boldsymbol{\theta})$

corresponds to the GMRF, whose precision matrix $\mathbf{Q}_{\boldsymbol{\theta}}$ depends on $\boldsymbol{\theta} = (\lambda, \kappa)$. As we do not have any prior knowledge on the parameters, we set uniform priors $p(\boldsymbol{\theta})$, $p(\beta_0)$, and $p(\beta_1)$. The `TMB` and Laplace approximations enable us to integrate out the spatial random effect $\mathbf{u}$ and to find the posterior densities $p(\boldsymbol{\theta}, \beta_0, \beta_1|\mathbf{y})$ and $p(\boldsymbol{\theta}|\mathbf{y})$. Thus, the EB estimates $\hat{\boldsymbol{\theta}}, \hat{\beta}_0, \hat{\beta}_1$ are obtained with a numerical optimisation of the posteriors. Since the EB estimates are also MAPs, which are invariant to transformations, we can easily recover the estimates of e^{β_0}, e^{β_1}, and of the functions of $\boldsymbol{\theta}$.

3 Results

3.1 Exploratory Analysis

Before implementing the pollution model (4) and the influenza model (6), described in the previous section, we conduct a quick exploration of the influenza data. Figure 3 depicts the weekly number of positive test-confirmed cases during the 2017-2018 influenza season across all 21 Swedish regions. Stockholm, Västra Götaland and Skåne regions have significantly higher figures for influenza, and most regions report less than 100 weekly cases. Overall, we can observe a common trend in the number of influenza cases across all regions. Specifically, it starts to increase at the beginning of the year, reaching its peak in late February, and it decreases afterwards. Note also the number of positive tests is very low during the first part of the considered period.

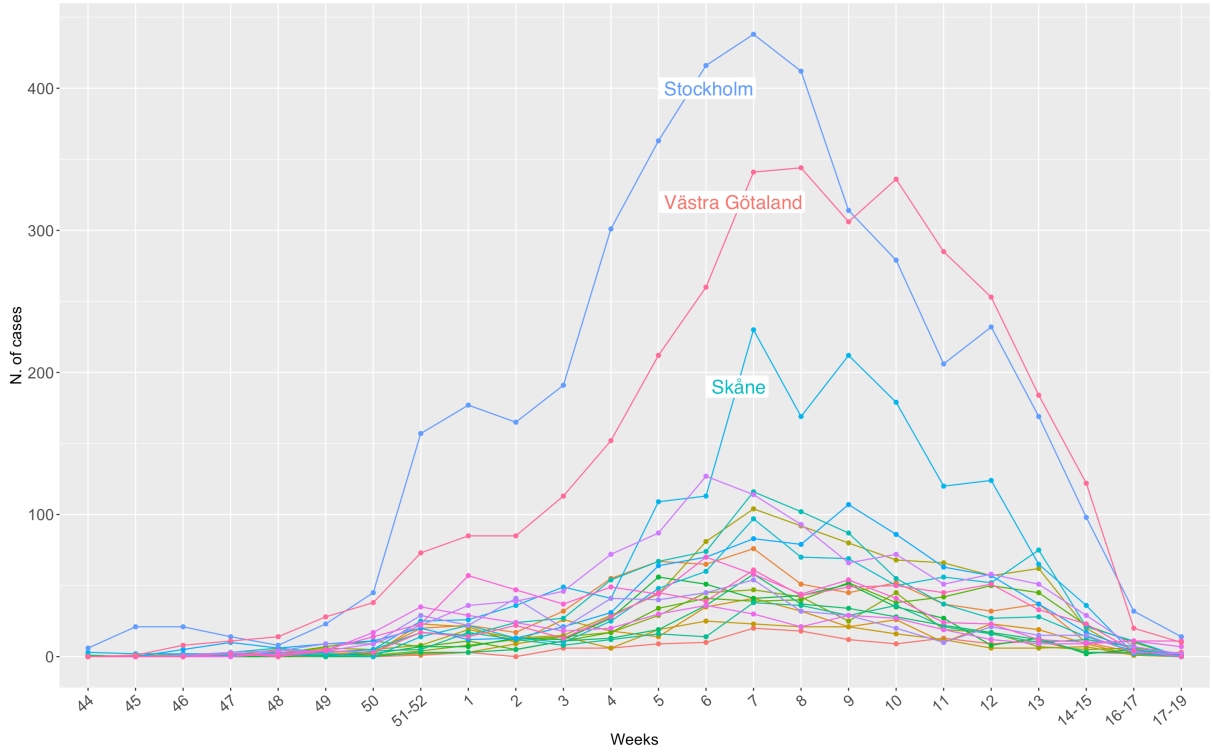

Figure 3: Weekly counts of positive influenza tests per region during season 2017-2018.

Notice the influenza cases are grouped during weeks 51-52, 14-15, 16-17, and 18-19. As mentioned in Section 2.3, our influenza model does not caputure the time dynamics, so (6) will not be implemented it during those weeks.

3.2 PM10 Model

As described in Section 2.1, the pollution model (2) utilises altitude and several weather variables as covariates. To enhance the accuracy and the stability of statistical analyses and to stabilise the variances, we both rescale the predictors and we logarithmically transform the PM10 values. As regards the temporal dynamics, we fit a first-order autoregressive GMRF using the whole 224-day duration of the 2017-2018 influenza season.

As a result, the pollution model performs adequately, although there is substantial noise in the approximations. A representative example of the predicted logPM10 concentrations during week 7 is displayed in Figure 4. As expected, the southern regions exhibit higher pollution levels than the northern areas. The former are indeed the most populated areas and with most factories. Notice also that the area around Bredkälen has notably lower pollution levels; however, the standard deviation at this location is relatively high. This behaviour suggests that the covariates are not predicting the pollution concentration well enough in that location, as the GMRF needs to have higher variability to match the true data. Furthermore, we observe extreme estimations of logPM10 values just below that monitoring station. Again, the standard deviation associated with that area is very high, leading to unreliable approximations. A similar pattern occurs along the northwestern border. Hence, focusing solely on the South may yield more precise approximations as the data are more circumscribed.

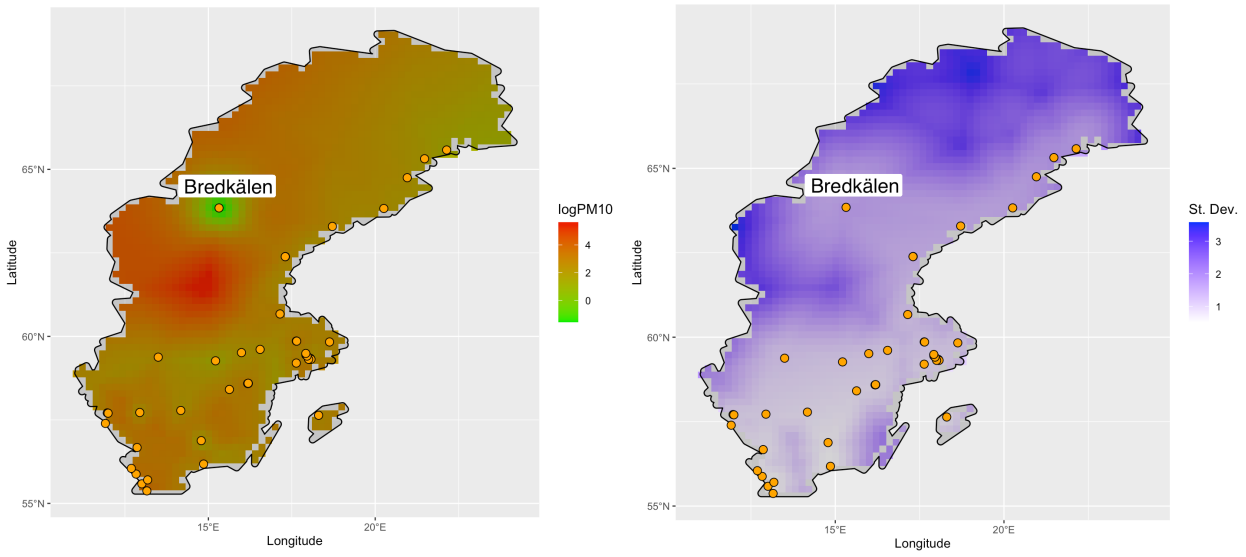

Figure 4: Left: predicted logPM10 with pollution stations (orange circles). Right: standard deviation of the predicted values with pollution stations (orange circles).

3.3 Influenza Model

We start by fitting the influenza model (6) in a simplified setting, specifically with uniform population density and with no pollution estimations. Figure 5 shows the estimates $e^{\hat{\beta}_0}$ throughout the season. Being the common factor to every region and representing the proportion of positive tests, $e^{\hat{\beta}_0}$ should follow a curve similar to the overall trend in Figure 3. Except for week 44, the estimation of e^{β_0} is 0 during the early part of the influenza

season, with a small peak at week 48. Then, the ∩-shape trend only occurs from week 3 to week 9, which corresponds to the period when the highest number of influenza cases have been registered during the season. After week 9, the estimate of e^{β_0} increases again, contradicting the decreasing behaviour of influenza cases. A reasonable explanation is the unstabity of the model, particularly when there are few influenza cases. This motivation is further enhanced when we fit the model with other conditions, such as using the true population density or assigning a normal prior on β_0, as estimations generally do not share the pattern. Nevertheless, we may have some confidence in the model during the weeks when the most influenza cases are recorded, as there seems to be sufficient data to make the basic model work during those weeks.

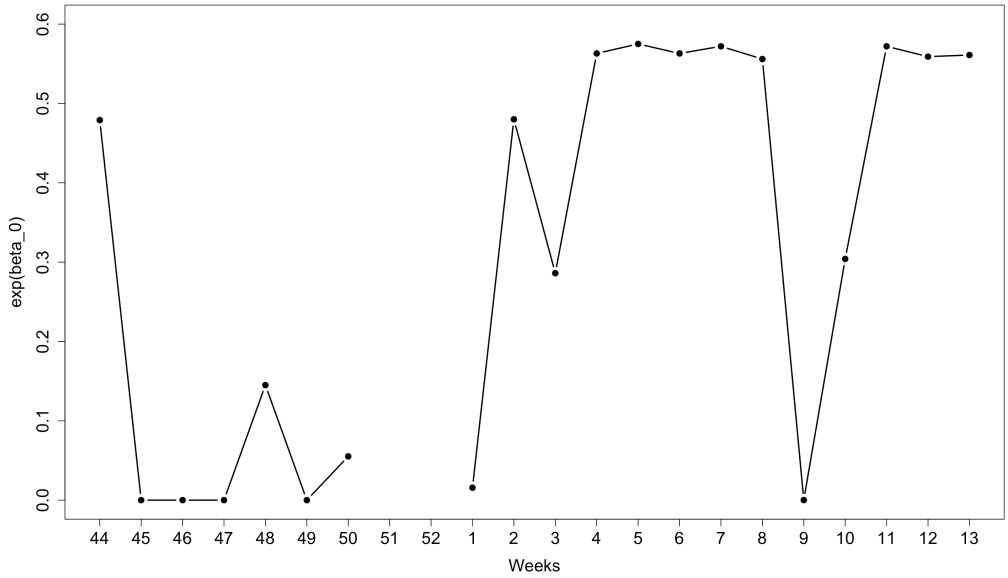

Figure 5: Estimation of e^{β_0} when (6) has uniform population density and no PM10.

Let us now focus on the influenza model as defined in Section 2.3. Recall g is a specific function that links the air pollution estimates to the number of influenza cases. Here, we investigate (6) when g is either the mean or the maximum of the logarithm of PM10 predictions during the specified week. We are interested in the estimations of β_1, as it is the coefficient linking pollution to the probability of being tested positive to influenza. Indeed,

$$\begin{aligned}\frac{\mathbb{P}(\text{person } i \text{ at location } j \text{ has influenza})}{\mathbb{P}(\text{person } i' \text{ at location } j' \text{ has influenza})} &= \frac{e^{\beta_0+\beta_1 g(Z(\mathbf{s}_{ij}))+\upsilon(\mathbf{s}_{ij})}}{e^{\beta_0+\beta_1 g(Z(\mathbf{s}_{i'j'}))+\upsilon(\mathbf{s}_{i'j'})}} \\ &= e^{\beta_1[g(Z(\mathbf{s}_{ij}))-g(Z(\mathbf{s}_{i'j'}))]} e^{\upsilon(\mathbf{s}_{ij})-\upsilon(\mathbf{s}_{i'j'})} \\ &\approx e^{\beta_1[g(Z(\mathbf{s}_{ij}))-g(Z(\mathbf{s}_{i'j'}))]},\end{aligned}$$

where we assume the two locations $\mathbf{s}_{ij}$ and $\mathbf{s}_{i'j'}$ are close enough to have almost the same spatial random effect, i.e. $\upsilon(\mathbf{s}_{ij}) - \upsilon(\mathbf{s}_{i'j'}) \approx 0$. Then, β_1 quantifies the magnitude of the 'pollution' difference $g(Z(\mathbf{s}_{ij})) - g(Z(\mathbf{s}_{i'j'}))$. For instance, if $g(Z(\mathbf{s}_{ij})) - g(Z(\mathbf{s}_{i'j'})) \approx 1$, we can interpret $e^{\beta_1} \approx 1.2$ as indicating that pollution increases the probability of having influenza by 20%.

We implemented (6) using the true population density of Sweden, as the uniform one would require too intensive computational efforts. The estimated values $e^{\hat{\beta}_1}$ throughout the

season are presented in Figure 6. The horizontal red dotted line separates the positive and negative impact of pollution. It can be observed that the estimates of e^{β_1} are quite different when we use the mean of logPM10 compared to when we use the maximum. Specifically, instability issues are apparent in the initial part of the plot, until the first weeks of January. As mentioned earlier, the results may be more reliable from week 3 to week 9. Within this period, the two estimates are still different, but relatively comparable. In particular, note the same increasing-decreasing trend from week 5 to week 9. However, the opposite behaviour during the final weeks is quite surprising and warrants further investigation into the underlying reasons.

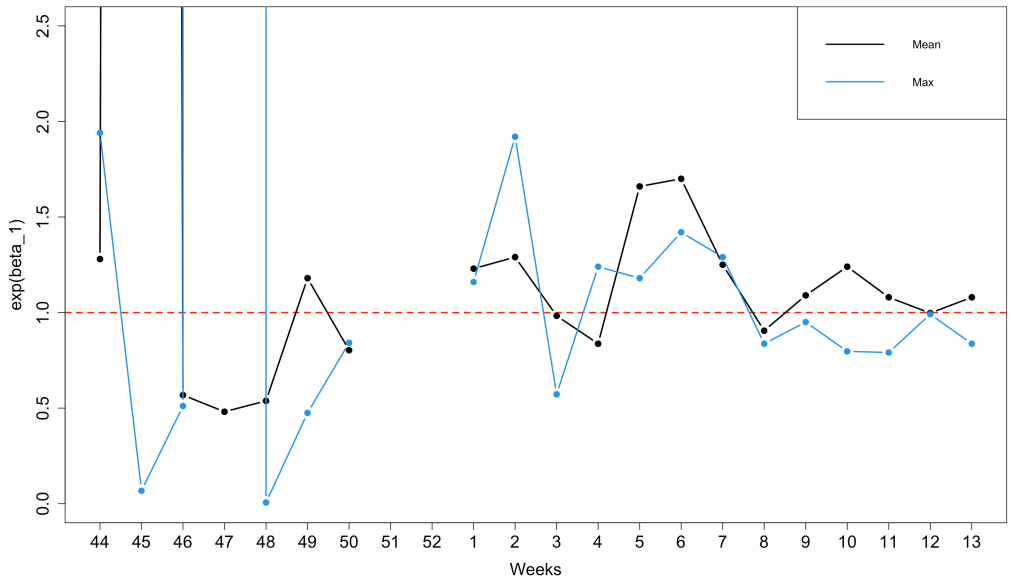

Figure 6: Estimation of e^{β_1}.

Taking everything into account, the theoretical formulation of this work is promising, but its implementation is not working well. Indeed, the influenza model implemented with the `TMB` approach is unstable and cannot provide reliable answers to assess the role of PM10 as a driver of influenza. Remember we chose to estimate our parameters through the `TMB` approach also because of its fast computations. A fully-Bayesian approach should provide better approximations, but implementing it would be unfeasible due to massive computational costs to obtain results. Thus, as proposed in Wilson & Wakefield (2018), a hybrid approach which combines both MCMC and Laplace approximations could be more efficient and effective. Further discussions about the future work will be provided in Section 5.

4 Connections among Strands

The topic of this section is the connection between the different research strands. Our research is primarily centered on the spatial aspect of the problem. Being unable to effectively study the time dynamics of influenza, we have excluded data from the fortnightly reports. However, both Patrick and Veronika have investigated the temporal aspects between pollution and the disease. Hence, we can leverage one of their methods to divide the data for weeks 51-52, 14-15, 16-17, and 18-19.

4.1 Connections with Time Series Analysis (Patrick's Strand)

Patrick conducted a study investigating the temporal correlation between COVID and pollution in several cities in England using a multivariate time series analysis. His study focused solely on temporal aspects while assuming independence of locations and no underlying spatial effect. To connect our two strands, we should apply our models to each other's data, and compare the results obtained from different perspectives. Exploring the time dynamics of the data could significantly improve our outputs. Indeed, this would help us to identify the flaws of our work, especially in the pollution model. Similarly, our spatial knowledge could be incorporated into the multivariate time series structure to define a more realistic setting.

Nevertheless, a more interesting and involved approach would be to develop a single spatio-temporal model by incorporating the theoretical equations defined previously with the time parameter t. For instance, the regional spatio-temporal influenza model could be defined as

$$\mu_i(t) = \sum_{j=1}^{N_i} e^{\beta_0 + \beta_1 g(Z(t,\mathbf{s}_{ij})) + \omega(t,\mathbf{s}_{ij})}, \qquad \forall i = 1, \ldots, 21. \tag{7}$$

This would allow the use of Patrick's estimated parameters to define a suitable model for both pollution and COVID/influenza data, resulting in a more comprehensive understanding of the relationship between the two variables.

4.2 Connections with SIR Model (Veronika's Strand)

Veronika's study focused on modelling the dynamics of the susceptible, infectious, and removed individuals over time using the SIR model. Her work included examining the optimal transmission rate, considering both the seasonality of influenza and pollution. However, such a model did not incorporate any spatial components. To address this limitation, we could provide insight into spatial variation and incorporate it into Veronika's strand. Then, by generating synthetic data based on her improved model, we can test the robustness of our own work.

Furthermore, age is known to have a significant impact on the spread of infectious diseases, yet neither of our models currently account for this. Exploring an age-structured SIR model may be beneficial, as it could provide estimates that we could integrate into our influenza model. For instance, let us define Y_{iaj} to be the binary influenza indicator per person j in age-band a of region R_i. Then, the disease counts are $Y_i = \sum_{a=1}^{A} \sum_{j=1}^{N_a} Y_{iaj}$, where A is the number of age groups and N_{ia} is the number of people in R_i within the stratum a. We define the expected number $E_i = \sum_{a=1}^{A} N_{ia} q_a$. The parameter q_a corresponds to the reference risk for age-band a and should be calculated in advance. If such information is available, then the influenza model becomes

$$\begin{aligned}
\mu_i &= \sum_{a=1}^{A} q_a \sum_{j=1}^{N_{ia}} e^{\beta_0 + \beta_1 g(Z(\mathbf{s}_{ij})) + \upsilon_a(\mathbf{s}_{ij})} \\
&\approx \sum_{a=1}^{A} N_{ia} q_a e^{\beta_0} \sum_{k=1}^{m_i} d_a(\mathbf{s}_{ik}) e^{\beta_1 g(Z(\mathbf{s}_{ik})) + \upsilon_a(\mathbf{s}_{ik})}
\end{aligned}$$

where d_a and υ_a are the population density and the spatial random effect for age group a, respectively.

5 Discussion and Future work

The findings presented in the previous sections demonstrate that ours is a promising framework for assessing the impact of air pollution on influenza transmissibility. However, to provide reliable outputs, further validation and additional modifications to the modelling approaches are required. This conclusive section presents some possible directions for future research.

5.1 Limitations of Current Work

One limitation of our work concerns the availability of the data. Specifically, the number and the locations of pollution stations were insufficient for carrying out a proper statistical analysis. We had data from 44 pollution stations, 36 of them located in the southern regions of Sweden. As a result, our approximations are coarse, and our confidence in the results is limited. Additionally, we have associated each pollution station with the closest available weather station. However, by doing so, some weather values were recorded more than 50 km away from the monitoring station.

Similarly, our study is limited by the availability of influenza data. We only had access to observations from week 44 to week 19, from which we have excluded several weeks. It would be interesting to investigate how pollution affects the transmissibility of influenza throughout the whole year.

5.2 Future Work

5.2.1 Testing the Models

We have not yet validated neither the pollution nor the influenza model. While we can utilise the `R-INLA` functions to effectively test the estimation of PM10, evaluating the influenza model may be more challenging (Plummer 2008). Cross-validation may be employed. For instance, we may remove an entire region and use it as a test set, but this method may be overly stringent. Alternatively, we can fit the model for the entire season, except for a few weeks, predict influenza during those weeks, and then compare the results with actual data.

5.2.2 Selection of Priors

In Section 2.3, we employed uniform priors on β_0, β_1, and $\boldsymbol{\theta}$ to define the EB estimates of the influenza model. Alternatively, informative priors, such as standard normal densities, may provide a more stable optimisation algorithm. Nonetheless, it would be valuable to explore the potential of priors within our framework, as incorporating prior knowledge could enhance the model's performance.

5.2.3 Spatio-Temporal Influenza Model

Currently, we are studying the impact of PM10 on the number of influenza cases through a *plug-in approach*. We begin by fitting the pollution model (4) and estimate the pollutant concentration during a given week. A summary statistic is then calculated for each grid point and included in the influenza model (6), which is fitted separately from the first stage, as a linear predictor. Although this method is computationally efficient, two problems arise.

Firstly, only the pollution model incorporates temporal dependency through an autoregressive latent model. Secondly, it does not take into account the uncertainty in predicting the PM10 concentration, leading to additional noise in (6). As a consequence, computations may be unstable and the resulting influenza risk $\hat{\mu}$ may be biased.

Combining the pollution and influenza models into a unified spatio-temporal model could greatly improve our outcomes. Model (7) presented in Section 4.1 is an example where we include a time-dependent structure in the spatial random field.
Another way is given by a two-stage Bayesian approach (Cameletti et al. 2019). In the first stage, we estimate the PM10 levels at the area level. Then, instead of averaging the point values in each area, we make use of a data fusion approach. Here, a Bayesian melding model allows us to combine pollutant measurements with the output of numerical dispersion models, as described in Moraga et al. (2017). This may improve the pollution estimates as it would lead to lower variability in the predictions. In the second stage, we link the areal PM10 measurements to the influenza outcomes, where we account for both the uncertainty of the pollution estimates and the additional spatial and temporal random effects (Villejo et al. 2023). As highlighted by Villejo et al. (2023), it may still be difficult to estimate the Matèrn field parameters, especially when using non-informative priors. Nevertheless, the two-stage approach could provide more accurate and stable estimates, even with few influenza cases.

5.2.4 Better Use of Data

To obtain estimates of PM10 for the entire region of Sweden, weather covariates should be defined in each point on the grid. As data were not available at the grid level, an inverse distance weighted interpolation was performed. However, more sophisticated techniques and increased data may improve the accuracy of the results.

Finally, the Västra Götaland region has four different laboratories located in distinct areas, highlighted in Figure 7. In the current framework, all observations have been aggregated into a single laboratory. Of course, incorporating additional information and creating an intricate model may improve the estimations. However, dividing the region into four partition sets is not reasonable as individuals may travel and be tested in laboratories outside their jurisdiction. Voronoi tessellation offers a potential solution to this problem. Specifically, a tessellation of either the centroid of each municipality or on uniform points in the Västra Götaland region could be conducted. The latter proposal seems to be more indicated as it does not require any municipality boundaries and allows trying different combinations of points. Then, a weighted

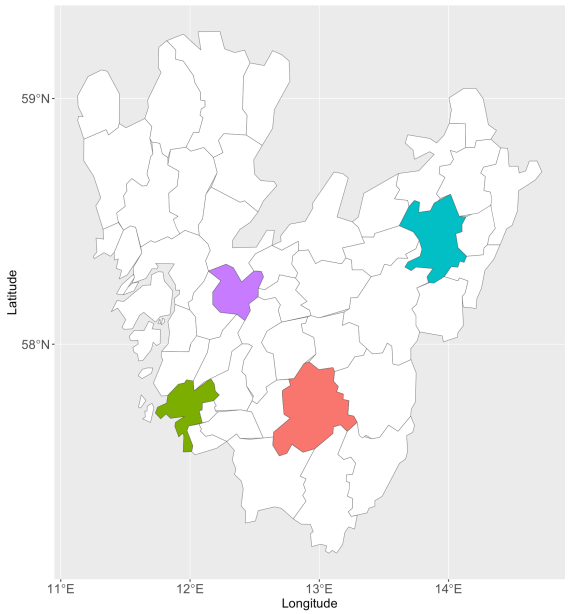

Figure 7: Västra Götaland region.

population density should be used in (6):

$$d'(\mathbf{s}_{ik}) = \frac{d(\mathbf{s}_{ik})}{|\mathbf{l}_i - \mathbf{s}_{ik}|},$$

where $\mathbf{l}_i$ is the location of the laboratory of partition i. We should apply this approach only to the Västra Götaland region, as we have been assuming that individuals can only be tested within their own region. The weighted population density takes into account this assumption, allowing for movement within the region but penalising areas that are farther away.

References

Banerjee, S., Carlin, B. P. & Gelfand, A. E. (2014), *Hierarchical modeling and analysis for spatial data*, CRC press.

Cameletti, M., Gómez-Rubio, V. & Blangiardo, M. (2019), 'Bayesian modelling for spatially misaligned health and air pollution data through the INLA-SPDE approach', *Spatial Statistics* **31**, 100353.

Cameletti, M., Ignaccolo, R. & Bande, S. (2011), 'Comparing spatio-temporal models for particulate matter in Piemonte', *Environmetrics* **22**(8), 985–996.

Cameletti, M., Lindgren, F., Simpson, D. & Rue, H. (2013), 'Spatio-temporal modeling of particulate matter concentration through the SPDE approach', *AStA Advances in Statistical Analysis* **97**(2), 109–131.

Chen, G., Zhang, W., Li, S., Zhang, Y., Williams, G., Huxley, R., Ren, H., Cao, W. & Guo, Y. (2017), 'The impact of ambient fine particles on influenza transmission and the modification effects of temperature in China: a multi-city study', *Environment international* **98**, 82–88.

Cocchi, D., Greco, F. & Trivisano, C. (2007), 'Hierarchical space-time modelling of PM10 pollution', *Atmospheric environment* **41**(3), 532–542.

Diggle, P. J., Moraga, P., Rowlingson, B. & Taylor, B. M. (2013), 'Spatial and spatio-temporal log-Gaussian Cox processes: extending the geostatistical paradigm'.

Domingo, J. L. & Rovira, J. (2020), 'Effects of air pollutants on the transmission and severity of respiratory viral infections', *Environmental research* **187**, 109650.

European Environment Agency (2019), 'Eea discomap: Air quality e-reporting'.
URL: *https://discomap.eea.europa.eu/map/fme/AirQualityExport.htm*

Fuentes, M. & Raftery, A. E. (2005), 'Model evaluation and spatial interpolation by Bayesian combination of observations with outputs from numerical models', *Biometrics* **61**(1), 36–45.

Hijmans, R. J. & University of California, Berkeley (2015), 'National Boundary, Sweden, 2015', TexasGeoDataPortal.
URL: *https://geodata.lib.utexas.edu/catalog/stanford-hd324xr2654*

IQAir (2023), 'Iqair'.
URL: *https://www.iqair.com/*

Kristensen, K., Nielsen, A., Berg, C. W., Skaug, H. & Bell, B. M. (2016), 'TMB: Automatic Differentiation and Laplace Approximation', *Journal of Statistical Software* **70**(5), 1–21.

Lindgren, F., Rue, H. & 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.

Moraga, P., Cramb, S. M., Mengersen, K. L. & Pagano, M. (2017), 'A geostatistical model for combined analysis of point-level and area-level data using INLA and SPDE', *Spatial Statistics* **21**, 27–41.

Plummer, M. (2008), 'Penalized loss functions for Bayesian model comparison', *Biostatistics* **9**(3), 523–539.

R-INLA (Version 22.05.07).
URL: *https://www.r-inla.org*

Rue, H., Martino, S. & Chopin, N. (2009), 'Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations (with discussion)', *Journal of the Royal Statistical Society, Series B* **71**(2), 319–392.

Sahu, S. K. (2012), Hierarchical Bayesian models for space–time air pollution data, *in* 'Handbook of Statistics', Vol. 30, Elsevier, pp. 477–495.

Singer, G., Zivin, J. G., Neidell, M. & Sanders, N. (2020), 'Air pollution increases influenza hospitalizations', *medRxiv* .

Swedish Meteorological and Hydrological Institute (2023), 'Download meteorological observations'.
URL: *https://www.smhi.se/data/meteorologi*

Swedish Public Health Agency (2023), 'Säsongsrapport för influensa'.
URL: *https://www.folkhalsomyndigheten.se/folkhalsorapportering-statistik/statistik-a-o/sjukdomsstatistik/influensa-veckorapporter/arkiv-for-influensa-veckorapporter/sasongsrapport-for-influensa/*

Villejo, S. J., Illian, J. B. & Swallow, B. (2023), 'Data fusion in a two-stage spatio-temporal model using the INLA-SPDE approach', *Spatial Statistics* **54**, 100744.

Wilson, K. & Wakefield, J. (2018), 'Pointless spatial modeling', *Biostatistics* **21**(2), 17–32.

World Health Organization (2018), 'Influenza (seasonal)'.
URL: *https://www.who.int/news-room/fact-sheets/detail/influenza-(seasonal)*

WorldPop & Bondarenko, M. (2020), 'Individual countries 1km UN adjusted population density (2000-2020)'.
URL: *https://hub.worldpop.org/geodata/summary?id=49011*