Spatio-temporal spread of COVID-19 and its associations with socioeconomic, demographic and environmental factors in England

Xueqing Yin[1]* John M. Aiken[2,3] Richard Harris[1] Jonathan L. Bamber[1,4]

[1]School of Geographical Sciences, University of Bristol, Bristol, UK
[2] Expert Analytics, Oslo, Norway
[3] Njord Centre, Departments of Physics and Geosciences, University of Oslo, Oslo, Norway
[4] Department of Aerospace and Geodesy, Technical University of Munich, Munich, Germany

Abstract

Exploring the spatio-temporal variations of COVID-19 transmission and its potential determinants could provide a deeper understanding of the dynamics of disease spread. This study aims to investigate the spatio-temporal spread of COVID-19 infection rate in England, and examine its associations with socioeconomic, demographic and environmental risk factors. Using weekly reported COVID-19 cases from 7 March 2020 to 26 March 2022 at Middle Layer Super Output Area (MSOA) level in mainland England, we developed a Bayesian hierarchical spatio-temporal model to predict the COVID-19 infection rates and investigate the influencing factors. The analysis showed that our model outperformed the ordinary least squares (OLS) and geographically weighted regression (GWR) models in terms of prediction accuracy. The results showed that the spread of COVID-19 infection rates over space and time was heterogeneous. Hotspots of infection rate exhibited inconsistent clustered patterns over time. Among the selected risk factors, the annual household income, unemployment rate, population density, percentage of Caribbean population, percentage of adults aged 45-64 years old, and particulate matter concentrations were found to be positively associated with the COVID-19 infection rate. The findings assist policymakers in developing tailored public health interventions for COVID-19 prevention and control.

Keywords: COVID-19, spatio-temporal patterns, risk factors, Bayesian hierarchical model, England

1 Introduction

Coronavirus Disease 2019 (COVID-19) has caused an enormous challenge to global public health, social security and the economy. The disease originated in Wuhan in China in December 2019, and was declared a pandemic by the World Health Organisation (WHO) in March 2020 [64]. As of February 2023, over 650 million people have been infected by the virus across the world, and the global death toll linked to COVID-19 has been recorded at over 6.7 million (`https://covid19.who.int/`). Although vaccination and public health measures have been successful in reducing the spread of COVID-19 in many areas, gaining a comprehensive understanding of the spatio-temporal transmission patterns and key determinants of the virus remains crucial in the ongoing public health fight against this global health threat. Recent research studies have proposed that COVID-19 transmission and severity may be influenced by various risk factors. Some researchers examined the association of COVID-19 mortality and infections data with socioeconomic status (e.g., deprivation index, income and unemployment), and found that the virus hit harder in socioeconomically disadvantaged communities [1, 9, 16, 58]. Some scholars indicated that environmental conditions such as air pollution, humidity and temperature could influence the virus transmission [4, 14, 41, 63, 70]. In addition, there have been papers on investigating the demographic influencers (e.g., race, gender, age distribution) of COVID-19 transmission [2, 18, 21, 23, 27, 38, 50, 60, 69]. However the available evidence on these determinants to date presents

*Corresponding author: de22002@bristol.ac.uk

a mixed and inconsistent picture. For example, Castr et al.[7] found a strong positive relationship of COVID-19 incidence and mortality with income, whereas Liu et al.[34] showed that higher income groups demonstrated lower incidence rates. A multi-country study carried out by Huang et al. [25] found significant effects of $PM_{2.5}$ on COVID-19 incidence in the USA but not in Italy or Canada. Furthermore, few studies have accounted for socioeconomic, demographic and environmental factors simultaneously to explain the spatio-temporal variations of COVID-19 infection. Therefore, this paper adds to the global evidence base on the topic of identifying influencing factors of COVID-19 spread, by presenting a new study that investigates the spatio-temporal spread of COVID-19 infection rates and its associations with socioeconomic, demographic and environmental factors in England. Our study is based on aggregated count data summarising COVID-19 infected cases between 7 March 2020 and 26 March 2022 at Middle Layer Super Output Area (MSOA) scale in England and by week. The data are modelled using a Bayesian hierarchical model, where the spatio-temporal variation in COVID-19 infection rate is explained by a set of risk factors of interest, important confounding factors, and random effects that account for the spatio-temporal autocorrelation in the data.

To the best of our knowledge, this study represents the first attempt to incorporate socioeconomic, demographic and environmental factors simultaneously to analyse the spread of COVID-19 infection rates at MSOA level across England. Furthermore, it is one of the most up-to-date and comprehensive studies in terms of its temporal duration. For example, Sun et al. [58] used data up to May 2020, Sartorius et al. [56] used data up to 23 August 2020, while the data used by Harris & Brunsdon [23] goes up to 21 May 2021. In addition, compared to previous studies that mostly used frequentist approaches, such as the geographically weighted regression [7, 71, 73], spatial error model [27, 39, 55] and spatial lag model [26, 29, 35], to identify the drivers of the spread of COVID-19, the Bayesian hierarchical model developed here has several advantages. Firstly, it allows for the incorporation of prior knowledge and a more robust assessment of the uncertainties in estimates by specifying prior distributions through a hierarchical modelling scheme [12, 19, 68]. Secondly, the model accounts for spatio-temporal autocorrelation in the data by modelling the random effects via a continuous Gaussian random field process. Thus it is able to account for unreported or missing case data by borrowing information from nearby locations in space and/or time based on the modelled spatio-temporal dependence structure. This enables predictions of COVID-19 infection rates in any geographical location and at any time point, even if the infection cases have not been recorded. This is important from a public health perspective, because COVID-19 is often asymptomatic and under-reported globally, especially during the early phase of the pandemic when there was a lack of resources to facilitate large-scale testing [23, 30]. Therefore, the ability of the model to predict the infection rate even in areas where data are incomplete could provide decision-makers with more comprehensive and accurate information to guide interventions and resource allocation in the fight against the pandemic. This study quantifies the spatio-temporal patterns and disparities in COVID-19 infection rates across England, while also identifying hotspot areas and providing additional evidence on the influencing factors of COVID-19 spread. The Bayesian model utilised in this analysis has demonstrated its effectiveness in accurately predicting COVID-19 infection rates, by exhibiting better modelling accuracy than the ordinary least squares (OLS) and geographically weighted regression (GWR) models. The obtained findings provide valuable insights for policymakers to optimise healthcare resources, establish targeted public health interventions, and improve epidemic prevention and control systems.

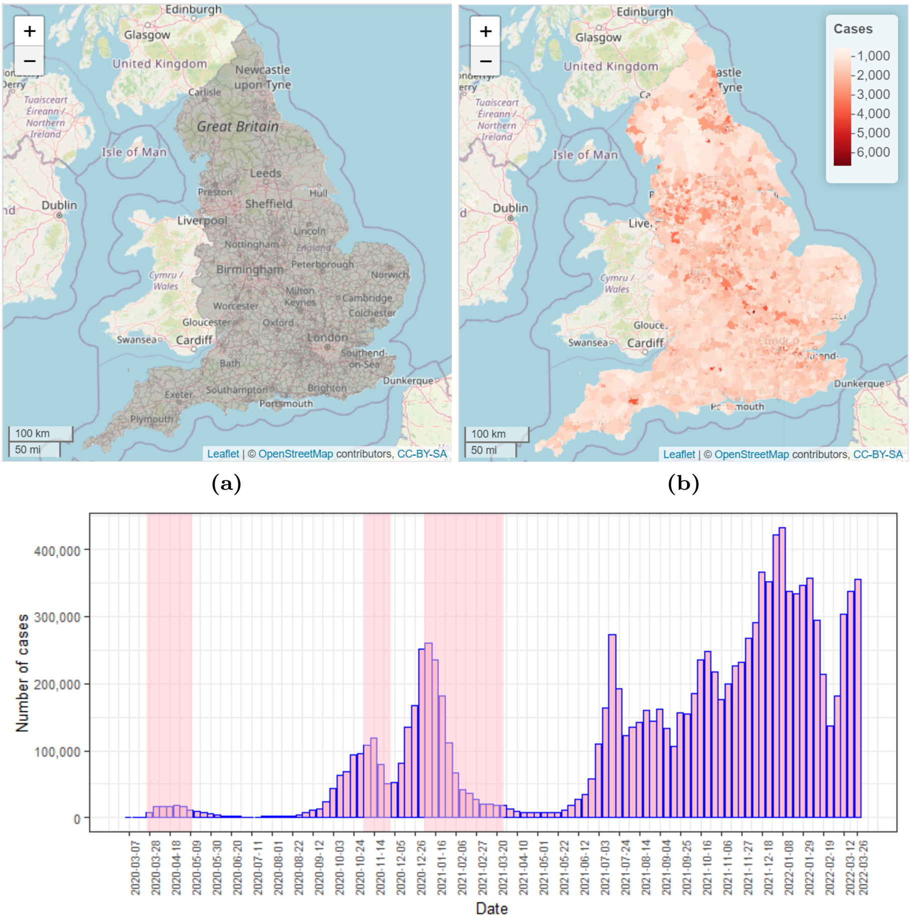
2 Materials and methods

2.1 Ethics statement

No specific permissions were required for the described study, because we did not carry out any experiments. All the work in this study was developed using published data from the official UK Government COVID-19 dashboard, Office for National Statistics and Department for the Environment, Food and Rural Affairs.

2.2 Study region and data

The study region, as shown in Fig 1(a), is mainland England, which is partitioned into $n = 6789$ small neighbourhood units called Middle Layer Super Output Areas (MSOAs). MSOAs are a statistical geography which are designed by English Government for reporting small-area statistics. They represent a formal, administrative specification of neighbourhoods and are developed with specific criteria, including broadly equal population size, socio-economic similarity based on accommodation type and tenure, and spatial compactness of the zones [10]. The shapefiles for MSOAs were obtained from the open geoportal platform (`https://geoportal.statistics.gov.uk/`) provided by the UK government. The median geographical size of an MSOA in England is 3.04 km^2, and as of the end of June 2020, the population size in a single MSOA ranges between 4843 and 27911 persons, with a median population size of 8123.



(a) (b)

(c)

Fig 1. Maps of the study region (a) and cumulative number of COVID-19 infection cases (b) at English MSOA level during the study period. (c): Plot of the reported number of COVID-19 cases by week for all MSOAs in mainland England between 7 March, 2020 and 26 March, 2022. Weeks of lockdown are highlighted in pink.

The COVID-19 infection data used for this study were obtained from the official UK Government COVID-19 dashboard (`https://coronavirus.data.gov.uk/details/download`) on the weekly number of reported COVID-19 cases for each MSOA. The infections data tally, by date reported, the number of individuals who had at least one positive COVID-19 test result, over that and the preceding six days. The observed time frame of the study spanned $T = 108$ weeks from the seven days leading up to 7 March 2020 (week 1) to the seven days to 26 March 2022 (week 108). The data only include pillar 1 cases until 2 July 2020, from when pillar 2 cases are also included. Pillar 1 cases come from the swab testing carried out in Public Health England lab and NHS hospitals for those with a clinical need, and health and care workers, while pillar 2 cases are also swab tests but done for the wider population. These tests are done at local and regional test sites, mobile testing units, satellite test centres and via home tests (https://www.gov.uk/government/publications/nhs-test-and-trace-statistics-england-methodology/). A total of 12 097 525 cases (MSOA and week combinations) were reported during the entire study period, whose values ranged from 3 to 978 cases in a single MSOA and week with a median of 28 cases. Note that the positive cases per MSOA range from 3 upwards is because of censoring of smaller numbers (between 0 and 2 cases) in the reported data. Fig 1(b) displays the spatial pattern of the cumulative COVID-19 infection cases over all weeks at English MSOA level, with higher numbers of infected cases in and around major metropolitan areas such as Newcastle, Manchester, Liverpool, Birmingham, Leicester, Sheffield, and London. Fig 1(c) presents the weekly number of reported COVID-19 cases for all MSOAs, with the x-axis labeled "date" indicating the start date of each observation week. During the COVID-19 pandemic, the UK government introduced various measures, such as the tiered system, the localise and nationwide lockdowns to limit the spread of COVID-19. Our observation window encompasses three separate periods of national lockdown in England, which were officially announced and took place from 26 March to 12 May 2020, 5 November to 2 December 2020, and 5 January to 28 March 2021, respectively [46]. To aid in visualising these periods, they have been shaded in pink in the figure. The figure shows that the first wave of the COVID-19 outbreak in England started in March 2020 and ended at the end of May 2020, whereas the second wave started at the beginning of September 2020 and then reached an initial peak in mid-November. After that the infection levels declined before drastically rising again in late December of 2020 and peaking in early January 2021. There were signs of decrease in the week starting 9 January 2021 until early June, after which the infections started to rise and then wavered until the end of March 2022, with the highest numbers of confirmed cases ($>$ 400 000 cases) observed in early January 2022 (i.e., the week beginning 8 January 2022). Overall, there was typically a reduction in the number of infection cases after the national lockdowns were in place, particularly during the second and third lockdown periods.

A variety of socioeconomic, demographic and environmental factors were selected as potential explanatory variables for the spread of COVID-19 infections, and all these variables were retrieved or prepared at MSOA level. Based on existing studies and data availability, our socioeconomic variables include annual household income, which was available for 2018 (`https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/earningsandworkinghours`), and unemployment rate, which was measured as the percent of people aged 16 years and over unemployed in the labour force, and was available from the 2021 UK census (`https://www.nomisweb.co.uk/census/2021/bulk`) held by the Office for National Statistics (ONS). The demographic factors of population density, ethnic and age distributions by MSOA were also available from the 2021 UK census. Population density was measured as the number of persons per square kilometer, and it was found to be highly skewed to the right. Thus a log-transformation was applied. Environmental variables include the annual mean concentrations of fine particulate matter ($PM_{2.5}$), coarse particulate matter (PM_{10}) and nitrogen dioxide (NO_2) for 2021, all of which were measured in $\mu g\ m^{-3}$. We used concentrations data for 2021 rather than for 2022 or 2020, because the data for 2022 were not available, while the data for 2020 were artificially lower than normal concentration levels due to the implementation of lockdowns and mobility restrictions across England in 2020, which significantly reduced transport usage and hence pollution concentrations of that year. Department for the Environment, Food and Rural Affairs (DEFRA) offers 1×1 km gridded annual mean $PM_{2.5}$, PM_{10} and NO_2 concentrations (`https://uk-air.defra.gov.uk/data/pcm-data`), which is spatially misaligned with the irregularly shaped MSOAs. Thus we adopted the simple averaging method [22] to obtain the concentrations by MSOA. Specifically, each 1 km^2 gridded concentration has an associated centroid, and the pollution concentration in an MSOA is computed by averaging the grid square concentrations whose centroids are located within the MSOA. If an MSOA does not contain a grid square centroid, it is assigned the pollution concentration from the nearest grid square. Table 1 lists the variables together with their source of data. Finally, we also considered two other confounding factors that may impact COVID-19 ill health, including the number of care home beds per adult population (`https://www.cqc.org.uk/`), and a binary variable indicating whether there is a hospital with emergency facilities or not in an MSOA (`https://www.nhs.uk/`).

Table 1. Summary of the socioeconomic, demographic and environmental variables used in this study.

Theme	Variable	Year	Source
Socioeconomic	Annual household income (in thousand, £)	2018	ONS
	Unemployment rate (%)	2021	ONS
Demographic	Log(Population density)	2021	ONS
	Percent of Chinese (%)	2021	ONS
	Percent of Indian (%)	2021	ONS
	Percent of Pakistani (%)	2021	ONS
	Percent of Bangladeshis (%)	2021	ONS
	Percent of African (%)	2021	ONS
	Percent of Caribbean (%)	2021	ONS
	Percent of white British (%)	2021	ONS
	Percent of age 18-29 (%)	2021	ONS
	Percent of age 30-44 (%)	2021	ONS
	Percent of age 45-64 (%)	2021	ONS
	Percent of 65 years old and over (%)	2021	ONS
Environmental	Annual mean $PM_{2.5}$ (µg m^{-3})	2021	DEFRA
	Annual mean PM_{10} (µg m^{-3})	2021	DEFRA
	Annual mean NO_2 (µg m^{-3})	2021	DEFRA

2.3 Statistical analysis

An infection rate is the probability or risk of an infection in a population, which in epidemiology is defined as the proportion of affected people in a population during a specific time period [52]. In this study, it was calculated as the ratio of the number of reported COVID-19 cases to the total population in each MSOA and week. However, naively using such infection rate as a measure of the variations in COVID-19 transmission ignores the spatio-temporal autocorrelation that characterises the disease dynamics. Moreover, it also ignores the potential effects of risk factors on COVID-19 infections. Therefore, it is necessary to develop a model-based approach that can capture the spatio-temporal variations of the disease spread, separate the variations from random noise and account for the spatio-temporal autocorrelation structure in the data.

2.3.1 Spatial and temporal autocorrelation analysis

The presence of positive spatial autocorrelation in the COVID-19 infections data was evidenced by performing the Moran's I test [45]. Moran's I is widely used to measure the level of spatial autocorrelation between adjacent locations globally. The computed Moran's I value was 0.242 and the associated P value was less than 0.0001, indicating statistically significant spatial correlation in the COVID-19 spread throughout England. To assess the temporal autocorrelation in the data, a Ljung-Box test [36] was conducted. The test showed strong evidence of positive temporal correlation in the COVID-19 infection rates over time, as indicated by a P value of less than 0.0001.

2.3.2 Variable selection

Before modelling the data, we examined the multicollinearity among the variables described in Table 1 as well as the confounding risk factors using a Pearson's correlation analysis, because the presence of multicollinearity among independent variables can cause overfitting and less reliable inferences about the associations between the response and predictor variables. The air quality components PM_{10} and NO_2 were found to be strongly correlated with $PM_{2.5}$, with the Pearson's correlation coefficients of 0.86 and 0.85, respectively. Given that previous studies have identified $PM_{2.5}$ as an important risk factor for COVID-19 infections [4, 14, 31], we selected $PM_{2.5}$ over the other two air pollutants for further analysis. The percent of each MSOA's population who were white British was highly correlated with the percent of African and Indian population, with the correlation coefficients of -0.7 and -0.6 respectively, and hence was not included in the analysis. The percent of the population between 30 and 44 years old was also not considered in the model due to its high correlations with the percent of age groups 45-64, and 65 years and over, with the correlation coefficients of -0.6 and -0.8, respectively. Simultaneously, the variance inflation factor (VIF) was used to verify multicollinearity. The VIF values for all the filtered variables were less than 5 (between 1.02 and 4.89), indicating no serious multicollinearity exists [42]. This means that all the predictor variables in the final model were not highly correlated to each other.

2.3.3 Bayesian spatio-temporal modelling

We employed a Bayesian hierarchical model to predict the COVID-19 infection rates at MSOA level over space and time and identify the associated risk drivers, by utilising a generalised linear mixed model with a combination of the selected risk factors, and random effects which account for any residual spatio-temporal dependence of COVID-19 transmission. Let $Y(\boldsymbol{s}_i, t)$ and $N(\boldsymbol{s}_i, t)$ be the number of reported COVID-19 cases and total population in MSOA $i \in (1, \ldots, n_t)$ during week $t \in (1, \ldots, T)$, respectively. Here $\boldsymbol{s}_i \in R^2$ denotes the geographical location for MSOA i, and n_t represents the number of MSOAs that have reported COVID-19 cases during week t. Let $\theta(\boldsymbol{s}_i, t)$ be the logarithm of the observed infection rate in MSOA i during week t, then it is defined as

$$\theta(\boldsymbol{s}_i, t) = \log\left(\frac{Y(\boldsymbol{s}_i, t)}{N(\boldsymbol{s}_i, t)}\right). \tag{1}$$

We chose to model the logarithm of the infection rates ($\theta(\boldsymbol{s}_i, t)$), because the raw infection rates ($\frac{Y(\boldsymbol{s}_i,t)}{N(\boldsymbol{s}_i,t)}$) were highly skewed to the right. Moreover, researchers [32, 59] have demonstrated that the logarithmic transformation is a superior choice over probit or logit transformations for explaining the spread of COVID-19. The first level of the Bayesian hierarchical model is the Gaussian specification given by

$$\begin{aligned} &\theta(\boldsymbol{s}_i, t) \sim \mathrm{N}(\mu(\boldsymbol{s}_i, t), \sigma_e^2), \quad i = 1, \ldots, n_t; \quad t = 1, \ldots, T, \\ &\mu(\boldsymbol{s}_i, t) = \boldsymbol{x}(\boldsymbol{s}_i)^\top \boldsymbol{\beta} + \xi(\boldsymbol{s}_i, t), \end{aligned} \tag{2}$$

where $\mu(\boldsymbol{s}_i, t)$ is the COVID-19 infection rate in MSOA i and week t on the log scale, and σ_e^2 is the variance of the measurement error. The log-infection rate is modelled by two components, the first of which is the vector of p known covariates $\boldsymbol{x}(\boldsymbol{s}_i) = (1, x_1(\boldsymbol{s}_i), \ldots, x_p(\boldsymbol{s}_i))$ related to location $\boldsymbol{s}_i$, including an intercept term, with regression parameters $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)$. As the temporally varying covariate information is not available, we cannot include this in the model and the regression parameters are assumed unchanged over time. The second component accounting for the variations in $\mu(\boldsymbol{s}_i, t)$ is the spatio-temporal random effect $\xi(\boldsymbol{s}_i, t)$, which represents the realization of a spatio-temporal process for the logarithm of COVID-19 infection rates after covariate adjustment. The spatial correlation is modelled by location specific random effects through a Gaussian random field process, which captures the correlation via a covariance matrix expressed as a function of distance between locations. The temporal correlation is modelled by a first-order autoregressive (AR1) process. Specifically, the second level of the model is given by

$$\xi(\boldsymbol{s}_j, t) = \alpha \times \xi(\boldsymbol{s}_j, t-1) + \omega(\boldsymbol{s}_j, t), \text{ and } \xi(\boldsymbol{s}_j, 1) \sim \mathrm{N}\left(0, \frac{\sigma_\omega^2}{1-\alpha^2}\right), \tag{3}$$

where α is a temporal dependence parameter such that $|\alpha| < 1$. $\boldsymbol{\omega}(\boldsymbol{s}_j, t) = (\omega(\boldsymbol{s}_1, t), \ldots, \omega(\boldsymbol{s}_{n_t}, t))^\top$ is a spatial random effect that is assumed to follow a multivariate Gaussian distribution and have $\boldsymbol{\omega}(\boldsymbol{s}_j, t) \sim \mathrm{N}(\mathbf{0}_{n_t}, \sigma_\omega^2 \boldsymbol{\Sigma}_\omega)$, where $\mathbf{0}_{n_t}$ is a $n_t \times 1$ vector of zeros and σ_ω^2 is the marginal variance of the spatial process. $\boldsymbol{\Sigma}_\omega$ is the $n_t \times n_t$ covariance matrix with elements $(\boldsymbol{\Sigma}_\omega)_{ij} = C(||\boldsymbol{s}_i - \boldsymbol{s}_j||)$, where $||\boldsymbol{s}_i - \boldsymbol{s}_j||$ is the distance between locations $(\boldsymbol{s}_i, \boldsymbol{s}_j)$, and $C(\cdot)$ is the Matern function [67] given by

$$C(||\boldsymbol{s}_i - \boldsymbol{s}_j||) = \frac{1}{2^{\nu-1}\Gamma(\nu)}(\kappa||\boldsymbol{s}_i - \boldsymbol{s}_j||)^\nu K_\nu(\kappa||\boldsymbol{s}_i - \boldsymbol{s}_j||), \tag{4}$$

where $K_\nu(\cdot)$ is the modified Bessel function of second kind, and $\Gamma(\nu)$ is the Gamma function. ν is the smoothness parameter of the Matern covariance function, which is fixed at 1. κ is a scaling parameter controlling the spatial correlation range ρ, which is the distance at which the correlation function has fallen to about 0.13 and is given by $\rho = \sqrt{8\nu}/\kappa$.

The third level of the model specifies the prior specifications for the model parameters. The regression parameters $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)$ were assigned independent weakly informative zero-mean Gaussian prior distributions with a large variance, i.e., $\beta_j \sim \mathrm{N}(0, 1000)$, to ensure their values are mainly informed by data. Penalised complexity priors [17] were specified for the correlation range ρ and the marginal standard deviation parameter σ_ω, with $p(\rho < 1.5) = 0.5$ and $p(\sigma_\omega > 1) = 0.01$, indicating a 0.5 probability of ρ being smaller than 1.5, and a low probability of σ_ω being greater than 1, respectively. The temporal autoregressive parameter α was also assigned a penalised complexity prior, with $p(\alpha > 0) = 0.9$. Finally, a weakly informative log-gamma prior was specified for σ_e^2, i.e., $\ln(\sigma_e^2) \sim \text{log-Gamma}(1, 0.00005)$.

All analyses were conducted using the statistical software R version 4.2.1 [49]. Model inference was implemented using the stochastic partial differential equations (SPDE) method and the Integrated Nested Laplace Approximations (INLA) algorithm via the R-inlabru package (2.5.2) [6, 33, 54]. This approach has significant advantages in terms of computational efficiency and accuracy when handling high-resolution spatio-temporal processes and large datasets. The INLA implementation of the SPDE approach approximates a continuous Gaussian random field process with the Matern covariance function by a discretely indexed spatial random process known as a Gaussian Markov random field (GMRF). The GMRF has zero mean and uses a sparse precision matrix, which thus substantially reduces the computational cost in matrix algebra operations compared to using dense covariance matrices [53]. To represent the Matern field as a GMRF, the SPDE approach discretizes the space by defining a mesh composed of non-intersecting triangles that partition the domain of the study area [33]. These triangles allow the spatial autocorrelation between observations to be calculated in the modelling process. Then the INLA algorithm estimates the posterior distribution of the latent Gaussian process and hyperparameters using the Laplace approximation [54].

Other R packages including spdep (1.2.5) [51], ggplot2 (3.3.6) [65], rgdal (1.5.32) [5], dplyr (1.0.10) [66], RColorBrewer (1.1.3) [48], leaflet (2.1.1) [8], and stats (4.2.1) have also been used for data analysis and visualisation. Model parameter estimates were summarised in posterior mean, standard deviation (SD) and 95% credible intervals. A risk factor is considered to have a statistically significant effect if the 95% credible interval of its estimated regression coefficient does not include zero. The main code to complete this analysis is available from `https://github.com/XueqingYin/COVID19.git`.

3 Results

3.1 Prediction performance

The model outlined above was applied to the COVID-19 infections data. For comparison purpose, the OLS and GWR models were also fitted to the same data to observe which one produces the most accurate predictions of COVID-19 infection rate. The OLS model is the most common and widely used global non-spatial regression model, while the GWR model is a spatial model that allows the relationships between the independent and dependent variables to vary by locality. Given the spatial nature of the GWR model, it was applied separately to the COVID-19 infections data for each time period, without considering the temporal autocorrelation in the data. To quantify the prediction accuracy of each model, statistical measures including root mean square error (RMSE) and mean absolute error (MAE) for the infection rate predictions were computed. Their mathematical equations are as follows:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n_{\mathrm{obs}}} \sum_{t=1}^{T} \sum_{i=1}^{n_t} (\hat{\theta}(\boldsymbol{s}_i, t) - \theta(\boldsymbol{s}_i, t))^2}, \tag{5}$$

$$\mathrm{MAE} = \frac{1}{n_{\mathrm{obs}}} \sum_{t=1}^{T} \sum_{i=1}^{n_t} |\hat{\theta}(\boldsymbol{s}_i, t) - \theta(\boldsymbol{s}_i, t)|, \tag{6}$$

where $\theta(\boldsymbol{s}_i, t)$ and $\hat{\theta}(\boldsymbol{s}_i, t)$ represent the observed and predicted values at a give location $\boldsymbol{s}_i$ and time t, and n_{obs} represents the total number of observations across all time steps.

It can be seen from Table 2 that our Bayesian model performed best in terms of prediction accuracy. The RMSE value for the Bayesian model was 72.5% lower than that of the OLS model and 15.4% lower than that of the GWR model. The MAE value for the Bayesian model outperformed the OLS model by 74.1% and the GWR model by 12.5%.

Table 2. Prediction accuracy of different models for the COVID-19 infection rates.

Model	RMSE	MAE
Bayesian	0.0011	0.0007
OLS	0.0040	0.0027
GWR	0.0013	0.0008

Therefore, the subsequent sections will present the primary results obtained from the proposed Bayesian model, because it is the best fitting model in terms of model accuracy. To explore the transmission of COVID-19 infection rate at MSOA level and understand its relationships with socioeconomic, demographic and environmental risk factors, we will focus on the following research questions of interest.

1. Are there health inequalities in COVID-19 infection rates among MSOAs in mainland England, and how do these inequalities evolve over time?

2. What impacts do the socioeconomic, demographic and environmental factors have on COVID-19 infection rate?

3. Where are the hotspots of COVID-19 infection rates?

4. How does the COVID-19 infection rate change over time in different regions of England?

3.2 Overall health inequalities and temporal evolution of COVID-19 infection rate

Table 3 provides a summary of the estimates for the spatial correlation range parameter ρ and temporal dependence parameter α from the model. It indicates that the spread of COVID-19 infections in England had a positive spatial and temporal autocorrelation, with $\rho = 0.568$ (SD = 0.004, 95% CI = 0.559-0.575) and $\alpha = 0.908$ (SD = 0.003, 95% CI = 0.905-0.913). These estimates lend further support to the choice of a spatio-temporal model as an appropriate framework for modelling the spread of COVID-19 infection rates.

Table 3. Summary of the spatial correlation range parameter ρ and the temporal dependence parameter α.

Parameter	mean	SD	0.025quant	0.975quant
Spatial range ρ	0.568	0.004	0.559	0.575
Temporal dependence α	0.908	0.003	0.905	0.913

Fig 2 shows two panels illustrating the estimated COVID-19 infection rates for all MSOAs by week. To improve the data visibility, we chose to plot the logarithm of the infection rates. The top panel reveals the presence of health inequalities in COVID-19 infection rates in England. There were substantial variations in the estimated infection rates and the extent of these spatial inequalities increased over time, because the interquartile range of the infection rate, which measures the spread of the distribution, widened from 0.0001 in the first observation week (starting on 7 March 2020) to 0.0028 in the last observation week (starting on 26 March 2022), indicating a potentially growing disparity in COVID-19 infection rates across different neighbourhoods over time.

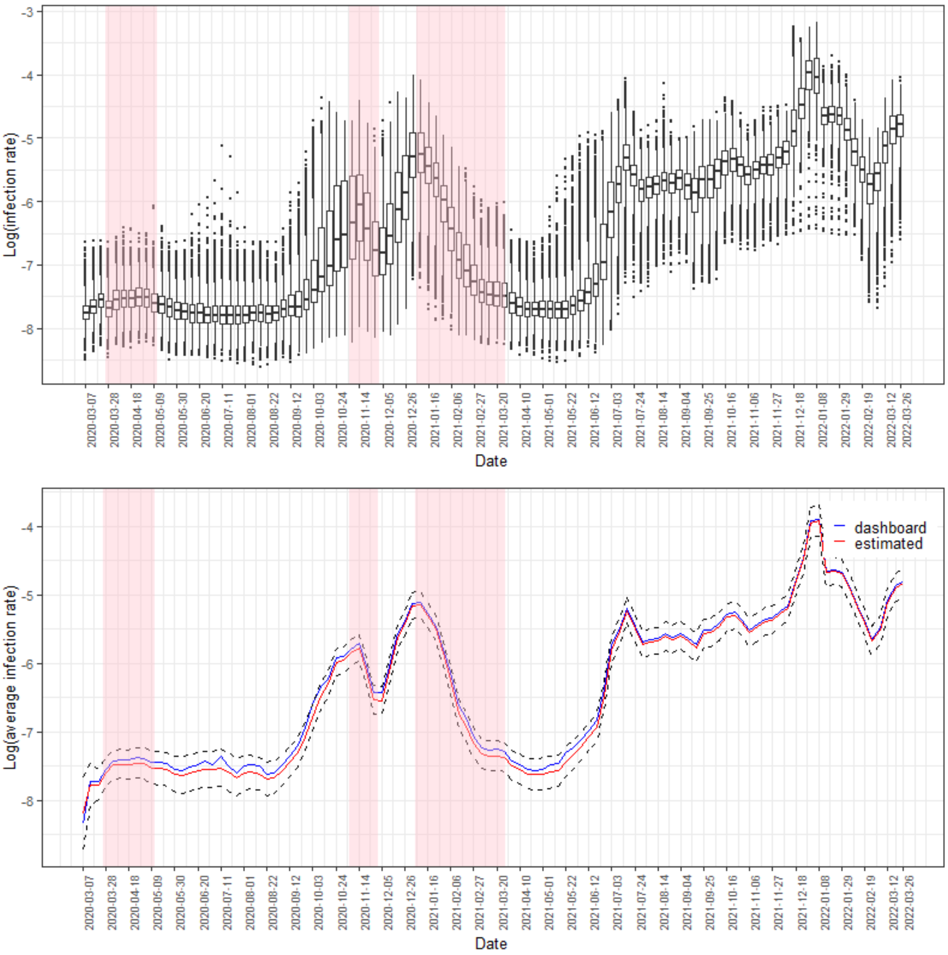


Fig 2. Summary of the estimated COVID-19 infection rates. The top panel displays boxplots of the logarithm of the estimated infection rates across all MSOAs over the 108 weeks, while the bottom panel compares the model's average estimated infection rate on the log scale across mainland England (with 95% credible intervals represented by dashed lines) to the log of the average infection rate computed based on UK COVID-19 dashboard reported cases in weeks. The x-axis label ”date” indicates the starting date of each observation week. Weeks of lockdown are highlighted in pink.

The bottom panel in Fig 2 shows that the average infection rates estimated by our model were highly consistent with the UK COVID-19 dashboard reported data. Over the study period, the COVID-19 infection rate displayed a series of fluctuations, but there was an overall upward temporal trend. The estimated infection rates ranged from 0.0002 to 0.0417, with a mean of 0.0028 (SD = 0.0037). The average infection rate curve shows an initial peak in mid-November 2020, followed by a decline in infection levels, which then surged in December 2020 due to the Alpha (also known as B.1.1.7) variant of coronavirus. Infections peaked in early January 2021, and then tapered off to almost baseline levels by April of the same year. From the week of 5 June 2021 to mid-July, the infection rate began to rise again, reaching an estimated 0.5%. The rate then fell and displayed a relatively stable trend with slight oscillations until November 2021. However, the infection rate reached its highest point in the week of 1 January 2022 at 1.9%, before substantially decreasing until early March 2022. The curve also suggests that the national lockdowns reduced the infection rate, particularly during the second and third lockdown periods, while individuals were at a higher risk of COVID infection during the months of July, December, and January. We also noted that in the first weeks of the study period both the estimated infection rates and their variations across MSOAs were quite low. This is likely because during the first wave of the pandemic between March and July 2020, COVID-19 testing capacity was strictly limited to priority groups due to a lack of infrastructure for large-scale testing, which means that a large number of infected individuals were not formally diagnosed with the virus. As a result, data on confirmed COVID-19 cases were under-reported and incomplete, thus leading to underestimated infection rates and small variation.

3.3 Risk factor effects

Table 4 presents the results of our Bayesian regression analysis, showing the estimated regression coefficients and relative risks (RR) for each of the selected explanatory variables in relation to COVID-19 infection rates. The estimated relative risks and 95% credible intervals were computed by exponentially transforming the regression coefficients associated with the variables. Note that the relative risks relate to realistic increases in each variable, which are given in brackets in column 1 of the table.

Table 4. Estimated regression coefficients, relative infection risks and 95% credible intervals for the effects of each risk factor on COVID-19 infection rates. The relative risks relate to realistic increases in each covariate, which are given in brackets in column 1 of the table.

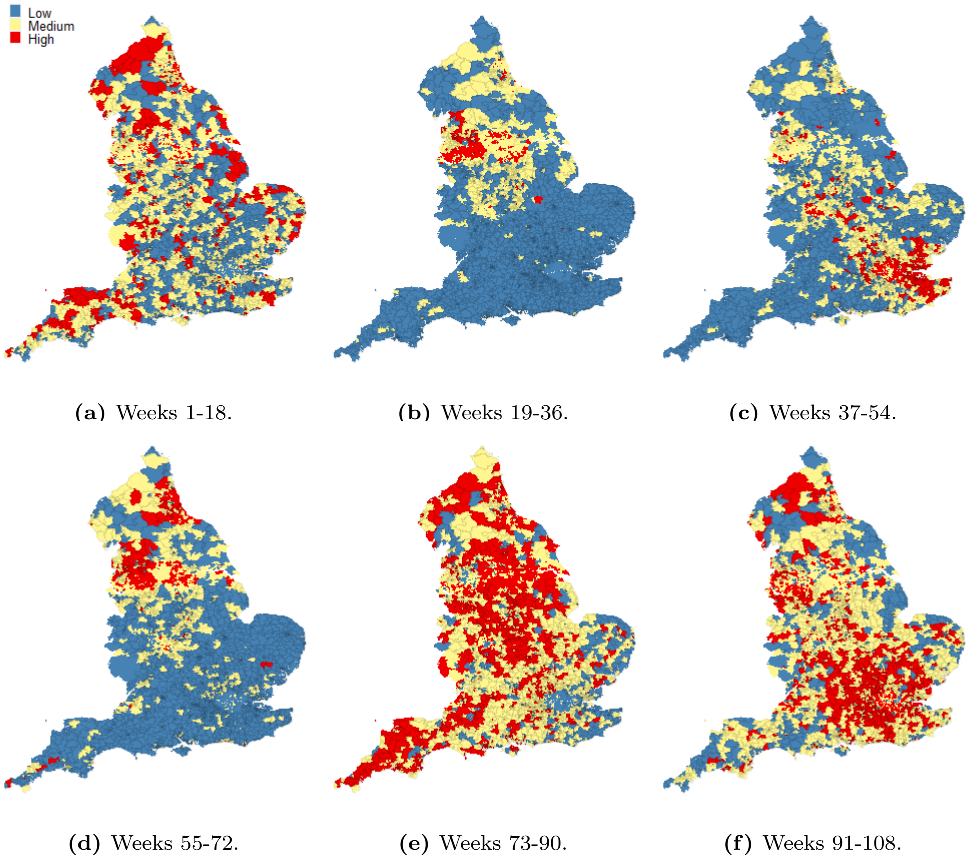
Variable	Regression coefficient	Relative risks (95% CI)
Annual household income (£1000)	0.0008	1.0008 (1.0005, 1.0012)
Unemployment rate (1%)	0.0027	1.0027 (1.0024, 1.0030)
Log(population density)	0.0145	1.0146 (1.0129,1.0164)
Percent of Chinese (1%)	-0.0161	0.9840 (0.9823, 0.9858)
Percent of Indian (1%)	-0.0005	0.9995 (0.9993, 0.9998)
Percent of Pakistani (1%)	-0.0016	0.9984 (0.9982, 0.9990)
Percent of Bangladeshis (1%)	-0.0014	0.9986 (0.9982, 0.9990)
Percent of African (1%)	-0.0093	0.9907 (0.9901, 0.9913)
Percent of Caribbean (1%)	0.0022	1.0022 (1.0009, 1.0036)
Percent of age 18-29 (1%)	-0.0049	0.9951 (0.9947, 0.9954)
Percent of age 45-64 (1%)	0.0031	1.0031 (1.0024, 1.0039)
Percent of 65 years old and over (1%)	-0.0067	0.9933 (0.9929, 0.9937)
Annual mean $PM_{2.5}$ (1 µg m^{-3})	0.0125	1.0126 (1.0083, 1.0167)
Care home beds (0.01)	1.3295	1.0134 (1.0121, 1.0147)
Emergency facilities (TRUE)	0.0007	1.0007 (0.9958, 1.0057)

The table clearly demonstrates that the selected variables significantly contribute to the spatio-temporal variations of COVID-19 infection rate, with the exception of the emergency facilities variable because its 95% credible intervals include the null risk of 1. One of the main drivers of elevated infection rates is socioeconomic factor. We found that both annual household income and unemployment rate were associated with higher infection rates. Specifically, an increase of £1000 in annual household income within an MSOA was associated with a 0.08% increased infection rate (RR = 1.0008, 95% CI = 1.0005-1.0012), while an increase of 1% in unemployment rate was associated with a 0.27% increased infection rate (RR = 1.0027, 95% CI = 1.0024-1.0030). Expectedly, the logarithm of the population density was positively correlated with the infection rate. A 1-unit increase in log(population density) was found to be associated with a 1.5% rise in COVID-19 infection rate (RR = 1.0146, 95% CI = 1.0129-1.0164). We also identified interesting patterns in the associations between ethnicity, age groups, and COVID-19 infection rates. Neighbourhoods with a greater percent of Chinese population tend to have lower rate of infections (RR = 0.9840, 95% CI = 0.9823-0.9858). The Indian, Pakistani and Bangladeshi ethnic groups had a lower level of infections, with rates decreasing by 0.05%, 0.16% and 0.14%, respectively, for every 1% increase in the percentage of these groups within an MSOA. The African population was also associated with a decreased rate of COVID-19 infection (RR = 0.9907, 95% CI = 0.9901-0.9913), whereas the Caribbean population had statistically significantly higher COVID-19 infection rates (RR = 1.0022, 95% CI = 1.0009-1.0036).

Table 4 further reveals that infection rates tend to be lower in MSOAs with higher percentages of adults aged 65 years and older (RR = 0.9933, 95% CI = 0.9929–0.9937), and those with higher percentages of adults aged 18-29 years (RR = 0.9951, 95% CI = 0.9947–0.9954). Conversely, the population aged between 45 and 64 years old was positively correlated with infection rates, with a 1% increase being associated with a 0.31% higher rate. Additionally, there was a positive relationship between $PM_{2.5}$ concentrations and COVID-19 infections, with a 1 µg m^{-3} increase in concentrations associated with between a 0.83% and a 1.67% increased rate (RR = 1.0126, 95% CI = 1.0083-1.0167). Finally, increasing the number of care home beds per adult population (RR = 1.0134, 95% CI = 1.0121-1.0147) was diagnosed positively linked to higher infection rates.

3.4 Hotspots of COVID-19 infection rates

It is of interest to identify which MSOAs showed the high infection rates during the study period, and if there were any areas that consistently exhibited high rates. Clustering methods can be used to identify the groups of MSOAs with high and low rates, and in this study, k-means clustering [24] was utilised. For ease of data visualization, we divided the 108-week time frame into six consecutive and non-overlapping intervals, each spanning 18 weeks. Since k-means clustering uses distance-based measurements to determine the similarity between data observations, we standardised the estimated infections rates during weeks 1 to 18, 19 to 36, 37 to 54, 55 to 72, 73 to 90 and 91 to 108, respectively, and then applied k-means clustering to the standardised rates. The MSOAs were assigned to between one and ten clusters based on their similarities in the estimated rates, and the optimal number of clusters was determined by using the elbow method. Here the optimal number of clusters is 3 clusters for each time interval, and hence the MSOAs were classified into three distinct clusters that have a high, medium, and low level of infection rate, respectively. Fig 3 displays the spatial patterns of the cluster memberships of MSOAs for each time interval.



(a) Weeks 1-18. **(b)** Weeks 19-36. **(c)** Weeks 37-54.

(d) Weeks 55-72. **(e)** Weeks 73-90. **(f)** Weeks 91-108.

Fig 3. Maps showing the clusters that were formed according to the estimated infection rates at MSOA level during weeks 1 to 18 (7 March 2020-4 July 2020), weeks 19 to 36 (11 July 2020-7 Nov 2020), weeks 37 to 54 (14 Nov 2020-13 March 2021), weeks 55 to 72 (20 March 2021-17 July 2021), weeks 73 to 90 (24 July 2021-20 Nov 2021) and weeks 91 to 108 (27 Nov 2021-26 March 2022), respectively.

Fig 3 illustrates that the distribution of high infection rates was not consistent across different time periods and exhibited clustering patterns, indicating that certain areas were more significantly impacted by the pandemic than others. During weeks 1 to 18 (7 March 2020-4 July 2020), the highest infection rates were mainly concentrated in northern England, such as districts of Cumbria, Lancashire, Northumberland and Newcastle (County Durham, Gateshead, Sunderland), and in the far southwest, such as Devon. From week 19 to week 36 (11 July 2020-7 Nov 2020), hotspots for COVID-19 infection rates were highly clustered in a contiguous aggregation of MSOAs spanning from Lancashire, Manchester and Liverpool through to Bradford, Leeds, Kirklees and Sheffield. However, this trend moved to the southeast, particularly in London and some surrounding areas such as Kent and Essex in the following 18 weeks. During weeks 55 to 72 (20 March 2021-17 July 2021), the clusters of MSOAs with high infection rates were mainly in and around metropolitan areas for example Manchester, Liverpool, Newcastle, and North Tyneside. Interestingly, weeks 73 to 90 (24 July 2021-20 Nov 2021) showed a more dispersed distribution of hotspots, with hotspots scattered throughout England and no clear clustering patterns observed. However, there was a higher number of hotspots in central England and surrounds (Nottingham, Birmingham, Leicester, Stoke, Coventry, Leeds, Sheffield, Doncaster and Hull), as well as in the southwest (Bath, Bristol, Plymouth and Exeter). Finally, in the weeks 91 to 108 (27 Nov 2021-26 March 2022), high infection rates shifted towards the southeast of England, with the regions of Greater London, South East and East of England most affected. Despite the varying spatial patterns over time, the hotspots tended to cluster in urban and populous areas, particularly in the northwest, central, and southeast. The shifting patterns of infection rates demonstrate the dynamic nature of the pandemic, and the importance of monitoring temporal trends in different regions. Thus in the next section we investigated the temporal trends of COVID-19 infection rates in diverse regions of England.

3.5 COVID-19 infection rates by region

Differences in COVID-19 infection rate trends can be explored at a larger geographical scale by examining regional data, which provides a more aggregated view of the results. The 6789 MSOAs are nested exactly within nine regions of England, comprising North East, North West, Yorkshire and The Humber, East Midlands, West Midlands, East of England, London, South East and South West. Fig 4 provides a geographical representation of these nine regions. To characterise the evolution of infection rates over time in each region, we computed the weighted averages of estimated infection rates for region r and week t as $\hat{\phi}_{rt} = \frac{\sum_{i \in r} p_i \exp(\hat{\theta}(\boldsymbol{s}_i, t))}{p_r}$, where $\hat{\theta}(\boldsymbol{s}_i, t)$ denotes the estimated log-infection rate in MSOA i and week t, p_i and p_r indicate the population sizes of MSOA i and region r respectively, and $i \in r$ means that MSOA i is geographically located within region r. We used population size as the weighting factor in the computation due to the assumption that neighbourhoods with larger populations are likely to make a greater contribution to the overall infection level of a region compared to those with smaller populations. Fig 5 depicts the line plots of the population-based weighted averages of estimated infection rates after logarithm transformation by week and region. Here we chose to plot the log-infection rates instead of the infection rates themselves, as this provides a more suitable scale that allows for better visualisation of the temporal trend curves.

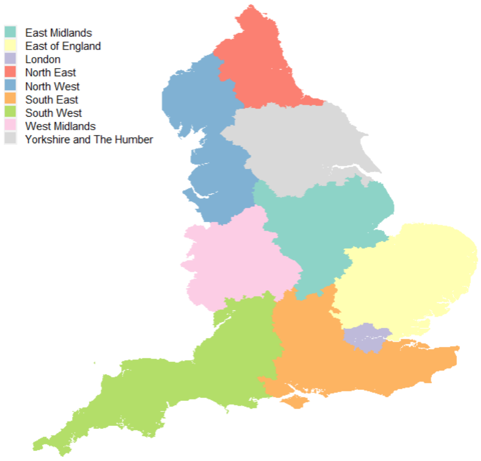


Fig 4. A map of the nine regions of mainland England.

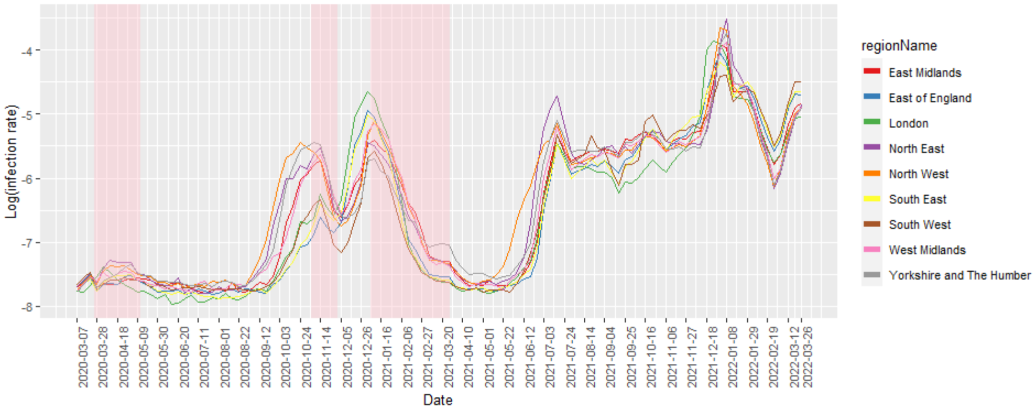


Fig 5. Population-based weighted averages of COVID-19 infection rates on the log scale by week and region in mainland England. Weeks of lockdown are highlighted in pink.

Fig 5 suggests that the national lockdown measures have been successful in containing the transmission of the virus, because in all regions the infection rates visibly declined following the initiation of any of the three lockdowns. During the first lockdown period, the Greater London region had the quickest decline in infection rate, while infection rates in other regions, particularly East Midlands, took longer to reduce. After the lockdown was lifted, infection levels remained stable in most regions until August 2020, but increased notably from September 2020 onwards, with peaks observed in mid-November 2020 in all regions except for North West. When the second lockdown was in place, infection rates decreased in all regions, but remained higher in the northern (i.e., North East, North West) and central regions (i.e., Yorkshire and The Humber, East Midlands and West Midlands) of England. After the lockdown was lifted, the rates surged particularly in London, South East, and East of England due to the dominant Alpha variant, which was estimated to be more transmissible than preexisting variants [13].

During the third lockdown, infection rates across England reduced substantially, with the Greater London region showing the most rapid decline as was the case in the first lockdown. A possible reason for this is that the percentage of key workers in London is relatively lower than that in other regions [3]. Key workers, such as healthcare workers and public transportation workers, are more likely to contract the virus as they continue to work and interact with the public during the imposition of restrictions. With fewer key workers in London, the lockdown measures may have been more effective and noticeable in reducing virus transmission compared to regions with larger numbers of key workers. There were no lockdown restrictions imposed from April 2021 to the end of the study, during which all regions experienced two infection peaks. One peak occurred in mid-July 2021, with rates ranging between 0.42% and 0.89%, while the other was in early Jan 2022, with rates ranging between 1.21% and 2.61%. Infection rates remained relatively stable between the peaks, with London having the lowest values. Infection levels sharply reduced from 8 Jan 2022, but rebounded by the end of Feb 2022. South West, South East, and East of England had the highest infection rates throughout the remainder of the study.

4 Discussion

This study aimed to explore the spatio-temporal spread of COVID-19 infection rate in mainland England from 7 March 2020 to 26 March 2022, and its associations with socioeconomic, demographic and environmental factors. It can provide information for public health policies at the neighbourhood level, offering valuable insights for policymakers to optimise healthcare resources, establish targeted public health interventions, and improve epidemic prevention and control systems. Additionally, the study contributes new cases and knowledge to the growing body of research on the space-time transmission of COVID-19.

A Bayesian hierarchical spatio-temporal model was used to predict the COVID-19 infection rates based on weekly reported COVID-19 cases from 7 March 2020 to 26 March 2022 at MSOA level. Note that we utilised the COVID-19 data at MSOA level rather than the other geographical levels such as the Lower Layer Super Output Area (LSOA) level due to data availability. At the time of the analysis, the MSOA level was the smallest geographic unit for which both COVID-19 data and risk factors data were accessible from open sources. We did not downscale the MSOA geographical information to a higher-resolution geographical level, because the downscaling procedure will likely induce bias and inflate the reported credible intervals for the predicted infection rate [28]. Moreover, the MSOA level is commonly used in the literature for studying the small-areas COVID-19 transmissions [56]. To evaluate the model's prediction performance, our Bayesian model was compared with the non-spatial regression model (OLS model) and the geographically weighted regression model (GWR model). The RMSE and MAE values of the Bayesian approach were lower than those of OLS and GWR models, indicating that the Bayesian model exhibited better prediction accuracy than the other models in this study.

The model estimation results showed that the level of spatial inequalities of COVID-19 infection rate increased over time, highlighting the need for effective strategies to address the disparities between different neighbourhoods. The COVID-19 infection rates in England exhibited spatio-temporal heterogeneity, with higher rates observed during the months of July, December and January. These findings could be explained by the increased travel, outdoor activities, and social gatherings during the summer months, and holiday celebrations and family gatherings during the winter months. These factors could contribute to higher levels of interpersonal contact within a population, potentially leading to increased transmission risk. Besides, COVID-19 virus may be more transmissible in colder and drier conditions [41, 62], contributing to higher infection rates during the winter months. The infection rates were not evenly distributed across the country, with certain areas more vulnerable to the pandemic. Furthermore, the hotspots of infection rates exhibited clustered patterns that changed over time, with a higher frequency of occurrence in and around urban areas such as Newcastle, Manchester, Birmingham, Liverpool, Nottingham, Sheffield, Leeds, and London. The analysis of regional data indicated that although identical national lockdowns were announced across England, different regions displayed varying impacts in response to these measures. This variation could be influenced by factors such as the proportion of essential workers and the emergence of new variants of the virus. Moreover, it is important to note that different regions in England experienced slight variations in the start and end dates of these national lockdowns due to local epidemiological conditions and the tier system. The tier levels, ranging from Tier 1 (medium alert) to Tier 4 (very high alert), determine the extent of restrictions imposed on social interactions, businesses, and public spaces (https://www.ageuk.org.uk/). Regions with higher infection rates and greater risk were typically assigned to higher tier levels, leading to more stringent measures and restrictions. Thus some regions may have entered lockdown earlier or experienced more prolonged periods of restrictions compared to those in lower tiers. Therefore, the observed differences in response to these measures may also be influenced by the region-specific timing and duration of the lockdown measures.

The analysis of socioeconomic, demographic, and environmental factors in relation to COVID-19 infections further indicated the key factors influencing the COVID-19 landscape in England. We found that MSOAs with higher annual household income had a higher infection rate, which is in agreement with previous investigations conducted in other countries or regions [7, 37]. It is assumed that the influence of annual household income on the infections is related to the presence of a better network of health services, expanding the population's access to carrying out diagnostic tests in the communities with high income, reducing underreporting, and contributing to increased COVID-19 incidence. Another socioeconomic indicator is unemployment rate, which was found to be positively associated with COVID-19 infection rate. Consistent findings have been found to exist in France and the United States [20, 72]. A positive relationship was found between the infection rate and population density. We uncovered that the percent of Asians, including Chinese, Indian, Pakistani, and Bangladeshis, was negatively related to COVID-19 infection rate. This corroborates with the study conducted by Lee et al. [31]. Conversely, MSOAs with a higher proportion of Caribbean population had elevated infection rates. These relationships may be related to the ethnic differences in COVID-19 vaccine uptake [40] and health behaviours (https://www.iser.essex.ac.uk/blog/). The age composition of MSOAs was also found to impact COVID-19 infection levels. Higher proportions of adults aged 65 and older were related to lower infection rates. This association could be linked to factors such as higher vaccination rates among older adults, as they were among the first groups prioritised for COVID-19 vaccination in the UK. MSOAs with a higher percentage of young adults aged 18-29 years old showed a lower level of infection, while populations between 45 and 64 years old were positively associated with infection rates. Finally, elevated concentrations of $PM_{2.5}$ were found to be positively associated with COVID-19 infection rates, which is consistent with existing studies that have reported a positive relationship between increasing $PM_{2.5}$ concentrations and COVID-19 infections in various regions such as Ohio, Colorado, and Scotland [4, 31, 43]. This could be related to the increased susceptibility to respiratory infections such as COVID-19 given exposure to air pollution [11, 15]. In addition, air pollution has been linked to underlying health conditions such as diabetes, cardiovascular, and respiratory diseases, which are known risk factors for severe COVID-19 [57]. Therefore, it may be necessary to implement more rigorous COVID-19 prevention measures in areas with higher $PM_{2.5}$.

This study provides evidence that local rates of COVID-19 infections are influenced by patterns of household income, unemployment rate, population density, ethnic composition, age population structure and exposure to air pollution. It provides a scientific basis for accurately predicting COVID-19 response and targeting recovery efforts in England based on community-specific risk factors. It is important to consider these risk factors when developing effective control measures and allocating resources in different communities to control the pandemic spread. For example, public health interventions, such as promoting measures to reduce personal exposure to fine particle pollution could be implemented in specific neighborhoods at higher infection risk, providing better protection for vulnerable populations. Perhaps, most importantly we show that our BHM framework is an effective and powerful tool for modelling and understanding the risk factors that influence the infectious disease dynamics in general. Nevertheless, this study has some limitations. The most significant is its ecological design, where the unit of inference is the group of individuals living in each MSOA rather than having data for each individual. The aggregation of data at the geographic level can cause loss or concealment of certain details about individuals, resulting in ecological fallacy [61] in the observed association. Thus, the estimated population-level associations should not be interpreted as cause-and-effect relationships at the individual level, because they may be influenced by factors such as within-area variation in either the exposure or the confounders. We used this ecological study design because the data required for an individual-level study were not available due to confidentiality reasons. We note, however, that this ecological approach has been predominant in the COVID-19 literature that focuses on exploring the distribution pattern of the pandemic and its influencing risk factors [44, 47, 58]. Our findings outlined above should be treated as indicative associations, rather than conclusive evidence of individual-level causation. To fully understand the causal relationships, further research, including tailored experimental designs combined with primary data collection, is necessary to establish causality definitively and investigate the mechanisms and biological pathways behind the observed associations. Secondly, since the annual household income data used were for 2018, the time gap between income data and other predictors and response data may influence the model estimation results. Thirdly, the reported COVID-19 infection cases were linked to the MSOA where the test was conducted, rather than the patient's place of residence. This is a concern for metropolitan areas where a considerable number of patients could have been admitted or transferred to neighboring healthcare facilities before being tested. Finally, the infections data collected in the early phase of the pandemic were often underestimated due to a lack of testing, which could potentially lead to biased estimates and uncertainty in the model predictions. Thus, the limitations of data quality and potential biases should be carefully considered when interpreting the modelling results for the initial pandemic stage.

We noted that in this study the effect sizes of the risk factors on the COVID-19 infection rate were relatively small. The small effect sizes are likely due to the small geographical level (MSOA) and the short time steps (weekly basis) for reporting the data. MSOAs represent small geographic areas with limited populations, therefore the spatial variations of risk factors and COVID-19 infections between English MSOAs naturally tend to be smaller than those between the larger geographical units such as counties or cities in other studies. Additionally, the use of weekly reporting interval tends to yield smaller infection rate variations compared to longer time steps. However, we believe that even these small effect sizes carry important information and implications for public health strategies. They contribute to our understanding of how specific risk factors influence infection rates at a fine-grained geographical level. By considering these effects, we can gain valuable insights for shaping localised interventions and preventive measures to reduce the spread of COVID-19 at neighbourhood level, and promote community health at this level of granularity. There are several avenues for future work. In this study we assumed fixed effects of risk factors over time, however, we also acknowledged that the associations between these factors and the outcome variable may vary with time, thus there is potential to further enhance the model by incorporating spatially and temporally varying effects of risk factors. In addition, future applications could consider other potentially important predictors for COVID-19 infections such as comorbidities, population mobility and and behavioural factors (e.g., alcohol consumption, smoking). The level of population immunity also plays a crucial role in shaping the disease dynamics and outcomes. To account for this factor in the model, vaccination data could be incorporated as a relevant variable, which is frequently cited as a proxy for population immunity. Lastly, extending the research to other countries and regions could offer comprehensive insights into the global spread of COVID-19 and inform the development of more targeted actions.

Acknowledgements The authors thank Dr. Richard Westaway for providing valuable comments on this study.

References

[1] Akinwumiju, A. S., Oluwafemi, O., Mohammed, Y. D. and Mobolaji, J. W. [2022], 'Geospatial evaluation of COVID-19 mortality: Influence of socio-economic status and underlying health conditions in contiguous USA', *Appl Geogr* **141**, 102671.

[2] Al Kindi, K. M., Al-Mawali, A., Akharusi, A., Alshukaili, D., Alnasiri, N., Al-Awadhi, T., Charabi, Y. and El Kenawy, A. M. [2021], 'Demographic and socioeconomic determinants of COVID-19 across Oman-A geospatial modelling approach', *Geospat Health* **16**(1).

[3] Batty, M., Murcio, R., Iacopini, I., Vanhoof, M. and Milton, R. [2021], London in lockdown: Mobility in the pandemic city, *in* 'COVID-19 pandemic, Geospatial Information, and Community Resilience', CRC Press, pp. 229–244.

[4] Berg, K., Present, P. R. and Richardson, K. [2021], 'Long-term air pollution and other risk factors associated with COVID-19 at the census tract level in Colorado', *Environ Pollut* **287**, 117584.

[5] Bivand, R., Keitt, T., Rowlingson, B., Pebesma, E., Sumner, M., Hijmans, R. et al. [2019], 'rgdal: Bindings for the geospatial data abstraction library', *R package version* **1**(3).

[6] Blangiardo, M. and Cameletti, M. [2015], *Spatial and spatio-temporal Bayesian models with R-INLA*, John Wiley & Sons.

[7] Castro, R., Santos, R., Sousa, G., Pinheiro, Y., Martins, R., Pereira, M. and Silva, R. [2021], 'Spatial dynamics of the COVID-19 pandemic in Brazil', *Epidemiol Infect* **149**.

[8] Cheng, J. [2019], 'Leaflet: create interactive web maps with the JavaScript'Leaflet'', *library* .

[9] Choi, K. H., Denice, P., Haan, M. and Zajacova, A. [2021], 'Studying the social determinants of COVID-19 in a data vacuum', *Can Rev Sociol* **58**(2), 146–164.

[10] Cockings, S., Harfoot, A., Martin, D. and Hornby, D. [2011], 'Maintaining existing zoning systems using automated zone-design techniques: methods for creating the 2011 Census output geographies for England and Wales', *Environment and Planning A* **43**(10), 2399–2418.

[11] Comunian, S., Dongo, D., Milani, C. and Palestini, P. [2020], 'Air pollution and COVID-19: the role of particulate matter in the spread and increase of COVID-19's morbidity and mortality', *Int J Environ Res Public Health* **17**(12), 4487.

[12] Cressie, N., Calder, C. A., Clark, J. S., Hoef, J. M. V. and Wikle, C. K. [2009], 'Accounting for uncertainty in ecological analysis: the strengths and limitations of hierarchical statistical modeling', *Ecol Appl* **19**(3), 553–570.

[13] Davies, N. G., Abbott, S., Barnard, R. C., Jarvis, C. I., Kucharski, A. J., Munday, J., Pearson, C. A., Russell, T. W., Tully, D. C., Washburne, A. D. et al. [2020], 'Estimated transmissibility and severity of novel SARS-CoV-2 Variant of Concern 202012/01 in England', *medRxiv* pp. 2020–12.

[14] Díaz-Avalos, C., Juan, P., Chaudhuri, S., Sáez, M. and Serra, L. [2020], 'Association between the new COVID-19 cases and air pollution with meteorological elements in nine counties of New York state', *Int J Environ Res Public Health* **17**(23), 9055.

[15] Fattorini, D. and Regoli, F. [2020], 'Role of the chronic air pollution levels in the Covid-19 outbreak risk in Italy', *Environ Pollut* **264**, 114732.

[16] Feng, C. [2022], 'Spatial-temporal generalized additive model for modeling COVID-19 mortality risk in Toronto, Canada', *Spat Stat* **49**, 100526.

[17] Fuglstad, G.-A., Simpson, D., Lindgren, F. and Rue, H. [2019], 'Constructing priors that penalize the complexity of Gaussian random fields', *J Am Stat Assoc* **114**(525), 445–452.

[18] Gebhard, C., Regitz-Zagrosek, V., Neuhauser, H. K., Morgan, R. and Klein, S. L. [2020], 'Impact of sex and gender on COVID-19 outcomes in europe', *Biol Sex Differ* **11**, 1–13.

[19] Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. [1995], *Bayesian data analysis*, Chapman and Hall/CRC.

[20] Goutte, S., Péran, T. and Porcher, T. [2020], 'The role of economic structural factors in determining pandemic mortality rates: Evidence from the COVID-19 outbreak in france', *Res. Int. Bus. Finance* **54**, 101281.

[21] Green, M. A. and Semple, M. G. [2023], 'Occupational inequalities in the prevalence of COVID-19: A longitudinal observational study of England, August 2020 to January 2021', *PLoS One* **18**(4), e0283119.

[22] Haining, R., Li, G., Maheswaran, R., Blangiardo, M., Law, J., Best, N. and Richardson, S. [2010], 'Inference from ecological models: estimating the relative risk of stroke from air pollution exposure using small area data', *Spat Spatiotemporal Epidemio* **1**(2-3), 123–131.

[23] Harris, R. and Brunsdon, C. [2021], 'Measuring the exposure of Black, Asian and other ethnic groups to COVID-infected neighbourhoods in English towns and cities', *Appl Spat Anal Policy* pp. 1–26.

[24] Hartigan, J. A. and Wong, M. A. [1979], 'Algorithm AS 136: A k-means clustering algorithm', *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **28**(1), 100–108.

[25] Huang, G., Blangiardo, M., Brown, P. E. and Pirani, M. [2021], 'Long-term exposure to air pollution and COVID-19 incidence: a multi-country study', *Spat Spatiotemporal Epidemio* **39**, 100443.

[26] Jackson, S. L., Derakhshan, S., Blackwood, L., Lee, L., Huang, Q., Habets, M. and Cutter, S. L. [2021], 'Spatial disparities of COVID-19 cases and fatalities in United States counties', *International Journal of Environmental Research and Public Health* **18**(16), 8259.

[27] Kim, B., Rundle, A. G., Goodwin, A. T. S., Morrison, C. N., Branas, C. C., El-Sadr, W. and Duncan, D. T. [2021], 'COVID-19 testing, case, and death rates and spatial socio-demographics in New York City: An ecological analysis as of June 2020', *Health Place* **68**, 102539.

[28] Konstantinoudis, G., Padellini, T., Bennett, J., Davies, B., Ezzati, M. and Blangiardo, M. [2021], 'Long-term exposure to air-pollution and COVID-19 mortality in England: a hierarchical spatial analysis', *Environment international* **146**, 106316.

[29] Kulu, H. and Dorey, P. [2021], 'Infection rates from Covid-19 in Great Britain by geographical units: A model-based estimation from mortality data', *Health & place* **67**, 102460.

[30] Lee, D., Robertson, C. and Marques, D. [2021], 'Quantifying the small-area spatio-temporal dynamics of the Covid-19 pandemic in Scotland during a period with limited testing capacity', *Spat Stat* p. 100508.

[31] Lee, D., Robertson, C., McRae, C. and Baker, J. [2022], 'Quantifying the impact of air pollution on Covid-19 hospitalisation and death rates in Scotland', *Spat Spatiotemporal Epidemio* **42**, 100523.

[32] Leffler, C. T., Ing, E., Lykins, J. D., Hogan, M. C., McKeown, C. A. and Grzybowski, A. [2020], 'Association of country-wide coronavirus mortality with demographics, testing, lockdowns, and public wearing of masks', *Am J Trop Med Hyg* **103**(6), 2400.

[33] 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', *J R Stat Soc Series B Stat Methodo* **73**(4), 423–498.

[34] Liu, C., Liu, Z. and Guan, C. [2021], 'The impacts of the built environment on the incidence rate of COVID-19: A case study of King County, Washington', *Sustainable cities and society* **74**, 103144.

[35] Liu, W., Wang, D., Hua, S., Xie, C., Wang, B., Qiu, W., Xu, T., Ye, Z., Yu, L., Yang, M. et al. [2021], 'Spatiotemporal analysis of COVID-19 outbreaks in Wuhan, China', *Sci. Rep* **11**(1), 1–9.

[36] Ljung, G. M. and Box, G. E. [1978], 'On a measure of lack of fit in time series models', *Biometrika* **65**(2), 297–303.

[37] Maiti, A., Zhang, Q., Sannigrahi, S., Pramanik, S., Chakraborti, S., Cerda, A. and Pilla, F. [2021], 'Exploring spatiotemporal effects of the driving factors on COVID-19 incidences in the contiguous United States', *Sustain Cities Soc* **68**, 102784.

[38] Mansour, S., Al Kindi, A., Al-Said, A., Al-Said, A. and Atkinson, P. [2021], 'Sociodemographic determinants of COVID-19 incidence rates in Oman: Geospatial modelling using multiscale geographically weighted regression (MGWR)', *Sustain Cities Soc* **65**, 102627.

[39] Martines, M. R., Ferreira, R. V., Toppa, R. H., Assunção, L., Desjardins, M. R. and Delmelle, E. M. [2021], 'Detecting space–time clusters of COVID-19 in Brazil: mortality, inequality, socioeconomic vulnerability, and the relative risk of the disease in Brazilian municipalities', *J Geogr Syst* **23**, 7–36.

[40] Mathur, R., Rentsch, C. T., Morton, C. E., Hulme, W. J., Schultze, A., MacKenna, B., Eggo, R. M., Bhaskaran, K., Wong, A. Y., Williamson, E. J. et al. [2021], 'Ethnic differences in SARS-CoV-2 infection and COVID-19-related hospitalisation, intensive care unit admission, and death in 17 million adults in England: an observational cohort study using the OpenSAFELY platform', *Lancet* **397**(10286), 1711–1724.

[41] Mecenas, P., Bastos, R. T. d. R. M., Vallinoto, A. C. R. and Normando, D. [2020], 'Effects of temperature and humidity on the spread of COVID-19: A systematic review', *PLoS One* **15**(9), e0238339.

[42] Menard, S. [2002], *Applied logistic regression analysis*, number 106, Sage.

[43] Mendy, A., Wu, X., Keller, J. L., Fassler, C. S., Apewokin, S., Mersha, T. B., Xie, C. and Pinney, S. M. [2021], 'Air pollution and the pandemic: Long-term PM2. 5 exposure and disease severity in COVID-19 patients', *Respirology* **26**(12), 1181–1187.

[44] Mollalo, A., Vahedi, B. and Rivera, K. M. [2020], 'Gis-based spatial modeling of covid-19 incidence rate in the continental united states', *Science of the total environment* **728**, 138884.

[45] Moran, P. A. [1950], 'Notes on continuous stochastic phenomena', *Biometrika* **37**(1/2), 17–23.

[46] Muegge, R., Dean, N., Jack, E. and Lee, D. [2023], 'National lockdowns in England: The same restrictions for all, but do the impacts on COVID-19 mortality risks vary geographically?', *Spat Spatiotemporal Epidemio* **44**, 100559.

[47] Nazia, N., Law, J. and Butt, Z. A. [2022], 'Identifying spatiotemporal patterns of covid-19 transmissions and the drivers of the patterns in toronto: a bayesian hierarchical spatiotemporal modelling', *Sci. Rep* **12**(1), 9369.

[48] Neuwirth, E. and Brewer, R. C. [2014], 'ColorBrewer palettes', *R package version* **1**, 4.

[49] R Core Team [2013], *R: A language and environment for statistical computing*, Vienna, Austria.

[50] Richardson, S., Martinez, J., Hirsch, J. S., Cerise, J., Lesser, M., Roswell, R. O., Davidson, K. W. and Consortium, N. H. C.-. R. [2022], 'Association of race/ethnicity with mortality in patients hospitalized with COVID-19', *PLoS One* **17**(8), e0267505.

[51] Roger Bivand [2022], 'R packages for Analyzing Spatial Data: A Comparative Case Study with Areal Data', *Geogr Anal* **54**(3), 488–518.

[52] Rothman, K. J., Greenland, S., Lash, T. L. et al. [2008], *Modern epidemiology*, Vol. 3, Wolters Kluwer Health/Lippincott Williams & Wilkins Philadelphia.

[53] Rue, H. and Held, L. [2005], *Gaussian Markov random fields: theory and applications*, Chapman and Hall/CRC.

[54] Rue, H., Martino, S. and Chopin, N. [2009], 'Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations', *J R Stat Soc Series B Stat Methodo* **71**(2), 319–392.

[55] Sarkar, S. K., Ekram, K. M. M. and Das, P. C. [2021], 'Spatial modeling of COVID-19 transmission in bangladesh', *Spat. Inf. Res* pp. 1–12.

[56] Sartorius, B., Lawson, A. and Pullan, R. [2021], 'Modelling and predicting the spatio-temporal spread of COVID-19, associated deaths and impact of key risk factors in England', *Scientific reports* **11**(1), 1–11.

[57] Semczuk-Kaczmarek, K., Rys-Czaporowska, A., Sierdzinski, J., Kaczmarek, L. D., Szymanski, F. M. and Platek, A. E. [2021], 'Association between air pollution and COVID-19 mortality and morbidity', *Intern Emerg Med* pp. 1–7.

[58] Sun, Y., Hu, X. and Xie, J. [2021], 'Spatial inequalities of COVID-19 mortality rate in relation to socioeconomic and environmental factors across England', *Sci Total Environ* **758**, 143595.

[59] Takagi, H., Kuno, T., Yokoyama, Y., Ueyama, H., Matsushiro, T., Hari, Y. and Ando, T. [2021], 'Meta-regression of COVID-19 prevalence/fatality on socioeconomic characteristics of data from top 50 US large cities', *J Med Virol* **93**(2), 595.

[60] Tamrakar, V., Srivastava, A., Saikia, N., Parmar, M. C., Shukla, S. K., Shabnam, S., Boro, B., Saha, A. and Debbarma, B. [2021], 'District level correlates of COVID-19 pandemic in India during March-October 2020', *PLoS One* **16**(9), e0257533.

[61] Wakefield, J. and Salway, R. [2001], 'A statistical framework for ecological and aggregate studies', *J R Stat Soc Ser A Stat Soc* **164**(1), 119–137.

[62] Wang, J., Tang, K., Feng, K., Lin, X., Lv, W., Chen, K. and Wang, F. [2021], 'Impact of temperature and relative humidity on the transmission of COVID-19: a modelling study in China and the United States', *BMJ Open* **11**(2), e043863.

[63] Wang, J., Tang, K., Feng, K., Lv, W. et al. [2020], 'High temperature and high humidity reduce the transmission of COVID-19', *Available at SSRN* **3551767**, 2020b.

[64] WHO [2020], 'World Health Organization: Coronavirus Disease 2019 (COVID-19) Situation Report'.

[65] Wickham, H. [2016], *ggplot2: Elegant Graphics for Data Analysis*, Springer-Verlag New York.
URL: *https://ggplot2.tidyverse.org*

[66] Wickham, H., François, R., Henry, L., Müller, K. et al. [2015], 'dplyr: A grammar of data manipulation', *R package version 0.4* **3**, p156.

[67] Williams, C. K. and Rasmussen, C. E. [2006], *Gaussian processes for machine learning*, Vol. 2, MIT press Cambridge, MA.

[68] Wintle, B. A., McCarthy, M. A., Volinsky, C. T. and Kavanagh, R. P. [2003], 'The use of Bayesian model averaging to better represent uncertainty in ecological models', *Conserv Biol* **17**(6), 1579–1590.

[69] Wong, D. W. and Li, Y. [2020], 'Spreading of COVID-19: Density matters', *PLoS One* **15**(12), e0242398.

[70] Wu, X., Nethery, R. C., Sabath, M. B., Braun, D. and Dominici, F. [2020], 'Air pollution and COVID-19 mortality in the United States: Strengths and limitations of an ecological regression analysis', *Sci Adv* **6**(45), eabd4049.

[71] Wu, X. and Zhang, J. [2021], 'Exploration of spatial-temporal varying impacts on COVID-19 cumulative case in Texas using geographically weighted regression (GWR)', *Environmental Science and Pollution Research* **28**, 43732–43746.

[72] Zhai, W., Liu, M., Fu, X. and Peng, Z.-R. [2021], 'American inequality meets COVID-19: Uneven spread of the disease across communities', *Ann Am Assoc Geogr* **111**(7), 2023–2043.

[73] Zhang, J., Wu, X. and Chow, T. E. [2021], 'Space-time cluster's detection and geographical weighted regression analysis of COVID-19 mortality on Texas counties', *International Journal of Environmental Research and Public Health* **18**(11), 5541.