Capturing Spatial Dependence of COVID-19 Case Counts with Cellphone Mobility Data

Justin J. Slater*[1,2], Patrick E. Brown[1,2], Jeffrey S. Rosenthal[1], and Jorge Mateu[3]

[1]Department of Statistical Sciences, University of Toronto

[2]Centre for Global Health Research, St. Michael's Hospital

[3]Department of Mathematics, University Jaume I of Castellon

Abstract

Spatial dependence is usually introduced into spatial models using measure of physical proximity. When analyzing COVID-19 case counts, this makes sense as regions that are close together are more likely to have more people moving between them, spreading the disease. However, using the actual number of trips between each region may explain COVID-19 case counts better than physical proximity. In this paper, we investigate the efficacy of using telecommunications-derived mobility data to induce spatial dependence in spatial models applied to two Spanish communities' COVID-19 case counts. We do this by extending Besag York Mollié (BYM) models to include both a physical adjacency effect, alongside a mobility effect. The mobility effect is given a Gaussian Markov random field prior, with the number of trips between regions as edge weights. We leverage modern parametrizations of BYM models to conclude that the number of people moving between regions better explains variation in COVID-19 case counts than physical proximity data. We suggest that this data should be used in conjunction with physical proximity data when developing spatial models for COVID-19 case counts.

Keywords: Bayesian hierarchical model, Besag York Mollié model, COVID-19, Gaussian Markov random field, Mobility data

Declarations of interest: None

*Address correspondence to `justin.slater@mail.utoronto.ca`

Funding Information: This research was funded by the Natural Sciences and Engineering Research Council (RGPIN-2017-06856) and the Ministry of Science and Innovation (PID2019-107392RB-I00).

1 Introduction

Spatial analyses of COVID-19 case data were first published as early as March of 2020 [1–3], in an attempt to characterize, predict, and attenuate the severity of the pandemic. Subsequent studies have noted substantial spatial dependence in COVID-19 case counts [4, 5]. This makes sense as regions that are close to each other likely have more people moving between them, spreading the disease to nearby regions.

Many groups have attempted to model COVID-19 case counts as a function of climate [6–8], healthcare quality [9], socioeconomic factors [10] and more. More recently, mobility data has become more abundant and popular for modeling COVID-19 transmission. This makes sense because the disease spreads through human contact, meaning that case counts are likely to be a function of the number of people moving around. Such mobility data has been used to model the evolution of the epidemic in Spain [11, 12], assess the effectiveness of the Spanish lockdown [13], monitoring the epidemic in Switzerland [14], identify at-risk populations in France during a lockdown [15], individual-level infection tracing in China [16], assess the timing of stay-home orders [17], and evaluating the effectiveness of social distancing in the United States [18]. This data can be found in many forms, but is commonly found in the form of aggregated areal *mobility matrices*. If we denote a mobility matrix $\boldsymbol{M}$, $[\boldsymbol{M}]_{ij}$ corresponds to the number of trips from region i to region j, and $\boldsymbol{M}_{ii}$ represents the number of trips within region i.

These data have been applied in a variety of different models to answer numerous questions, but lack of available methods makes it difficult for researchers to use this data to its full potential. In this paper, we demonstrate a novel method for analyzing this data, whereby the mobility data is used as edge weights in a Gaussian Markov random field (network) model. Previous work using network models have been applied to mobility data in the form of a network compartment model [19] which was used to conduct inference regarding societal inequities, and inform reopening. This work does not aim to make such claims, but rather demonstrate the efficacy of mobility data in modern parametrizations of Besag, York, and Mollié (BYM) models [20] and their extensions.

BYM models have been used frequently in the spatial analysis literature due to their effectiveness and computational efficiency. In these models, the spatial component is comprised of Conditional Autoregressive (CAR) [21] models and conventional random effects. This means that the spatial effect of region i depends only on its "neighbours". Neighbours could be defined by any quantity the analyst has access to, but is most often defined by physical adjacency, i.e. if two regions share a common border, they are considered neighbours. Several ICAR/BYM models have been applied to COVID-19 data with neighbours defined in this way [22–24]. Although these spatial model components based on physical adjacency are powerful and computationally efficient, it makes more sense to use mobility between regions to induce spatial dependence in COVID-19 models because the disease spreads via person-to-person contact.

In this paper, we build a BYM model where mobility data is used to induce spatial dependence between regions. Using mobility data within two Communities in Spain, Madrid and Castilla-Leon, we demonstrate the value of mobility data for COVID-19 spatial modeling applications. Furthermore, we extend modern parametrizations of BYM models to account for both physical adjacency and mobility simultaneously, and show that mobility data captures spatial variation in COVID-19 case counts much more accurately than physical adjacency alone.

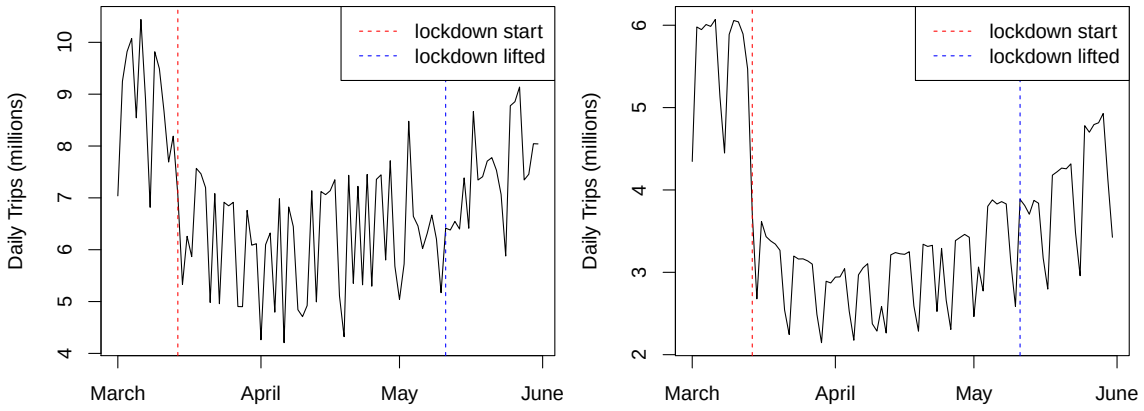
This is a short focused paper with the following plan. Section 2 presents the data and the modeling strategy based on particular parametrizations of BYM models. The results come in Section 3, and the paper ends with a final discussion in Section 4.

2 Methods

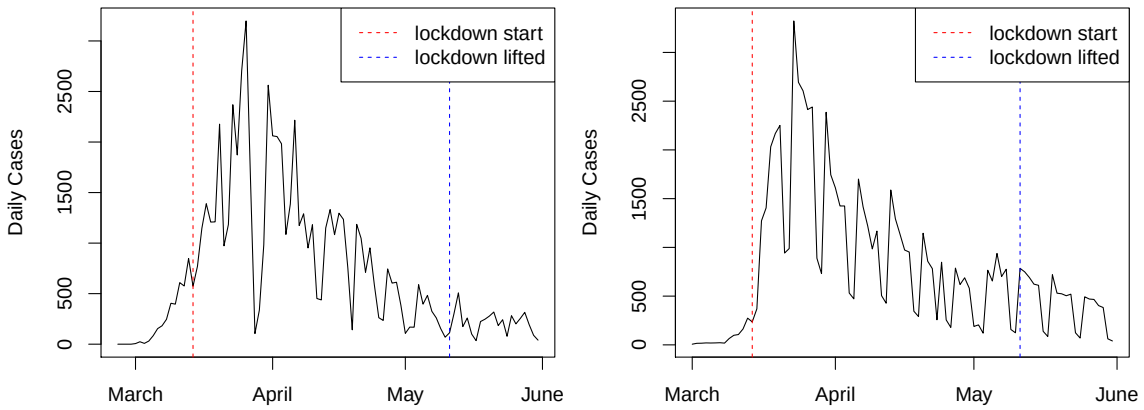
2.1 Data

This paper is focused on two regions in Spain. Castilla-Leon is the largest Community in Spain by area and is located in the northwest part of Spain, with a population of 2.5 million. The Community of Madrid is located in the central part of Spain and has a population of around 6.8 million, and it is home of the capital of the country, Madrid City, with 3.3 million inhabitants.

The human mobility data was obtained from Barcelona Supercomputing Center Flow-map dashboard [25]. Trips within Madrid and Castilla-Leon were extracted from over 13 million phone records provided by a Spanish cellphone company. Both passive



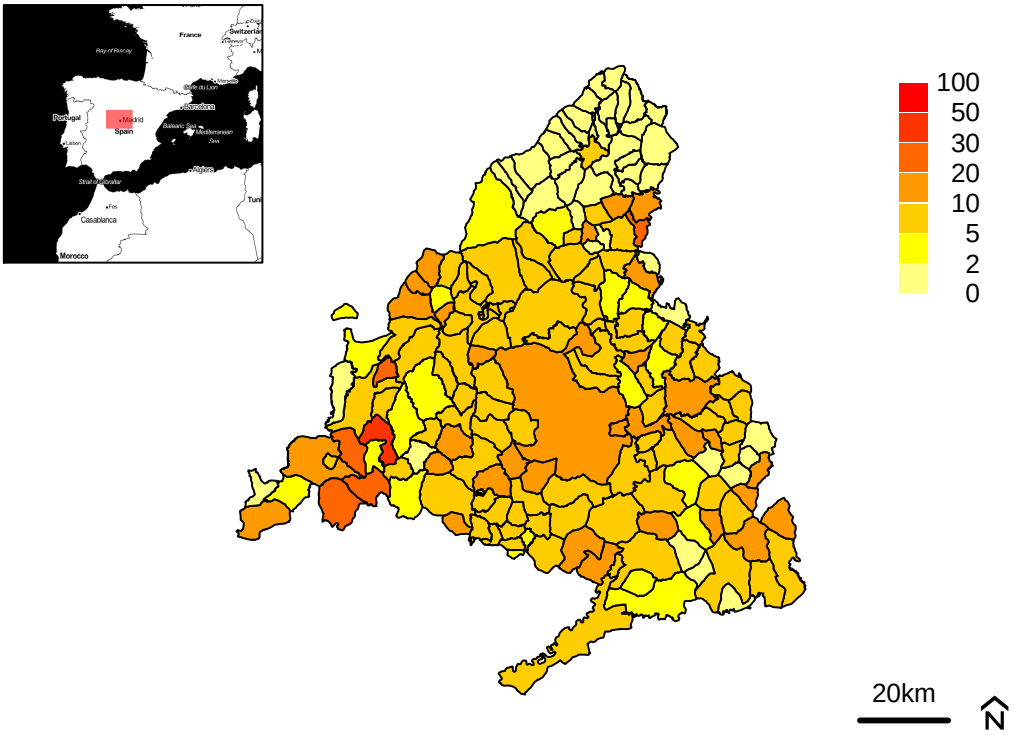
(a) Madrid (b) Castilla-Leon



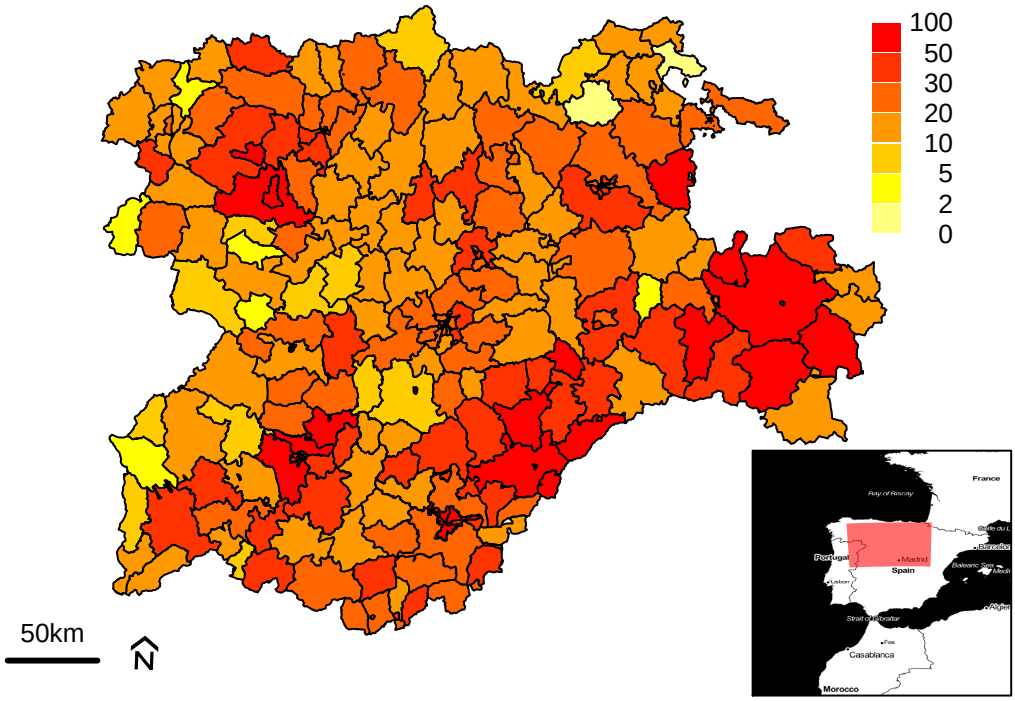
(c) Madrid (d) Castilla-Leon

Figure 1: Number of trips greater than 500 metres (a and b) and daily case counts (c and d) in the

two Communities of Spain from March to June 2020.



(a) Madrid



(b) Castilla-Leon

Figure 2: COVID-19 cases per thousand, up to May 31 2020 for two communities in Spain. Background map ©Stamen Design.

(GPS) and active (text messages, calls etc.) data were aggregated to construct daily movement matrices in each of the Communities, prior to the authors acquisition of the data. Given that trips were only recorded from one cellphone company, adjustment was made to estimate the number of total trips between each region. As a result, the entries of the mobility matrices are non-integer values.

Figures 1a and 1b show the total daily movement between regions in Madrid and Castilla-Leon. There is a sharp drop in the number of trips around March 14th 2020, which corresponds to a nation-wide lockdown. Lockdown restrictions began to ease around May 11th, where the number of trips slowly began to rise. Figures 1c and 1d show the number of cases of COVID-19 cases in both Communities. COVID-19 daily cases data were retrieved from the open data portal of Castilla-Leon [26] and from the Epidemiological Surveillance Network of Madrid [27]. Notice that the movement drops as cases rise, because a lockdown was implemented in response to the increasing severity of the epidemic. In order to avoid this potential "reverse causality" problem, we will only use movement data in the first week of March. Our justification for this is that there is a time lag between when the virus spreads and the resulting COVID cases are confirmed. That is, the "first wave" of the epidemic was likely influenced mostly by the movement that occurred prior to the peak in cases, and less by the movement that occurred during it.

Figure 2 shows the spatial distribution of the COVID-19 case rates up until May 31, 2020. The cases per thousand people range from (approximately) $0-30$ in Madrid, and $0-100$ for Castilla-Leon. We can see that there is substantial variation in the case rates within each of these Communities. Note that the extreme values in these plots are mostly small regions, which makes sense since the variance of case rates is higher when population is small. In the north of Madrid, there is a cluster of municipalities that have very low case rates. In Castilla-Leon, case rates are highest near the southeast border, which is the border to Madrid.

Figure 3 shows the number of trips to, from, and within each Municipality of Madrid (there are 179 of these small regions), and Castilla-Leon (there are 245 health zones). Madrid and Castilla-Leon are considered separately throughout this paper. Although they are adjacent, data on movements between the two communities are not available. In Madrid, there is a lot of movement in and around Madrid City, and less movement in the more rural areas. Castilla-Leon shows a less predictable movement pattern, as there

is not a single capital city that accounts for most of the movement. This movement data will be used to induce spatial correlation between regions, as described in Section 2.3.

2.2 Spatial autoregressive models

Besag, York, and Mollié (BYM) models [20] are widely used in spatial epidemiology and disease mapping due to their simplicity and computational efficiency. They assume the incidence of disease in region i follows a Poisson distribution

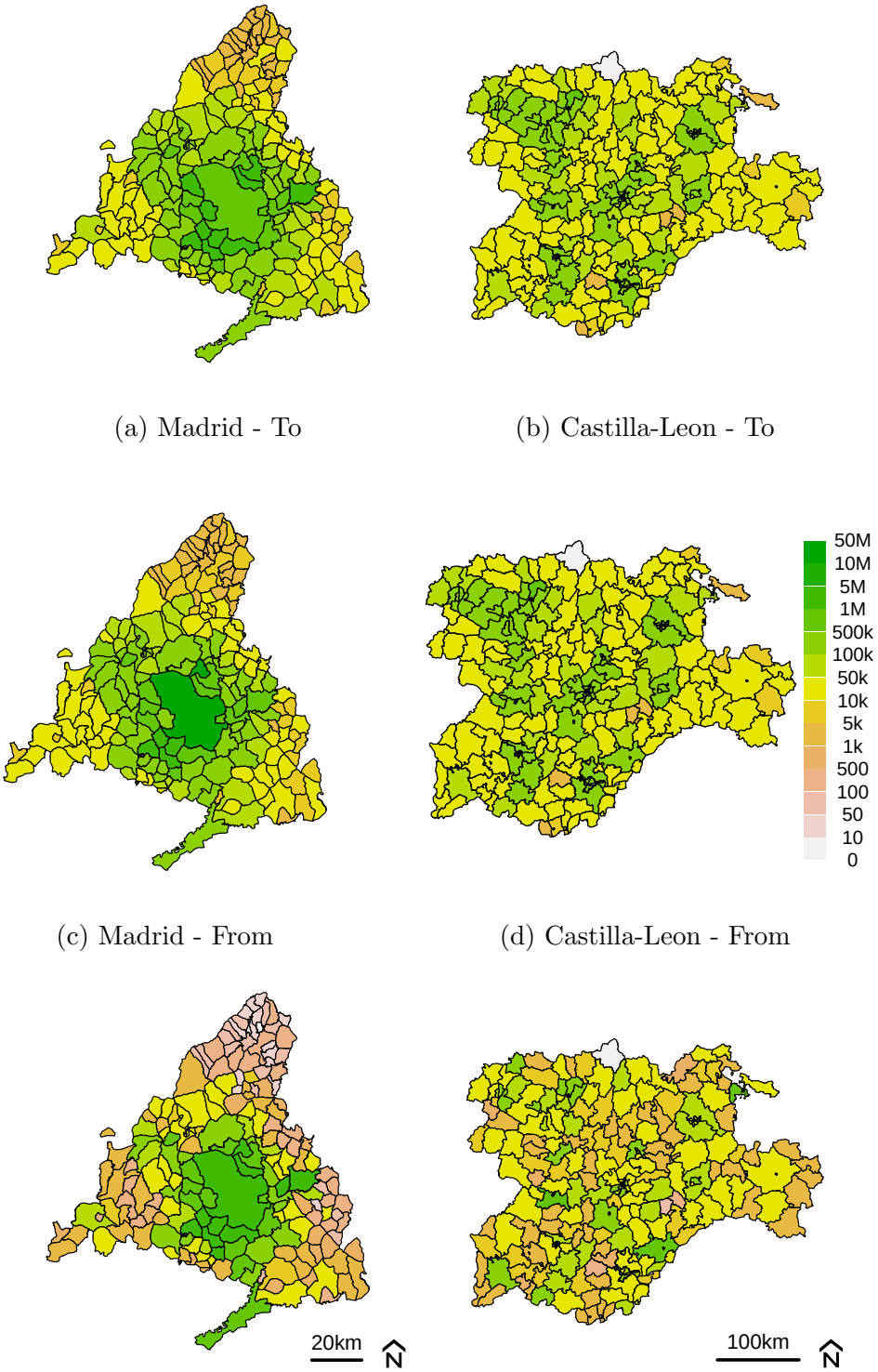
$$Y_i \sim \text{Pois}(E_i \lambda_i)$$

where Y_i is the number of infected cases in region i, and E_i is some form of expected count or offset, which could be the at-risk population, exposure time, etc. The log-relative risk, λ_i, is often modeled as

$$\begin{aligned} \log(\lambda_i) &= \mu + \beta X + \phi_i + \theta_i \\ \phi_i | \boldsymbol{\phi}_{-i} &\sim N\left(\frac{1}{\sum_j w_{ij}} \sum_j w_{ij} \phi_j,\ \frac{\sigma_\phi^2}{\sum_j w_{ij}} \right) \\ \theta_i &\overset{i.i.d}{\sim} N(0, \sigma_\theta^2) \end{aligned} \tag{1}$$

where μ is the overall intercept, β is the effect of spatial covariates, ϕ_i is the structured spatial random effect, and θ_i is the unstructured spatial random effect which allows for overdispersion in the response. In the spatial formulation of the BYM model, w_{ij} = 1 when regions i and j share a common border, and 0 otherwise. That is, region i's structured spatial effect is only conditionally dependent on its neighbours, given all other regions. The distributions $\{\phi_i | \boldsymbol{\phi}_{-i}\}_{i=1}^n$ are known as the *full conditionals*, where $\boldsymbol{\phi}_{-i}$ is short hand for the set $\{\phi_1, \phi_2, ...\phi_{i-1}, \phi_{i+1}, ...\phi_n\}$. We can see from (1) that $E(\phi_i | \boldsymbol{\phi}_{-i})$ is a weighted average of its neighbours, resulting in spatial smoothing. These full conditionals correspond to the joint distribution of the ϕ's being a Gaussian Markov random field (GMRF) [28], with

$$\begin{aligned} \boldsymbol{\phi} &\sim \boldsymbol{MVN}(\mathbf{0}, \boldsymbol{Q}^{-1}) \\ \boldsymbol{Q} &= \sigma_\phi^{-2} \boldsymbol{D}(\boldsymbol{I} - \boldsymbol{W}) \end{aligned}$$



(a) Madrid - To

(b) Castilla-Leon - To

(c) Madrid - From

(d) Castilla-Leon - From

(e) Madrid - Within

(f) Castilla-Leon - Within

Figure 3: Number of trips (incoming, outgoing, and within) the 179 regions of Madrid, and 245 health zones of Castilla-Leon, for the period March 1 to March 7 2020.

where $\boldsymbol{W}$ is a matrix of weights such that $w_{ij} > 0$ for $i \neq j$ and $w_{ii} = 0$, and σ^2 is a variance parameter to be estimated. $\boldsymbol{D}$ is a diagonal matrix such that $\boldsymbol{D}_{ii} = \sum_j w_{ij}$. This definition ensures that the precision matrix, $\boldsymbol{Q}$, is both symmetric and positive definite. In addition to the 0-1 weights based on regions being adjacent, other weighting schemes, such as inverse of euclidean distance between regions, have been used. For a comparison of common weighting schemes, see [29]. When we specify $\boldsymbol{Q}$ in this way, we refer to this as an Intrinsic Autoregressive (ICAR) model for $\boldsymbol{\phi}$. The joint density function has a computationally convenient form with

$$p(\boldsymbol{\phi}) \propto \exp\Big[-\frac{1}{2\sigma_\phi}\sum_{i<j} w_{ij}(\phi_i - \phi_j)^2\Big]$$

which is sometimes referred to as *the pairwise difference formula.* Notice that this density is invariant to the addition of a constant to each ϕ_i, leaving the spatial random effects unidentifiable up to a constant. This is typically remedied by imposing the constraint $\sum_i \phi_i = 0$ [29]. We will now modify this BYM model to account for movement between regions, in addition to physical adjacency.

2.3 Movement augmented BYM model

In order to extend the BYM model to allow for spatial correlation based on movement data, a second ICAR term, γ_i, with dependence structure governed by the movement data is added to the model. We also retain an adjacency-determined spatial effect ϕ_i in order to infer the relative importance of mobility-based and adjacency-based spatial dependence in determining COVID-19 case counts. The resulting model is

$$\begin{aligned}
\log(\lambda_i) &= \mu + \beta X_i + \phi_i + \gamma_i + \theta_i \\
\phi_i|\boldsymbol{\phi}_{-i} &\sim N\left(\frac{1}{\sum_j w_{ij}}\sum_j w_{ij}\phi_j, \frac{\sigma_\phi^2}{\sum_j w_{ij}}\right) \\
\gamma_i|\boldsymbol{\gamma}_{-i} &\sim N\left(\frac{1}{\sum_j v_{ij}}\sum_j v_{ij}\gamma_j, \frac{\sigma_\gamma^2}{\sum_j v_{ij}}\right) \\
\theta_i &\sim N(0, \sigma_\theta^2)
\end{aligned}$$

where ϕ_i and γ_i are the spatial random effects with priors based on the physical data and movement data respectively. The geographically-defined process ϕ_i has weights $w_{ij} = 1$ if regions i and j share a common border and are 0 otherwise, while the

movement-defined process γ_i has weights v_{ij} representing the number of trips between regions i and j. Using mobility as edge weights in network models has shown to be effective in the context of infectious diseases [30–32]. [30] used mobility weights in an autoregressive term, which allowed the weights matrices to be asymmetric. However, given that our mobility data is being used in a Gaussian prior for a random effect, the precision matrices of $\boldsymbol{\phi}$ and $\boldsymbol{\gamma}$, Q_ϕ and Q_γ, must be symmetric. Therefore we require $w_{ij} = w_{ji}$ and $v_{ij} = v_{ji}$. While the first equality will always be true, the mobility matrices are not perfectly symmetric, thus symmetry was induced by defining v_{ij} as the sum of the numbers of trips from i to j and from j to i. The GRMF does not account for the movement within a region, so the movement within a region was included in the model as a spatial covariate X_i (fixed effect). That is, X_i was computed as

$$X_i = \frac{\frac{v_{ii}}{E_i} - \underset{j}{\text{mean}}(\frac{v_{jj}}{E_j})}{\underset{j}{\text{sd}}(\frac{v_{jj}}{E_j})}$$

where v_{ii}/E_i is the number of trips per person within a region, and $\text{mean}(v_{jj}/E_j)$ and $\text{sd}(v_{jj}/E_j)$ are the mean and standard deviations of the trips per person in all other regions. This model was run on both the Madrid and Castilla-Leon data.

There are two main drawbacks with the formulations of BYM models presented thus far. Firstly, the interpretation of the parameters σ_γ and σ_ϕ depend on the average number of neighbours and the total number of trips for each region, and hence their magnitudes are not comparable [33]. Secondly, $\sigma_\phi, \sigma_\gamma$, and σ_θ are hard to estimate without very careful choices of hyperpriors [34]. We will now address these shortcoming via reparametrizations.

2.4 Reparametrizations and Priors

In order to solve issues with comparability, interpretability, and estimation, we apply a reparameterization of our model that is inspired by [35] with

$$\begin{aligned}
\sigma^2 &\approx \text{Var}(\phi_i + \gamma_i + \theta_i) \\
\phi_i^* | \boldsymbol{\phi}_{-i}^* &\sim N\left(\frac{1}{\sum_j w_{ij}} \sum_j w_{ij}\phi_{ij}^*,\ \frac{\rho_\phi \sigma^2}{s_\phi \sum_j w_{ij}}\right) \\
\gamma_i^* | \boldsymbol{\gamma}_{-i}^* &\sim N\left(\frac{1}{\sum_j v_{ij}} \sum_j v_{ij}\gamma_{ij}^*,\ \frac{\rho_\gamma \sigma^2}{s_\gamma \sum_j v_{ij}}\right) \\
\theta_i &\sim N(0, \rho_\theta \sigma^2)
\end{aligned}$$

where $\rho_\phi + \rho_\gamma + \rho_\theta = 1$ and $0 < \rho_\gamma, \rho_\phi, \rho_\theta < 1$. The priors for σ and $\boldsymbol{\rho}$ are

$$\begin{aligned}
\sigma &\sim N_+(0, 1) \\
\boldsymbol{\rho} &\sim \text{Dirichlet}(1, 1, 1)
\end{aligned}$$

Note that

$$\begin{aligned}
\phi_i^* &= \sigma\left(\sqrt{\rho_\phi / s_\phi}\right)\phi_i \\
\gamma_i^* &= \sigma\left(\sqrt{\rho_\gamma / s_\gamma}\right)\gamma_i.
\end{aligned}$$

Here, σ is the combined variance of the spatial effects, and the ρ's are mixing parameters, interpreted as the proportion of the combined spatial variance explained by each model component. Note that $\rho_\theta = 1$ reduces the spatial component to purely overdispersion, $\rho_\phi = 1$ reduces the spatial component of the model to an adjacency ICAR model for the spatial effects, and $\rho_\gamma = 1$ reduces the spatial component to a mobility ICAR model. Most importantly, if $\rho_\gamma > \rho_\phi$ then this means that the mobility data better explains variation in COVID-19 case counts than the adjacency data. As long as the spatial weights matrix and the mobility weights matrix are linearly independent, then having both spatial and mobility terms in our model present no issues with identifiability [36]. Finally, s_γ and s_ϕ are scaling factors, such that the geometric means of $s_\gamma^{-1}\text{Var}(\gamma_i)$ and $s_\phi^{-1}\text{Var}(\phi_i)$ are both ≈ 1 for each i, meaning that γ_i^* and ϕ_i^* are the log relative risk contributions from the movement data and physical data respectively [33]. Scaling is absolutely necessary in order to conduct inference on the ρ's. We compute

the scaling factors as follows

$$s = \exp\left(\frac{1}{n}\sum_{i=1}^{n}\log[\boldsymbol{Q}^{-}]_{ii}\right)$$

where $\boldsymbol{Q}^{-}$ is the generalized inverse of the $n \times n$ precision matrix [37]. In order to scale the precision matrices of the spatial effects, the generalized inverse for sparse matrices from [38] was used. The diagonal elements, $[\boldsymbol{Q}^{-}]_{ii}$, of $\boldsymbol{Q}^{-}$ are referred to as the *marginal variances* of the structured spatial effects, i.e $\text{var}(\phi_i) = [\boldsymbol{Q}_{\phi}^{-}]_{ii}$ and $\text{var}(\gamma_i) = [\boldsymbol{Q}_{\gamma}^{-}]_{ii}$.

As was the case with the ICAR model in (1), we can derive the full conditionals of the combined spatial effect, $\tau_i = \phi_i^* + \gamma_i^* + \theta_i^*$, for the model described in Section 2.3

$$\tau_i|\boldsymbol{\tau}_{-i} \sim N\left[\frac{\sum_j(\frac{\rho_\phi}{s_\phi}w_{ij} + \frac{\rho_\gamma}{s_\gamma}v_{ij})\tau_j}{\frac{\rho_\phi}{s_\phi}\sum_j w_{ij} + \frac{\rho_\gamma}{s_\gamma}\sum_j v_{ij} + \rho_\theta}, \frac{\sigma^2}{\frac{\rho_\phi}{s_\phi}\sum_j w_{ij} + \frac{\rho_\gamma}{s_\gamma}\sum_j v_{ij} + \rho_\theta}\right] \tag{2}$$

These full conditionals can help provide some intuition as to the mechanism by which this model provides spatial smoothing. As $\rho_\gamma \to 1$, τ_i is simply the weighted sum of the other regions, where the weights are the proportion of region i's total movement between each other region. If $\rho_\phi \to 1$, the conditional mean of τ_i reduces to the arithmetic average of the spatial effects of its neighbours. If $\rho_\theta \to 1$, then the conditional mean shrinks to 0 (remember that $\rho_\phi + \rho_\gamma + \rho_\theta = 1$). Given that ρ_θ is positive, the conditional mean is always shrunk towards 0, resulting in spatial smoothing. In practice, the conditional mean will be a weighted average of the estimates smoothed by the movement GMRF, the physical GMRF and 0. It is important to note here that the w_{ij}/s_ϕ and v_{ij}/s_γ are relative measures due to the scaling factors. That is, doubling the total amount of movement has no effect on the conditional mean or variance of τ_i. This is in contrast to the combined spatial effects in the commonly used Leroux model [34]. Additionally, the variance of $\tau_i|\boldsymbol{\tau}_{-i}$ is lower when region i has a lot of movement or many neighbours, relative to the other regions.

2.5 Inference, computation, and validation

Four chains each with 3000 iterations of No U-Turn Sampling were used for parameter estimation within Stan [39]. The first 1500 iterations were used as a warm-up, the 1500 remaining iterations from each chain were thinned by a factor of 10, leaving 600 total posterior samples to perform inference. As mentioned in Section 2.2, we require

$\sum_i \phi_i = 0$. In practice, we use the soft constraint

$$\sum_i \phi_i \sim N(0, 0.001)$$

for computation purposes (as recommended by the Stan team [40]). To complete the model, priors for β and μ were $N(0, 1)$. To ensure the robustness of our results, we also ran BYM models using the adjacency data and the movement data separately. That is, for both Madrid and Castilla-Leon, we ran a model where we assumed $\rho_\gamma = 0$, and a separate model where $\rho_\phi = 0$. The results of these four models are presented in Section 3.2.

Our code and posterior samples are posted at `https://github.com/cghr-toronto/public/tree/master/covid/spain_public_code`.

3 Results

3.1 Joint model

Table 1 shows posterior medians and credible intervals for the mixing parameters for the model with both movement and adjacency spatial effects. For both Madrid and Castilla-Leon, the proportion of spatial variation explained by $\boldsymbol{\gamma}$ is much higher than that of $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$. The posterior probability that $\rho_\gamma > \rho_\phi$ was 0.997 for Madrid, and 0.998 for Castilla-Leon. However, $\boldsymbol{\phi}$ does seem to account for a non-trivial amount of spatial variation in both Madrid and Castilla-Leon. This means that although movement data is likely more explanatory, adjacency data can help with explaining variation in COVID-19 cases. Additionally, there is a substantial amount of spatial variation explained by the unstructured spatial effect for Madrid. This is not the case for Castilla-Leon, as most of the mass of the posterior of ρ_θ is near 0. This makes sense given that Madrid has a large metropolitan centre surrounded by a mix of suburbs and rural areas, so there are probably spatial confounders that our model is missing. For a plot of the posterior densities of $\boldsymbol{\rho}$, see Appendix A.

Figures 4a through 4d show the spatial distribution $\boldsymbol{\gamma}^*$ and $\boldsymbol{\phi}^*$, plotted using the same colour scale for comparability. We can see that $\boldsymbol{\gamma}$'s log-relative risks have a lot more spatial variation in both Commmunities. The log-relative risks for $\boldsymbol{\phi}$ tend to have smooth spatial gradients, while $\boldsymbol{\gamma}$ tends to identify clusters of regions as high-risk areas.

	Parameter	Madrid Est (95% CrI)	Castilla-Leon Est (95% CrI)
$\boldsymbol{\rho}$	Movement	0.76 (0.54, 0.89)	0.88 (0.66, 0.98)
	Neighbour	0.13 (0.01, 0.39)	0.09 (0.01, 0.30)
	Independent	0.10 (0.02, 0.25)	0.02 (0.00, 0.09)
	μ	-5.36 (-5.51, -5.24)	-3.75 (-3.78, -3.73)
	β	0.12 (0.05, 0.20)	-0.01 (-0.04, 0.02)
	σ	0.65 (0.55, 0.78)	0.72 (0.63, 0.83)

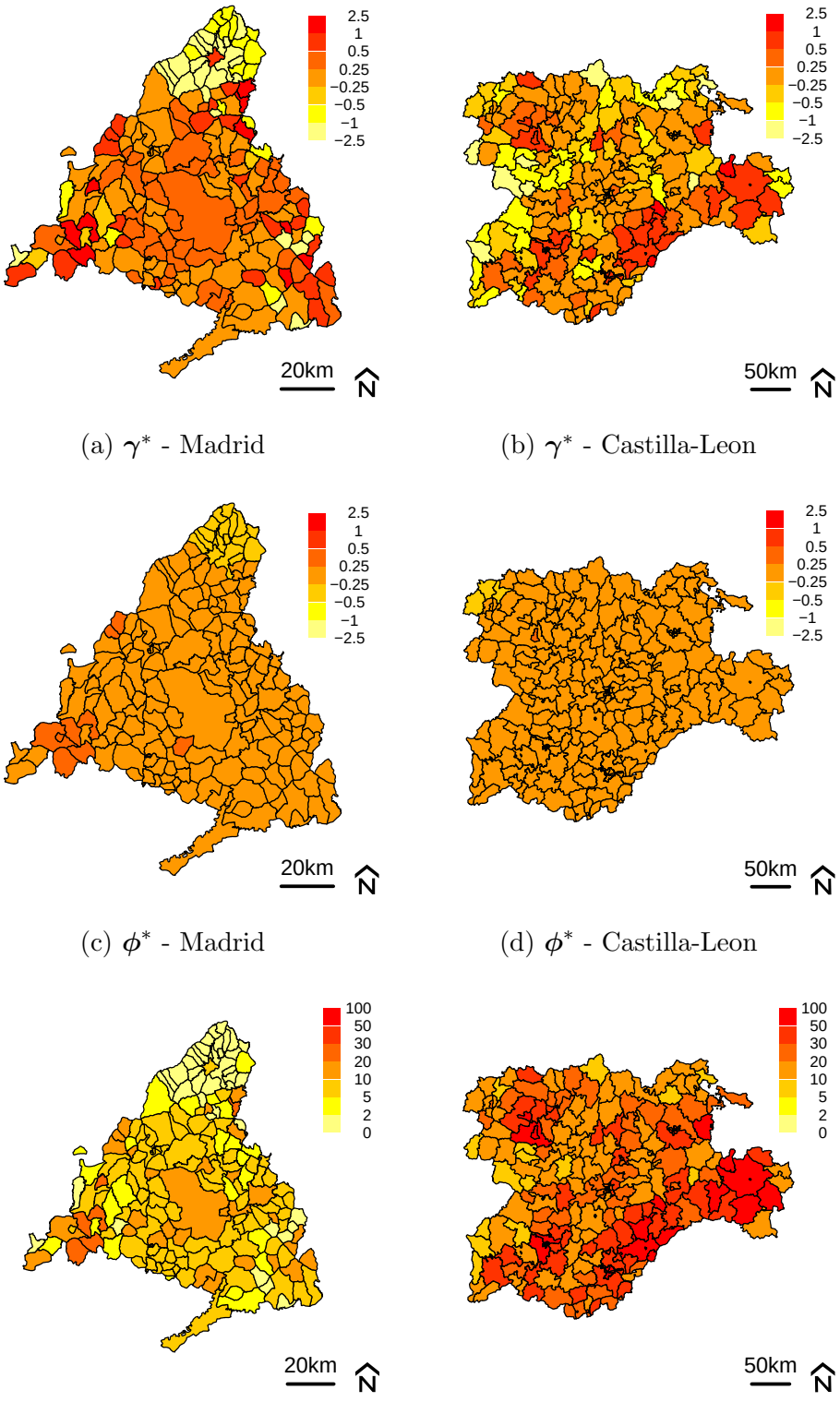
Table 1: Posterior medians, and 95% credible intervals for $\boldsymbol{\rho}$ in BYM models using movement and physical (adjacency) data in the same model.

As seen in equation 2, the expectation of the combined spatial effects are a weighted average of these spatial effects, and 0 (notice that the numerator can be rewritten as $\sum_j (\frac{\rho_\phi}{s_\phi} w_{ij} + \frac{\rho_\gamma}{s_\gamma} v_{ij} + \rho_\theta \cdot 0)\tau_j$ where $\rho_\theta > 0$). Figures 4e and 4f show the predicted cases per 1000 people per region, showing highly similar patterns to the observed values in Figure 2.

The standard deviation was slightly larger for Castilla-Leon than it was for Madrid. Figure B.2 shows the the spatial distribution of the standard deviation of the cases per thousand people in both communities. Here, we can see that the standard deviation is pretty small in and around Madrid-city, because the movement to and from Madrid-city is causing a high-degree of spatial smoothing in the surrounding area. The effect of movement within regions, β, is associated with larger case counts in Madrid, but this is not the case for Castilla-Leon. This small covariate effect could result in more variance being attributable to the random effects, potentially contributing to the larger σ in Castilla-Leon.

3.2 Model Validation - Individual models

Table 2 shows posterior medians and credible intervals for the ρ parameter from the movement and physical BYM models described in Section 2.5, fit separately to Madrid and Castilla-Leon (four models total). In both regions, the model where spatial smoothing is induced by population movement explains a higher proportion of the variation in the outcome, indicated by the posterior density of ρ having more mass near 1. Additionally, the BYM model that used physical adjacency as a spatial smoother had a much wider credible interval for ρ, indicating more model uncertainty. Both models show more uncertainty in the region of Madrid than for Castilla-Leon, likely due to the fact that Madrid is more heterogeneous in terms of population density and other



(a) γ^* - Madrid

(b) γ^* - Castilla-Leon

(c) ϕ^* - Madrid

(d) ϕ^* - Castilla-Leon

(e) λ^*1000 - Madrid

(f) λ^*1000 - Castilla-Leon

Figure 4: Log-relative risk contributions (a-d) from the movement effects (γ^*) and spatial effects effects (ϕ^*). The predicted cases per thousand people are also presented (e-f).

	Parameter	Madrid Est (95% CrI)	Castilla-Leon Est (95% CrI)
ρ	movement	0.82 (0.66, 0.91)	0.95 (0.89, 0.98)
	neighbour	0.56 (0.22, 0.83)	0.77 (0.58, 0.91)
μ	movement	-5.34 (-5.48, -5.23)	-3.75 (-3.78, -3.73)
	neighbour	-5.18 (-5.30, -5.09)	-3.74 (-3.78, -3.70)
β	movement	0.12 (0.05, 0.18)	-0.02 (-0.05, 0.02)
	neighbour	0.13 (0.01, 0.24)	-0.01 (-0.05, 0.04)
σ	movement	0.63 (0.55, 0.76)	0.74 (0.65, 0.83)
	neighbour	0.66 (0.56, 0.83)	0.58 (0.51, 0.66)

Table 2: Posterior medians, and 95% credible intervals for ρ in BYM models using movement and physical (adjacency) data in separate models.

factors. For full posterior densities of the ρ parameter, see Appendix A.2.

4 Discussion

In this paper, we have demonstrated that there is much value in using mobility data in combination with geographical proximity for defining correlation structures COVID-19 incidence data. We showed that even while using only one week of movement data, we were able to explain the spatial variation in COVID-19 counts better than using the classic BYM model. Additionally, we showed that the model can be re-parametrized so that the means by which smoothing occurs in these mobility models is intuitive.

A key limitation of this work is that the models presented in this paper do not serve as individual-level infectious disease models, as correlation is induced by a latent effect rather than direct dependence between the counts. However, this will be a natural extension of this work and would require the addition of many more parameters, including multiple mobility network components at various time points. This will ultimately pose a computational challenge as well.

An additional limitation of this work is that the availability and structure of mobility data will vary across data sources, and may only be available in higher income countries. Furthermore, there is selection bias in the movement data, as it only tracks those who actually have a cellphone, which may tend to be younger and more economically advantaged individuals. Given potential differences in quality of this data, its efficacy in spatial models may need to be assessed on a case by case basis.

Furthermore, the models presented in this paper may suffer from overfitting. A potential remedy for this would be to put a penalized complexity prior [41] on the mixing

parameters, which may improve inference by shrinking ρ_γ (and perhaps ρ_ϕ) towards 0. An interesting area for future work would be to combine Dirichlet and penalized complexity priors to specify a joint prior for the mixing parameters as described in [42], which can be implemented using the *makemyprior* R package [43]. This was deemed unnecessary for this work, as we were mainly interested in comparing ρ_γ to ρ_ϕ, and felt that our prior should not favour either one of these terms.

Despite these limitations, this work demonstrates the value of mobility data and provides the foundation for various extensions and future work. This data is only becoming more abundant as time passes, and methods that allow for efficient use of this data are essential to model the current epidemic, and any spatial epidemiological application where population movement is likely a predictor of disease.

References

[1] H. Huang, Y. Wang, Z. Wang, Z. Liang, S. Qu, S. Ma, G. Mao, and X. Liu, "Epidemic features and control of 2019 novel coronavirus pneumonia in Wenzhou, China," *Preprints with The Lancet*, 2020. http://dx.doi.org/10.2139/ssrn.3550007.

[2] Z. Arab-Mazar, R. Sah, A. A. Rabaan, K. Dhama, and A. J. Rodriguez-Morales, "Mapping the incidence of the covid-19 hotspot in iran – implications for travellers," *Travel Medicine and Infectious Disease*, vol. 34, p. 101630, 2020.

[3] D. Giuliani, M. M. Dickson, G. Espa, and F. Santi, "Modelling and predicting the spatio-temporal spread of coronavirus disease 2019 (COVID-19) in Italy," *Preprints with The Lancet*, 2020. http://dx.doi.org/10.2139/ssrn.3559569.

[4] D. Kang, H. Choi, J.-H. Kim, and J. Choi, "Spatial epidemic dynamics of the COVID-19 outbreak in China," *International Journal of Infectious Diseases*, vol. 94, pp. 96–102, 2020.

[5] U. Bilal, S. Barber, L. Tabb, and A. V. Diez-Roux, "Spatial inequities in COVID-19 testing, positivity, incidence and mortality in 3 US cities: a longitudinal ecological study," *medRxiv*, 2020. https://doi.org/10.1101/2020.05.01.20087833.

[6] J. Liu, J. Zhou, J. Yao, X. Zhang, L. Li, X. Xu, X. He, B. Wang, S. Fu, T. Niu, *et al.*, "Impact of meteorological factors on the COVID-19 transmission: A multi-city study in China," *Science of the Total Environment*, vol. 726, p. 138513, 2020.

[7] P. Shi, Y. Dong, H. Yan, C. Zhao, X. Li, W. Liu, M. He, S. Tang, and S. Xi, "Impact of temperature on the dynamics of the COVID-19 outbreak in China," *Science of the Total Environment*, vol. 728, p. 138890, 2020.

[8] Á. Briz-Redón and Á. Serrano-Aroca, "A spatio-temporal analysis for exploring the effect of temperature on COVID-19 early evolution in Spain," *Science of the Total Environment*, vol. 728, p. 138811, 2020.

[9] M. M. Sugg, T. J. Spaulding, S. J. Lane, J. D. Runkle, S. R. Harden, A. Hege, and L. S. Iyer, "Mapping community-level determinants of COVID-19 transmission in nursing homes: A multi-scale approach," *Science of the Total Environment*, vol. 752, p. 141946, 2021.

[10] C. F. Baum and M. Henry, "Socioeconomic factors influencing the spatial spread of COVID-19 in the United States," *Preprints with The Lancet*, 2020. http://dx.doi.org/10.2139/ssrn.3559569.

[11] F. Aràndiga, A. Baeza, I. Cordero-Carrión, R. Donat, M. C. Martí, P. Mulet, and D. F. Yáñez, "A spatial-temporal model for the evolution of the COVID-19 pandemic in Spain including mobility," *Mathematics*, vol. 8, no. 10, p. 1677, 2020.

[12] S. M. Iacus, C. Santamaria, F. Sermi, S. Spyratos, D. Tarchi, and M. Vespe, "Human mobility and COVID-19 initial dynamics," *Nonlinear Dynamics*, vol. 101, no. 3, pp. 1901–1919, 2020.

[13] L. Orea and I. C. Álvarez, "How effective has the Spanish lockdown been to battle COVID-19? a spatial analysis of the coronavirus propagation across provinces," *Documento de trabajo FEDEA*, vol. 3, pp. 1–33, 2020.

[14] J. Persson, J. F. Parie, and S. Feuerriegel, "Monitoring the COVID-19 epidemic with nationwide telecommunication data," *arXiv preprint arXiv:2101.02521*, 2021.

[15] G. Pullano, E. Valdano, N. Scarpa, S. Rubrichi, and V. Colizza, "Evaluating the effect of demographic factors, socioeconomic factors, and risk aversion on mobility during the COVID-19 epidemic in France under lockdown: a population-based study," *The Lancet Digital Health*, vol. 2, no. 12, pp. e638–e649, 2020.

[16] M. U. Kraemer, C.-H. Yang, B. Gutierrez, C.-H. Wu, B. Klein, D. M. Pigott, L. Du Plessis, N. R. Faria, R. Li, W. P. Hanage, *et al.*, "The effect of human mobility and control measures on the COVID-19 epidemic in China," *Science*, vol. 368, no. 6490, pp. 493–497, 2020.

[17] M. Audirac, M. Tec, L. A. Meyers, S. Fox, and C. Zigler, "How timing of stay-home orders and mobility reductions impacted first-wave COVID-19 deaths in US counties," *medRxiv*, 2020. https://doi.org/10.1101/2020.11.24.20238055.

[18] H. S. Badr, H. Du, M. Marshall, E. Dong, M. M. Squire, and L. M. Gardner, "Association between mobility patterns and COVID-19 transmission in the usa: a mathematical modelling study," *The Lancet Infectious Diseases*, vol. 20, no. 11, pp. 1247–1254, 2020.

[19] S. Chang, E. Pierson, P. W. Koh, J. Gerardin, B. Redbird, D. Grusky, and J. Leskovec, "Mobility network models of COVID-19 explain inequities and inform reopening," *Nature*, vol. 589, no. 7840, pp. 82–87, 2021.

[20] J. Besag, J. York, and A. Mollié, "Bayesian image restoration, with two applications in spatial statistics," *Annals of the institute of statistical mathematics*, vol. 43, no. 1, pp. 1–20, 1991.

[21] J. Besag, "Spatial interaction and the statistical analysis of lattice systems," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 36, no. 2, pp. 192–225, 1974.

[22] C. DiMaggio, M. Klein, C. Berry, and S. Frangos, "Blacks/African Americans are 5 times more likely to develop COVID-19: spatial modeling of New York city zip code-level testing results," *MedRxiv*, vol. 14, p. 2020, 2020.

[23] G. Huang and P. E. Brown, "Population-weighted exposure to air pollution and COVID-19 incidence in Germany," *Spatial Statistics*, vol. 41, p. 100480, 2021.

[24] J. S. Brainard, S. Rushton, T. Winters, and P. R. Hunter, "Spatial risk factors for pillar 1 COVID-19 case counts and mortality in rural eastern England, U.K," *medRxiv*, 2020. https://doi.org/10.1101/2020.12.03.20239681.

[25] A. Valencia, "COVID-19 Flow Maps." https://flowmaps.life.bsc.es/flowboard/. Accessed: Jan 10, 2021.

[26] General Directorate of Information Systems, Quality and Pharmaceutical Provision, "Open Data of Castile and Leon." https://datosabiertos.jcyl.es/web/es/datos-abiertos-castilla-leon.html. Accessed: Jan 10, 2021.

[27] "Epidemiological surveillance network of Madrid." https://datos.gob.es. Accessed: Jan 10, 2021.

[28] H. Rue and L. Held, *Gaussian Markov random fields: theory and applications*. CRC press, 2005.

[29] E. W. Duncan, N. M. White, and K. Mengersen, "Spatial smoothing in bayesian models: a comparison of weights matrix specifications and their impact on inference," *International Journal of Health Geographics*, vol. 16, no. 1, pp. 1–16, 2017.

[30] B. Schrödle, L. Held, and H. Rue, "Assessing the impact of a movement network on the spatiotemporal spread of infectious diseases," *Biometrics*, vol. 68, no. 3, pp. 736–744, 2012.

[31] V. V. Volkova, R. Howey, N. J. Savill, and M. E. Woolhouse, "Sheep movement networks and the transmission of infectious diseases," *PloS one*, vol. 5, no. 6, p. e11185, 2010.

[32] M. Geilhufe, L. Held, S. O. Skrøvseth, G. S. Simonsen, and F. Godtliebsen, "Power law approximations of movement network data for modeling infectious disease spread," *Biometrical Journal*, vol. 56, no. 3, pp. 363–382, 2014.

[33] S. H. Sørbye and H. Rue, "Scaling intrinsic Gaussian Markov random field priors in spatial modelling," *Spatial Statistics*, vol. 8, pp. 39–51, 2014.

[34] B. G. Leroux, X. Lei, and N. Breslow, "Estimation of disease rates in small areas: a new mixed model for spatial dependence," in *Statistical models in epidemiology, the environment, and clinical trials*, pp. 179–191, Springer, 2000.

[35] A. Riebler, S. H. Sørbye, D. Simpson, and H. Rue, "An intuitive bayesian spatial model for disease mapping that accounts for scaling," *Statistical methods in medical research*, vol. 25, no. 4, pp. 1145–1165, 2016.

[36] E. C. Rodrigues and R. Assunçao, "Bayesian spatial models with a mixture neighborhood structure," *Journal of Multivariate Analysis*, vol. 109, pp. 88–102, 2012.

[37] A. Freni-Sterrantino, M. Ventrucci, and H. Rue, "A note on intrinsic conditional autoregressive models for disconnected graphs," *Spatial and spatio-temporal epidemiology*, vol. 26, pp. 25–34, 2018.

[38] H. Rue, A. Riebler, S. H. Sørbye, J. B. Illian, D. P. Simpson, and F. K. Lindgren, "Bayesian computing with INLA: a review," *Annual Review of Statistics and Its Application*, vol. 4, pp. 395–421, 2017.

[39] Stan Development Team, "Stan Modeling Language Users Guide and Reference Manual, version 2.26," 2021. https://mc-stan.org.

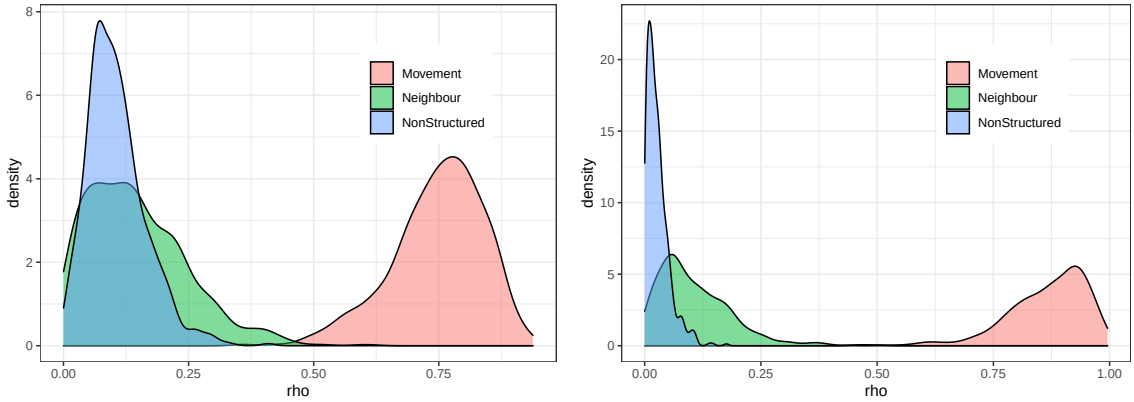
[40] M. Morris, K. Wheeler-Martin, D. Simpson, S. J. Mooney, A. Gelman, and C. DiMaggio, "Bayesian hierarchical spatial models: Implementing the Besag York Mollié model in Stan," *Spatial and Spatio-temporal Epidemiology*, vol. 31, p. 100301, 2019.

[41] D. Simpson, H. Rue, A. Riebler, T. G. Martins, and S. H. Sørbye, "Penalising model component complexity: A principled, practical approach to constructing priors," *Statistical science*, vol. 32, no. 1, pp. 1–28, 2017.

[42] G.-A. Fuglstad, I. G. Hem, A. Knight, H. Rue, and A. Riebler, "Intuitive joint priors for variance parameters," *Bayesian Analysis*, vol. 15, no. 4, pp. 1109–1137, 2020.

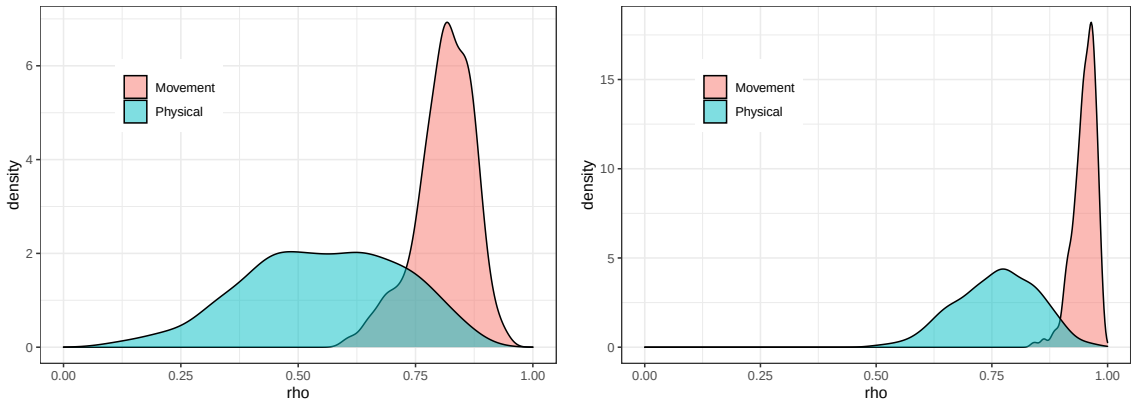
[43] I. G. Hem, G.-A. Fuglstad, and A. Riebler, "makemyprior: Intuitive construction of joint priors for variance parameters in r," *arXiv preprint arXiv:2105.09712*, 2021.

A Appendix Posterior Densities of ρ for various models



(a) Madrid (b) Castilla-Leon

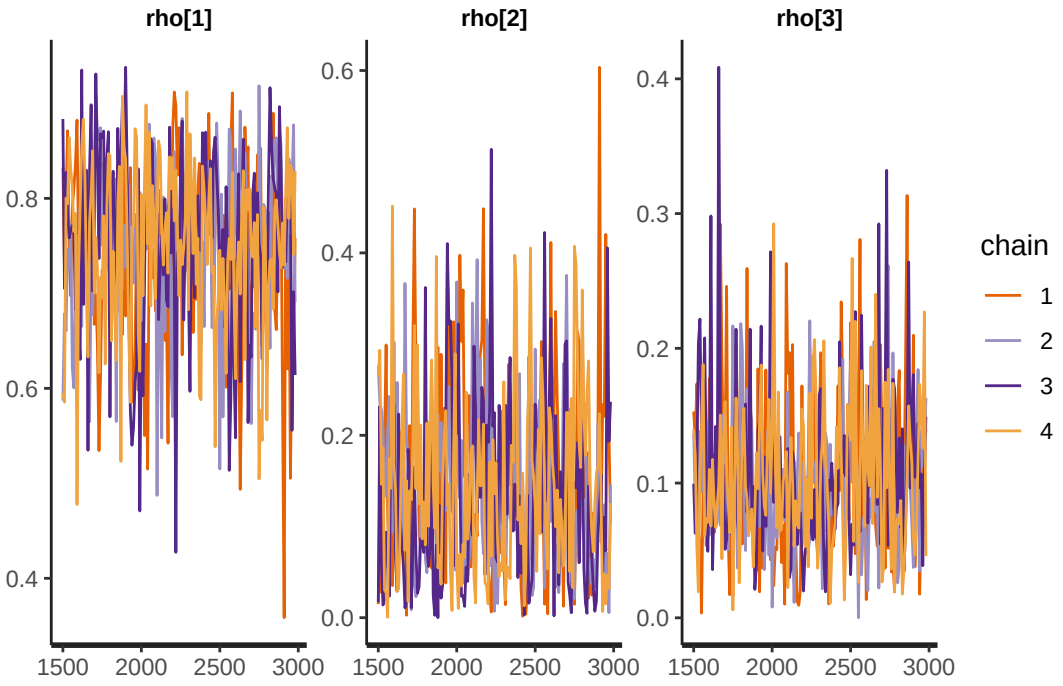
Figure A.1: Posterior Density of the proportion of variance explained by each of the 3 spatial parameters when adjacency and movement data are included in the same model



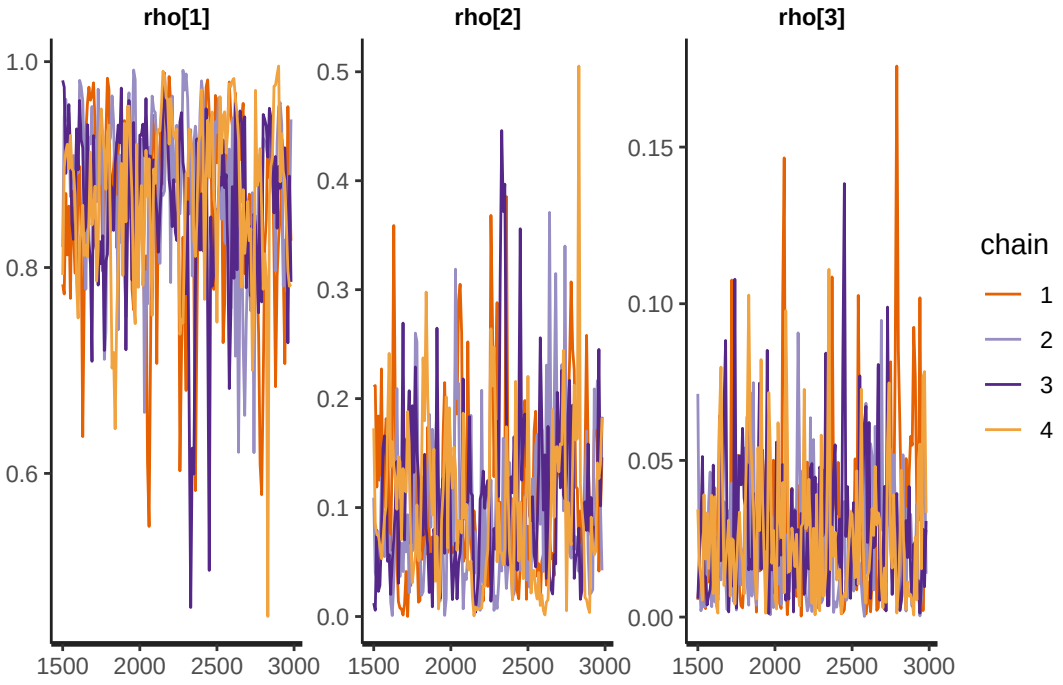
(a) Madrid (b) Castilla-Leon

Figure A.2: Posterior Density of the proportion of variance explained by spatial components when adjacency and movement data are used in separate models (model validation).

B Additional Spatial plots

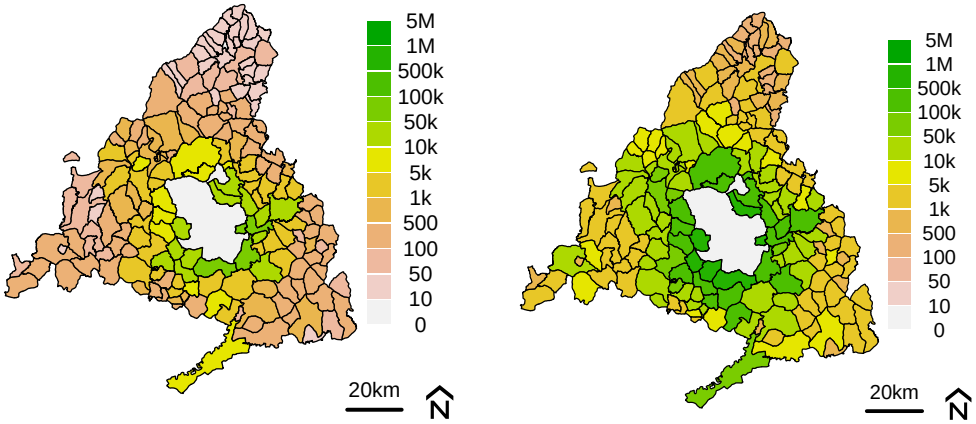


(a) Madrid



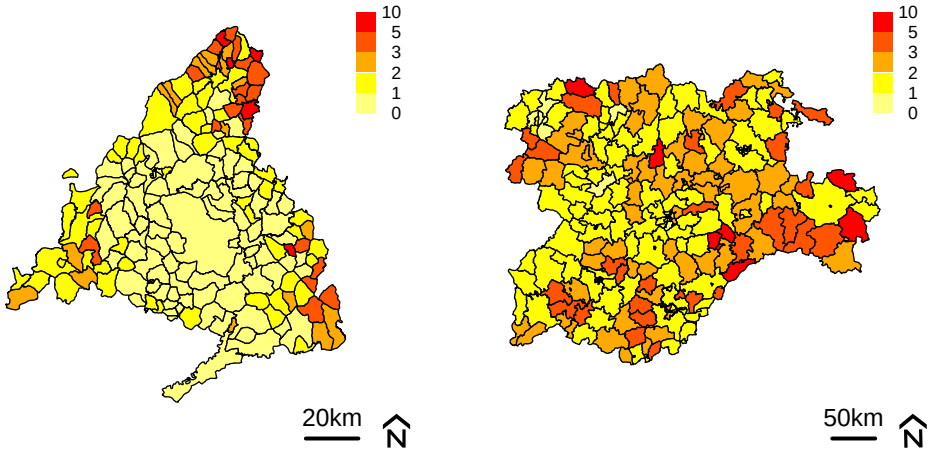
(b) Castilla-Leon

Figure A.3: Traceplots of $\boldsymbol{\rho}$



(a) To (b) From

Figure B.1: Number of trips to and from Madrid City (white).



(a) Madrid (b) Castilla-Leon

Figure B.2: Standard deviations of predicted cases per thousand people.