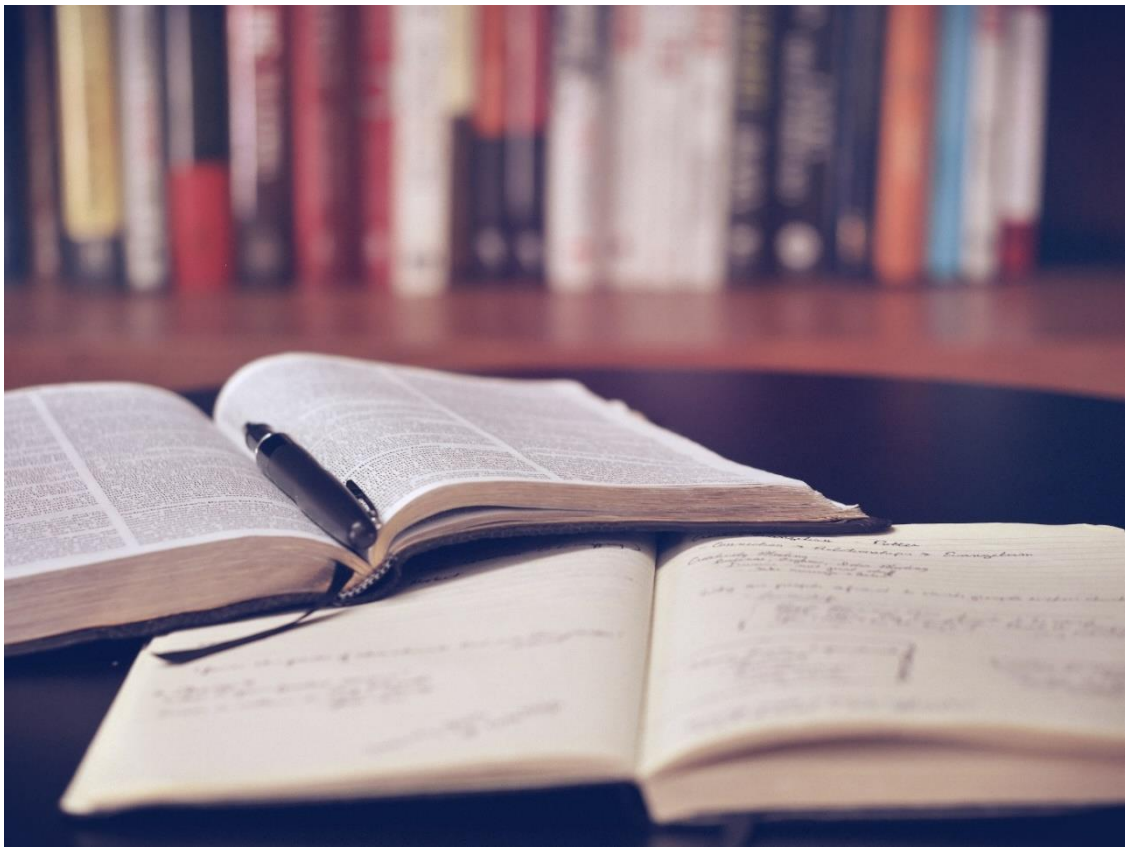
UMR CNRS 6240 LISA

Working Paper TerRA n°22

01-12-2021



A tale of two “AR” models: a spatial analysis of Corsican second home incidence

Yuheng Ling

A tale of two "AR" models: a spatial analysis of Corsican second home incidence

Yuheng LING[1] *†

[1] UMR CNRS 6240 LISA, University of Corsica Pascal Paoli, Corte, France

*Corresponding author.

†E-mail: lyhneo@gmail.com

Abstract

Spatial autoregressive (AR) models can accommodate various forms of dependence among data with discrete support in a space, and hence are widely used in economics and social science. We examine the relationship between spatial (autoregressive) error models and conditional autoregressive models, considered to be the two main types of spatial AR models. This topic is likely incomplete in the literature and is often overlooked by econometricians. To further develop and broaden this topic, we demonstrate that spatial error and conditional autoregressive models can be made equivalent via hierarchal models, but have different variance-covariance matrices. We then propose a Bayesian approach, known as integrated nested Laplace approximations (INLA), to produce accurate estimates for these models and to speed up inferences. We also discuss how to interpret model coefficients, especially estimates of spatial latent effects. We illustrate the two AR models with the proposed methodology in an application to the second home incidence rates of Corsica, France in 2017. We find that both models can capture spatial dependence, but conditional autoregressive models perform slightly better and produce a higher spatial autocorrelation coefficient. We further illustrate estimates of latent effects by identifying several "hot spots" and "cold spots" in terms of second home incidence rates.

Keywords: spatial error model; conditional autoregressive model; hierarchical model; integrated nested Laplace

approximation; Corsican second home incidence rate

Introduction

After decades of development, spatial econometrics has become the dominant approach in exploring spatial interactions in economics and other social science, such as identification and quantification of neighbourhood effects, spatial externalities and network effects (Anselin, 2010; Brueckner, 2003; Dubin et al., 1999). Tobler's first law of geography states that spatial units that are close together tend to have more similar values than those farther away. As a result, ignoring underlying spatial processes, omitting unobserved externalities or misspecifying spatially delineated variables in classical linear regression should yield non-spherical residuals, and further larger standard errors. That is to say, the classical estimator, like the OLS, is inconsistent and inefficient here. To address this issue, researchers often use spatial autoregressive ("AR") models with a specific variance-covariance matrix that is assumed to follow the spatial ordering of the observations.

The two most common spatial "AR" models refer to spatial autoregressive error models[1] (SEMs) (Anselin, 1988) and conditional autoregressive models for latent variables[2] (CAR) (Besag, 1974). The SEM, incorporating simultaneous autoregression on the random disturbances, is often referred to as a spatial econometric model (LeSage and Pace, 2009). It offers a straightforward way to address spatial effects. The CAR model, however, is popular in disease mapping and ecological research (Lee, 2011). As a spatial statistical model, it handles spatial effects via conditional distributions and allows for good prediction (See Figure 1 for a detail description.).

[1] Hereafter the spatial error model.

[2] Hereafter the conditional autoregressive model.

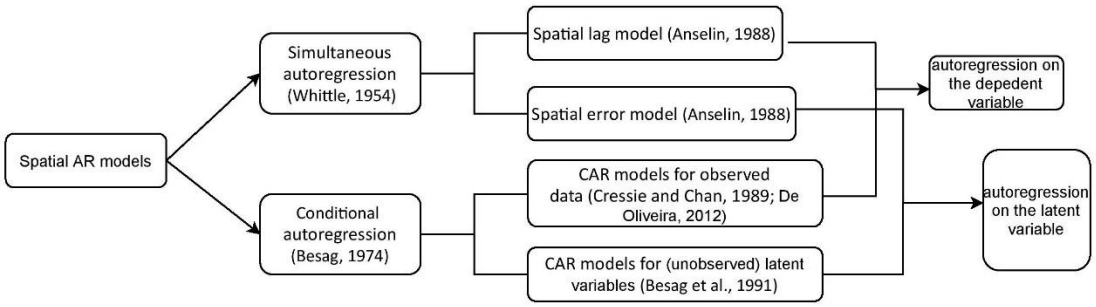


Figure 1. The relationship among spatial "AR" models. According to the nature of the spatial interaction effects considered, spatial "AR" models can be divided into different classes, such as the model contains spatially lagged dependent variables (spatial lag model) and spatially auto-correlated disturbances. In this study, we focus on the model incorporating spatial error autocorrelation.

Several scholars (Griffith and Paelinck, 2007; Kauermann et al., 2012) have argued that in a broad sense, it is difficult to bridge the gap between spatial econometrics and spatial statistics, since they have different objectives and further lead to different interpretations. Nevertheless, as the two most common classes of models for areal data, it deserves to compare them. To our knowledge, few researchers have focused on this topic. Wall (2004) looked into the relationship between SEM and CAR models regarding their spatial autocorrelation parameters, and showed that for a given ρ, the marginal correlation augmented faster in SEMs than in CAR models. Ver Hoef and his collegues (2018) investigated the two models via mathematical properties of their covariance matrices and demonstrated the equivalence of SEMs and CAR models under several conditions. However, econometricians often employ SEMs rather than CAR models, and look over the implicit relation between the two models. Moreover, inference on these models often falls into a classical framework, such as the maximum likelihood method or the generalized method of moments, but Bayesian approaches to spatial models are relatively few. Last but not least, little attention is paid to how the result of these models should be interpreted. Hence, the objective of this article is threefold.

First, we integrate both SEMs and CAR models into a hierarchical modeling framework. More importantly, we show that the SEM model, with the exception of its spatial covariance structure, is identical to the CAR model for modeling non-Gaussian distributed geographic data, i.e., data with a Poisson, binomial, or negative binomial distribution. Second, we investigate the estimation techniques for the two models. In particular, with the recent advances in Bayesian computing, the integrated nested Laplace approximation (INLA) is proposed to fit these

models. Third, we discuss the economic interpretation of spatial components.

Regarding the contribution of this study, we initially bridge the gap between SEMs and CAR models by employing the hierarchical modeling framework for several count data families. For statistical inference, we use hierarchical models coupled with Bayesian approaches as an alternative to likelihood-based estimation for obtaining the coefficients of SEMs. In particular, INLA has been used in various disciplines, including ecology (Beguin et al., 2012) and epidemiology (Moraga, 2019), but to the best of our knowledge, it has been overlooked in spatial econometrics. As such, it is novel for spatial econometrics in general. Lastly, we go into the economic interpretation of spatially-structured random components, which has received little attention in the literature. Although there are no indirect (spillover) effects (Elhorst, 2010) in these models, we suggest that the spatial component can be interpreted as local or neighbourhood shocks diffused as a result of geographical closeness (Ertur et al., 2006), rather than the common shocks in the literature (Andrews, 2005). With the greater availability of geo-referenced data nowadays and a growing interest in spatial modeling, our work would advance the knowledge of spatial autoregressive models and thus enrich the field of spatial econometrics.

Review of SEM and CAR models

Before delving into the two "AR" models, we examine the associated data type. First, AR models are designed for data or a process that contains the information from observations in a space of interest (spatial domain). Second, these data or processes have discrete spatial support, meaning that the information $p(\boldsymbol{k})$ $\boldsymbol{k} = 1 \dots, K,$ is collected correspondingly to areal units $\mathbf{S} = \{S_1, \dots, S_K\}$ with exact boundaries. The aggregation of these areas makes up the entire spatial domain $\mathcal{D}$ ($\boldsymbol{s} \in \mathcal{D} \subset \mathbb{R}^2$). In spatial statistics, these data are called areal data or lattice data (Wikle et al., 2019). In economics, examples of areal data include the regional GDP, the poverty rate in census tracts, or bilateral trade volumes per country (Baltagi et al., 2008; Le Gallo and Ertur, 2003). Due to the Tobler's law, it can be expected that spatial autocorrelation appears in the data and should be tackled through modeling.

Simultaneous autoregression

For notation, let Z be an n-length random vector, and $\mathrm{z_i}$ is a random variable for observations at the i th area. W is a $\mathrm{n} \times \mathrm{n}$ non-stochastic spatial weights matrix with zero diagonals. ρ is the spatial autocorrelation parameter. Simultaneous autoregression (Whittle, 1954) represents that the outcome at region i is simultaneously dependent on that at other (neighbouring) regions j $(\mathrm{i} \neq \mathrm{j})$. Following the notation above, $\mathrm{z_i}$ causes the simultaneous autoregression of each random variable on its neighbours ($\mathrm{z_i} = \rho \sum_{j \in \partial_i}^{m} w_{ij} z_j + \eta_\mathrm{i}$)[3]. In matrix notation, this is equivalent to $\mathrm{Z} = \rho\mathrm{WZ} + \mathrm{H}$, with the vector of disturbances Z. We typically rely on normal assumptions for the residual term H, with $\mathrm{E(H)} = 0$ and $\mathrm{Cov(H)} = \sigma_\eta^2 I_n$. After some matrix algebra, this equation follows as:

$$\mathrm{Z} - \rho\mathrm{WZ} = \mathrm{H}.$$

If the matrix $\mathrm{I} - \rho\mathrm{W}$ is invertible, we have

$$\mathrm{Z} = (\mathrm{I} - \rho\mathrm{W})^{-1}\mathrm{H}.$$

Thus, Z has mean $\mathrm{E(Z)} = 0$, but its covariance takes the form $\mathrm{Cov(Z)} = \sigma_\mathrm{Z}^2\big((I - \rho\mathrm{W})(I - \rho W')\big)^{-1}$.

Conditional autoregression

The conditional autoregression was proposed by Besag (1974). Wall (2004) stated that both simultaneous and conditional autoregression took ideas from autoregressive time series models. Simultaneous autoregression is analogous in its functional form, but conditional autoregression is analogous in its Markov property. As such, in an auto-normal CAR process, the spatial dependence between a pair of region i and region j is modelled conditionally as a normal random variable. In the full conditional distribution of z_i, the mean value is a linear combination of the neighbouring values. Its variance is associated with the values in the ith row of W, and hence is nonstationary.

---

[3] LeSage and Pace (2009) called it the spatial autoregressive process.

$$z_i | z_{j,i\neq j} \sim \mathcal{N}\left(\rho \sum_{j\in\partial_i}^{m} w_{ij} z_j, \sigma_{ii}^2\right)$$

where ∂_i denotes the index of neighbours of region i. Besag (1974) proved that the joint distribution was multivariate normal using Hammersley–Clifford theorem,

$$Z \sim \mathrm{MVN}(0, \sigma_z^2 \Sigma_{CAR})$$
$$\Sigma_{CAR} = (D - \rho W)^{-1},$$

where $D = diag(m_{ii})$ is a $n \times n$ diagonal matrix with m_{ii} indicating the number of neighbours for region i. For a validated conditional autoregressive covariance matrix, Σ_{CAR} must be symmetric and positive definite.

In short, the conditional autoregression implies the full distribution of a random vector that after a spatial transformation yields a multivariate normal distribution with the CAR covariance specification. While, the simultaneous autoregression begins with $i.i.d.$ disturbances that after a spatial transformation yields a multivariate normal distribution with the SAR covariance specification.

SEMs for count data

Considering that some hidden effects spread across (spatial) units, but are omitted in regression models. To account for this bias, classical regression models incorporate simultaneous autoregression for disturbances, resulting in the SEM.

To illustrate the specification of SEMs, we consider the first case of using a binomial distribution as the likelihood function for the outcome variable $Y \sim \mathrm{Binomial}(n, \theta)$. The classical approach to induce the structural form of SEMs is given,

$$\mu = X\boldsymbol{\beta} + Z \qquad 1$$
$$Z = \rho W Z + H, H \sim N(0, \sigma_Z^2)$$

μ denotes the log odds ($\log(\frac{\theta}{1-\theta})$). Linear predictors contain a series of covariates X and a SAR component (Z) to capture the spatially-structured latent effects.

The second approach derives the reduced form of SEMs. That is, the above-mentioned random vector Z is inserted in the Eq.1 directly,

$$\mathrm{Y} \sim \mathrm{Binomial}(\mathrm{n}, \theta),$$
$$\log\left(\frac{\theta}{1-\theta}\right) = \mu,$$
$$\mu = \mathrm{X}\boldsymbol{\beta} + \mathrm{Z},$$
$$\mathrm{Z} \sim \mathrm{MVN}\left(0, \sigma_{\mathrm{Z}}^2 \Sigma_{SAR}\right),$$
$$\Sigma_{SAR} = \sigma_{\mathrm{Z}}^2\left((I - \rho\mathrm{W})(I - \rho W')\right)^{-1}.$$

The second representation reduces to a generalized linear mixed model (GLMM) (Mcculloch and Neuhaus, 2014), where $\mathrm{X}\beta$ is the fixed covariate effect and Z is the individual-specific random intercepts assumed to follow a multivariate normal distribution. The positive definite covariance matrix Σ_{SAR} captures the spatial dependence across the individuals. As such, the similarity between the SEM and the GLMM can be seen. In fact, for the Bayesian version of these GLMMs, a latent variable representation can be used to facilitate computation when fitting the model.

CAR models

As previously stated, conditional/simultaneous autoregression can be used to assess spatial dependence, and SEMs for count data are shown to be a subset of GLMMs. CAR models can easily be applied in a GLMM, and act as a prior distribution for random effects. Such a specification has been popular for decades in the statistical literature, known as the "BYM" model (Besag et al., 1991) and its variants. Similar to the reduced form of SEMs, the GLMM with CAR random effects for binomial outcomes takes the form:

$$\mathrm{Y} \sim \mathrm{Binomial}(\mathrm{n}, \theta),$$
$$\log\left(\frac{\theta}{1-\theta}\right) = \mu,$$
$$\mu = \mathrm{X}\boldsymbol{\beta} + \mathrm{Z},$$
$$\mathrm{Z} \sim \mathrm{MVN}(0, \sigma_z^2 \Sigma_{CAR})$$
$$\Sigma_{CAR} = (D - \rho W)^{-1}. \qquad (2)$$

In short, both SEMs and CAR models for count data belong to the GLMM framework in terms of the model specification, and spatial dependence is captured by a random effect component.

Conditional autoregression and simultaneous autoregression in hierarchical models

In what follows, we focus on the use of SEMs and CAR models within the hierarchical modeling framework. As previously stated, both SEMs and CAR models can be represented as GLMMs. It is also easy to rewrite a GLMM in a three-stage hierarchical formulation. Following the binomial case above, we have

$$\begin{aligned} Y \sim \pi(y_k|\boldsymbol{x}) \;\; - \;\; data \;\; model \\ \boldsymbol{x}|\theta_1 \sim \mathcal{N}\left(0, \boldsymbol{Q}^{-1}\right) \;\; - \;\; process \;\; model \\ \boldsymbol{\theta} = \theta_1 = \{\sigma_\beta^2, \rho, \sigma_z^2\} \;\; - hyperparameters \end{aligned} \qquad 3$$

This formulation avoids complications arising in fitting SEMs and CAR models directly, since the spatial dependence structure of the outcome is modelled through the process model (i.e., the latent process) rather than the data model (i.e., the likelihood). More specifically, in the data model, y_k is assumed to be conditionally independent, given the latent Gaussian field $\boldsymbol{x}$. The latent field $\boldsymbol{x} = \{\boldsymbol{\beta}, \mathrm{Z}\}$ contains all parameters to be estimated in the process model. We then impose a vague Gaussian prior to $\boldsymbol{\beta}$ and a CAR or SAR prior model to Z. As such, the process model reads $\boldsymbol{x}|\theta_1 \sim \mathcal{N}(0, \boldsymbol{Q}^{-1})$, given the set of hyperparameters $\theta_1 = \{\sigma_\beta^2, \rho, \sigma_z^2\}$. Note that the CAR and SAR priors are also known as Gaussian Markov Random Fields (GMRFs) (Rue and Held, 2005), which are simply high dimensional multivariate normal prior distributions with a sparse precision matrix. Due to binomial-distributed observations, the hyperparameter vector for the whole model is equivalent to the hyperparameters for the latent field, thus denoting $\boldsymbol{\theta} = \theta_1 = \{\sigma_\beta^2, \rho, \sigma_z^2\}$. Lastly, together with the likelihood, latent field and hyperparameters, we can derive the posterior distribution of the joint distribution,

$$\pi(\boldsymbol{x}, \boldsymbol{\theta} \mid \boldsymbol{Y}) = \pi(\boldsymbol{\theta})\pi(\boldsymbol{x} \mid \boldsymbol{\theta}) \prod_{k=1}^{K} \pi(\boldsymbol{y}_k \mid \boldsymbol{x}, \boldsymbol{\theta}). \qquad 4$$

Estimation

Two traditional techniques for fitting SEMs and CAR models are likelihood-based methods and Markov chain

Monte Carlo (McMC) methods (Banerjee et al. 2014).

Researchers can apply the conventional maximum likelihood method to fit SEMs, but modeling count data through SEMs requires more complex likelihood methods, such as restricted maximum likelihood (Wood, 2011) or h-likelihood (Rönnegård et al., 2010) approaches. Regarding conditional autoregression, researchers indicated that applying the standard maximum likelihood was not feasible due to an awkward normalizing term appearing in likelihood functions, and hence Besag (1974) proposed maximum pseudo-likelihood estimation in the early years. With the availability of computational power, several researchers are inclined to fit SEMs through the McMC approach (LeSage, 2000). This approach can also be used to estimate parameters in CAR models. According to Besag (1974), the joint distribution shown in Eq.2 is a (Gaussian) Markov random field (MRF). Such a MRF can be sampled from a Gibbs distribution, and hence Gibbs sampling is suitable to fit these models. However, a major drawback of this approach is the speed of computation. Calculating the probability density for CAR random effects needs to compute the determinant of its auto-covariance matrix. This process requires $O(n^3)$ operations, and computation is, in particular, expensive for a large number of spatial observations. To efficiently compute coefficients in Eq.3, we take advantage of Markovian properties and make use of INLAs.

The Markovian property implies that the parameter θ_i for the ith area is independent of all the other parameters $\boldsymbol{\theta}_{-i}$, given a set of its neighbours $\mathcal{N}(i)$,

$$\theta_i \perp\!\!\!\perp \boldsymbol{\theta}_{-i} \mid \boldsymbol{\theta}_{\mathcal{N}(i)}.$$

This ensures the sparsity of the precision matrix $\boldsymbol{Q}$, where zero elements Q_{ij} are only derived from θ_i and θ_j that are independent on the planar, forming $\theta_i \perp\!\!\!\perp \theta_j \mid \boldsymbol{\theta}_{-ij} \Leftrightarrow Q_{ij} = 0$. Additionally, zero elements are outside the diagonal and first off-diagonals, which leads to the precision matrix tridiagonal. With this setting, computation is tractable even for extremely high dimensional parameters. To be more specific, the tridiagonal precision matrix enables efficient matrix operations, such as Cholesky decomposition, for calculating the determinant of the auto-covariance matrix. For example, the probability density for that CAR random effect takes the form:

$$\pi(\boldsymbol{\theta}) \propto |\boldsymbol{Q}|^{\frac{1}{2}} \exp\left(-\frac{1}{2}\boldsymbol{\theta}^T \boldsymbol{Q} \boldsymbol{\theta}\right). \quad 5$$

The Cholesky factorization gives $\boldsymbol{Q} = \boldsymbol{L}\boldsymbol{L}^T$, where $\boldsymbol{L}$ denotes the Cholesky triangle, which remains sparse. As a result, the complexity for calculating Eq.5 decreases from $O(n^3)$ to $O(n^{3/2})$. We refer the readers to Rue and Held (2005) for a more detailed discussion.

The following task is taken by INLA (Rue et al., 2009), which is an efficient, full Bayesian inference approach based on sufficiently accurate numerical approximations. INLA derives the marginal posterior distributions of the hyperparameter $\pi(\theta_j|\boldsymbol{Y})$ and the latent variable $\pi(\boldsymbol{x}_j|\boldsymbol{Y})$ for Eq.4.

This is done by integration,

$$\pi(x_i \mid \boldsymbol{Y}) = \int \pi(x_i \mid \boldsymbol{\theta}, \boldsymbol{Y}) \pi(\boldsymbol{\theta} \mid \boldsymbol{Y}) d\theta, \quad 5$$

$$\pi(\theta_j \mid \boldsymbol{Y}) = \int \pi(\boldsymbol{\theta} \mid \boldsymbol{Y}) d\theta_{-j}. \quad 6$$

where x_i is the i^{th} latent variable. θ_j is the j^{th} hyperparameter. θ_{-j} is the complement hyperparameter vector to θ_j. INLA applies a three-step procedure to achieve the integration. The first step aims to calculate $\pi(\boldsymbol{\theta} \mid \boldsymbol{Y})$, since this term serves to compute the marginal distribution of both the hyperparameter and the latent variable. It can be approximated by,

$$\tilde{\pi}(\boldsymbol{\theta} \mid \boldsymbol{Y}) \propto \left. \frac{\pi(\boldsymbol{Y} \mid \boldsymbol{x}, \boldsymbol{\theta}) \pi(\boldsymbol{x} \mid \boldsymbol{\theta}) \pi(\boldsymbol{\theta})}{\tilde{\pi}_G(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{Y})} \right|_{x = x^*(\boldsymbol{\theta})}.$$

$\tilde{\pi}_G(\mathbf{x} \mid \theta, \mathbf{y})$ denotes the Gaussian approximation[4] to $\pi(\mathbf{x} \mid \theta, \mathbf{y})$ by mathching the mode of the full conditional of $\boldsymbol{x}$ for a given $\boldsymbol{\theta}$.

The second step aims to approximate $\pi(x_i \mid \boldsymbol{\theta}, \boldsymbol{Y})$. Rue et al. (2009) offered three possible approaches: a Gaussian approximation, a Laplace approximation and a simplified Laplace approximation. The simplified Laplace approximation, as the default option, compromises estimate accuracy and computational costs.

Once the two terms in Eq.5 are approximated, we can integrate out θ by applying numerical integration. This can be done by an iterative algorithm with respect to suitable evaluation points θ_m and corresponding area weights Δ_m. Putting all things together, we have

$$\tilde{\pi}(x_i \mid \mathbf{y}) \approx \sum_{m=1}^{M} \tilde{\pi}(x_i \mid \theta_m, \mathbf{y}) \tilde{\pi}(\theta_m \mid \mathbf{y}) \Delta_m,$$

as the final approximation of the posterior marginal density of Eq.5.

Despite the complexity of the INLA algorithm, the developers of INLA provide a feasible way for users to

[4] Tierney and Kadane (1986) proved that this equation is equivalent to the Laplace approximation of the marginal posterior distribution.

implement INLA. INLA is a C program that is further packaged into a R library. The R-INLA supports a variety of standard and custom models and can process the results directly in R.

A rethinking on Interpretation

We now move on to the interpretation of latent spatial random effects (Z). To the best of our knowledge, only a few studies shed light on interpreting latent spatial random effects. Unlike the interpretation of spatial lag models, which has been well documented in the theoretical and empirical literature, the interpretation of SEMs was rarely seen. Le Gallo et al. (2005) paved the way for its interpretation and highlighted the spatial properties of shocks. Since both CAR models and SEMs pertain to hierarchical modeling, following the interpretation of SEMs should help explain the random effects in CAR models. However, due to the omission of the structural form for CAR models, we cannot directly borrow the interpretation from SEMs. This leads us to focus on the work of Andrews (2005) and Bai (2009). They demonstrated that the latent random field for observations was used for capturing the varying response to some common shocks. Notably, the latent field in these models is built on a covariance structure that has uncorrelated diagonal elements and homogeneous variance. That is to say, the latent field does not incorporate any spatial information. By contrast, in SEMs and CAR models, spatial information is contained in the covariance matrix of the random field. High covariance between neighbouring regions is expected, while covariance between non-neighbouring regions is expected to be zero. In this way, the interpretation of a spatial random field is based on the ideas that each observation should react differently to neighbourhood shocks, but observations that are close to each other should respond similarly to those shocks due to similar unobserved characteristics.

More specifically, the latent spatial random effects are associated with random intercepts. Following this logic, we can posit that the individual-level random intercepts v_i reveal some grouping patterns. Some random intercepts have similar positive values, others have similar negative values. When these values are plotted on a graph, we may observe several clusters having similar values. Clusters with similar positive values are recognized as “hot spots”,

as the random intercept positively contributes to the composite intercept[5]. In contrast, we can identify some "cold spots".

This phenomenon should reflect the existence of spatial autocorrelation defined as the coincidence of value similarity with locational similarity. For example, geographical areas tend to be surrounded by neighbours with very similar values. This phenomenon is also a reflection of spatial heterogeneity, since the estimate value of random intercepts is not constant across space, that is a cluster of high values being distinguished from a cluster of low values (Anselin, 2010).

A simulation study

In this section, we illustrate and compare the SEM and CAR model by conducting a simulation study based on artificial data.

The artificial dataset is built on the freely available R package "spData", which gives topographic information on Syracuse, New York State. We create two uncorrelated ($\text{Pearson's } \mathrm{r} = 0.084,\ \mathrm{p} - \text{value} = 0.509$) variables ($x1, x2$), available on 63 lattices of Syracuse City. Further, the data for the dependent variable Y ($\mathrm{Y} = (y_1 \dots y_i)'$ with i being an indicator for lattices) are assumed to follow binomial distribution, and are generated as a function of the two covariates above. Onto this functional relationship, we add a spatially correlated disturbances referred to as error Ω. Ω is simulated based on the distance between data points on the surface of Syracuse City. d_{ij} denotes the Euclidean distance between the centroid of lattice i and j. Then, the spatial covariance matrix these centroids is defined as $\sigma^2 \exp\left(-\frac{d_{ij}}{\rho}\right)$. ρ is the range that is set to 2500 meters. σ^2 is the variance that is set to 1.

The data generating process is expressed as：

$$\mathrm{Y} \sim Binomial(n, p)$$

$$logit(p) = 2 + 1 \cdot \mathrm{x1} + 0.5 \cdot \mathrm{x2} + \Omega.$$

[5] The composite intercept relies on an overall intercept β_0 and individual intercepts v_i.

Table 1. The descriptive statistics for the artificial data.

Statistic	n	Mean	SD	Min	Pctl(25)	Pctl(75)	Max
Y	63	55.9	8.27	28	53.5	62	63
x1	63	0.055	0.894	-1.97	-0.485	0.635	2.17
x2	63	-0.032	0.903	-2.31	-0.645	0.485	2.19

Based on the artificial data set (See Table 1), we conduct a statistical analysis via SEMs and CAR models with the same spatial weight matrix.

- SEM: $\text{logit}(\pi_i) = x_1\beta_1 + x_2\beta_2 + \Gamma_{SAR}$ with $\Gamma_{SAR} \sim SAR(\rho_{SAR}, \tau_{SAR})$
- CAR: $\text{logit}(\pi_i) = x_1\beta_1 + x_2\beta_2 + \Gamma_{CAR}$ with $\Gamma_{CAR} \sim pCAR(\rho_{CAR}, \tau_{CAR})$

Regarding the spatial weight matrix, binary matrices based on the distance threshold are considered (see Figure 2). These matrices are then rescaled by its maximum eigenvalues (Haining, 2003). To avoid the inappropriate specification of spatial weight matrices, we consider three candidate matrices with different thresholds, which generate three scenarios. We adopt a $\text{logGamma}(1,0.00005)$ prior to the precision parameters τ for both models. A uniform prior on the interval $(-1,1)$ is then assigned to ρ. β has an independent, zero mean normal prior distribution with the precision $1/3$. Analyses are then carried out using the software package R-INLA.

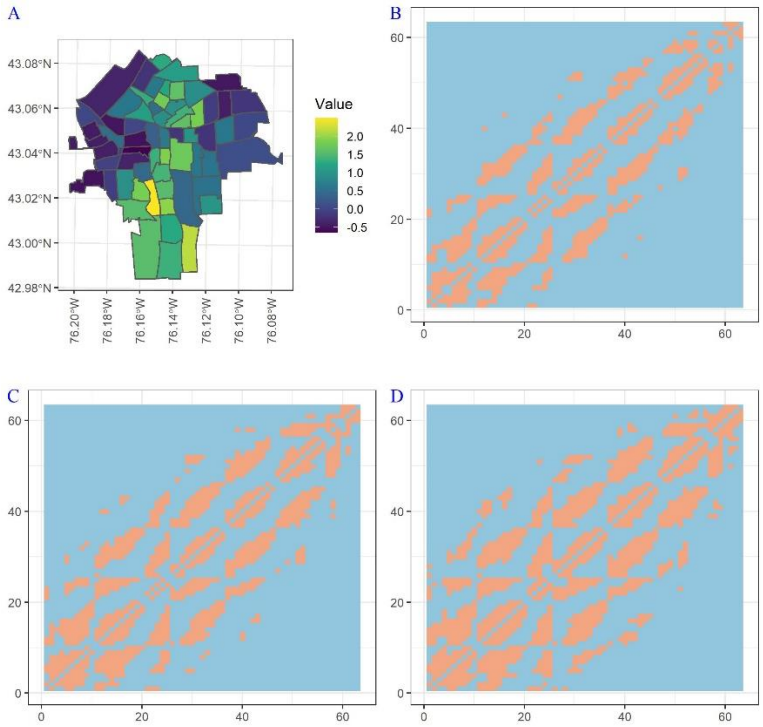


Figure 2. Spatial distribution of variables and different spatial weight matrices. (A) Map of Syracuse with census tracts and the spatial visualization of the dependent variable. Nearest neighbouring relationships between census tracts for generating spatial weight matrices. (B) The threshold for the nearest neighbour is set to 1000 meters; (C) The threshold is 1250 meters; (D) The threshold is 1500 meters.

The relative performance of the two models is assessed by a series of criteria[6], 1) the Deviance Information Criterion (DIC) (Spiegelhalter et al., 2002), the Watanabe–Akaike (or "widely applicable") information criterion (WAIC) (Watanabe, 2010) and the mean logarithmic score of CPO (LCPO) (Roos and Held, 2011) (A detailed explanation of these criteria is found in Appendix A.); these classical criteria help to select the model that fits the data well. 2) posterior distribution; since the true parameters are known, we can directly evaluate the quality of parameter estimates. The two types of assessment tools allowed us to identify the model configuration that successfully accounts for spatial autocorrelation in the data and is able to provide precise parameter estimates.

[6] According to Bakka et al. (2018), the utilization of several criteria can give a more robust result in term of model fits.

Some interesting results emerged from our analysis. Figure 3 displays the model assessment result using the classical criteria. We observe that the CAR model produces lower values in terms of the DIC and WAIC, but the LCPO value for the both models are almost identical across the three scenarios. As such, both models perform well, but the performance of the CAR model is slightly better, irrespective of the matrix type. Furthermore, the assessment criteria indicate that the CAR model with W3 performs best in most scenarios.

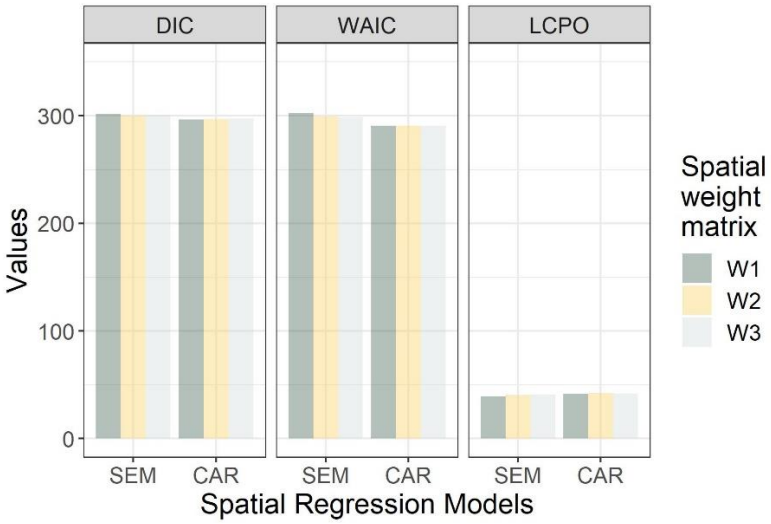


Figure 3. The comparison of SEMs and CAR models based on the classical criterion.

Based on the assessment results using the second type of criterion, both models can identify the statistical significance of independent variables (see Figure 4). Regarding x_1, the CAR model yields better estimates than the SEM, since its posterior mean estimate is close to the true value across the three scenarios. While, regarding x_2, the CAR model and SEMs produce similar posterior estimates and credible intervals. Within each model, the credible interval is fairly consistent for each scenario (with the exception of the SEM with W1), but the posterior mean estimate varies a little. For the CAR model, the deviation between the posterior mean estimate and the true value of x_1 in scenarios 1 and 2 is almost identical, but is slightly smaller than in scenario 3. For x_2, such a deviation is almost identical in scenario 1 and 2, but is slightly larger than in scenario 3. Our results suggest that the CAR model is more reliable in terms of parameter estimates, independence of the spatial weight matrix.

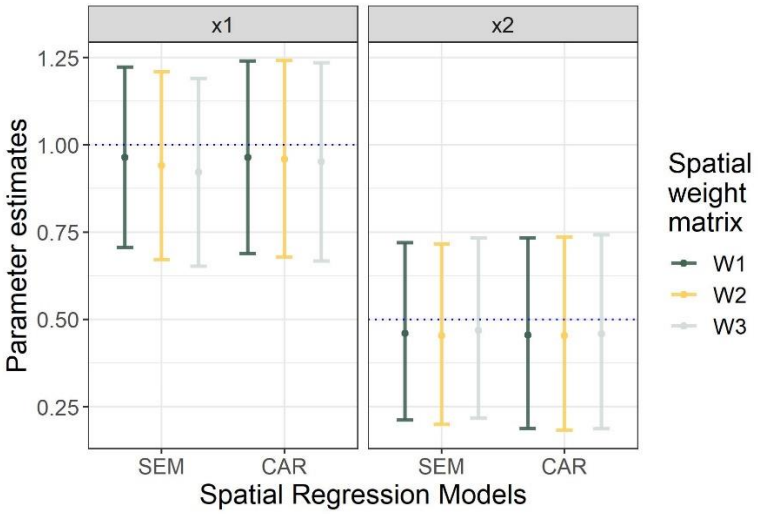


Figure 4. Parameter estimates, such as mean values (colored dots) and 95% credible intervals (vertical segments), from SEMs and CAR models with three different spatial weight matrices. The three scenarios are marked in green, yellow and cadet blue, respectively. Dotted lines indicate the true value.

Figure 5 shows the posterior distribution of spatial autocorrelation parameters and variance in each scenario. The statistical significance of ρ demonstrates that both models have potential to remove spatial autocorrelation in the data, and thus to provide precise parameter estimates. For the three scenarios, the mean estimate of the spatial parameter in the SEM is always inferior to that in the CAR model, but the two models have comparable posterior marginals of variance. According to Ver Hoef et al. (2018), we may anticipate the spatial autocorrelation parameter in the CAR model to be larger in absolute value than that in the SEM. The fact is that, using the same construction W, the precision structure for CAR component includes the term $I - \rho W$, while that for SEMs includes the product term $(I - \rho W)(I - \rho W)'$, which implies higher order dependence. In other words, the SAR component in SEMs average over more neighbours to smooth out the impact of the spatial dependence.

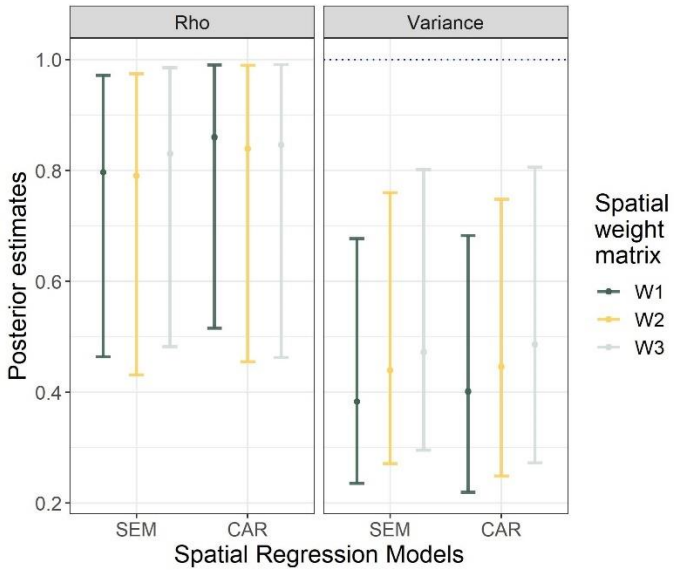


Figure 5. Posterior estimates of spatial autocorrelation parameters (the right panel) and of variance (the left panel).

We also notice that the INLA takes on average 7.6 and 4.4 seconds to fit the two models. This shows that SEMs require more computing power than CAR models. This finding also corresponds to Ver Hoef's statement (2018) that converting from simultaneous autoregression to conditional autoregression models should offer computational benefits.

The simulation study above is meant to illustrate the application of the SEM and CAR model to count data. As such, it remained a cartoon of the complexity and challenges posed by real data. In the following section, these two models are employed to investigate a real-world case.

An empirical study of second home rates

Background

Our application builds on the abovementioned models for the incidence rate of Corsican (France) second homes in 2017. Corsica is one of the 13 French administrative regions and locates in the Mediterranean Sea. Its white sand beaches in the south, towering cliffs in the east, and rocky highlands in the centre draw millions of tourists each

year. Corsica has seen a significant growth in the number of second homes, in parallel with the growing number of tourists. According to a local division of the INSEE (French National Institute of Statistics and Economic Studies), Corsica had 91,622 second houses in 2015, accounting for 37.2% of the region's housing stock. It is one of the French regions most affected by second houses. Residents are anxious about how their lives might change, since high second home rates and housing prices may raise the cost of living and lead to the loss of locally-derived revenue (Marcelin, 2018). This phenomenon has attracted the attention of local government agencies and economists.

In addition to the Corsican case, second home phenomena in different regions and countries have been examined from various aspects for decades (Müller and Hoogendoorn, 2013). Several scholars have examined the impact of public policy or planning on second home occurrences (Norris and Winston, 2009). Paris (2009) argue that second home studies should concentrate on assessing the influence of second homes on local housing markets. As more and more social scientists intend to understand how space or location translates into socioeconomic phenomena, some scholars also focus on the spatial property of second homes. Back and Marjavaara (2017) found that there was a significant difference in the spatial distribution of second house stocks between urban and rural areas, as well as between tourist destinations and the hinterlands. They also showed the impact of second homes on housing markets depending on local contexts. Mowl et al. (2020) performed an empirical study indicating the heterogenous nature of second home stocks in rural mountainous areas (Alpujarra Granadina) in Spain. In a word, it is necessary to recognize the importance of locations on second home stocks.

Our case study investigates the incidence rate of second homes in Corsica. We intend to demonstrate the role that both local characteristics and spillover play in determining the incidence rate. To achieve this, the two AR models based on local indicators are developed.

Spatial Dependence

Apart from the above-mentioned empirical studies, there is an intuitive motivation for considering spatial dependence in modeling the incidence rate. The dependence should arise due to unobserved neighbourhood shocks associated with scenery, availability of housing sale information and even the impact of zoning regulation

transmitted over space.

To be more specific, these amenity-related characteristics are not always limited within a single spatial unit, but frequently spread across adjacent units, resulting in spatial dependence among these units. Furthermore, pioneer second homes develop in a specific area with desirable features for buyers and then the phenomenon spreads out to neighbouring counties over time as the information becomes public. However, housing supply is typically constrained by zoning policies and economic profitability. That is, housing supply responds more slowly to housing demand. Due to the lack of supply and a high price, buyers should be forced to relocate to surrounding counties with comparable qualities but lower housing prices. Of course, the incidence rate in the neighbouring counties increases. We also refer to this kind of spatial dependence as the "spatial error effect"[7].

Data and models description

To model the second home incidence rate, we initially assume that $Y_i \sim \text{Binomial}(n_i, \pi_i)$, where Y_i is the number of second homes[8] in county i $(i = 1, \cdots, 360)$ in 2017, π_i is the second home rate and n_i is the total number of housing stocks. n_i also acts as an offset term in the regression, which allows the count of second homes in each county to be comparable.

A set of control variables are considered in the models to capture the relevant amenities, social-economic factors and accessibility that are expected to influence the second home occurrence within each county. All covariates are found as important in the literature.

---

[7] Another realization of spatial dependence is known as spatial spillover effects, which are generated only through observed factors of neighbourhoods, and are modelled by a spatial lag in the dependent variable. However, in our particular case, there is no convincing economic theoretical explanation for the presence of strong spatial spillover effects. Furthermore, factors directly affecting second homes in proximal counties are also hard to recognize by second home buyers. Even though some factors may show spatial dependence implicitly, it is still hard to measure, for example, the scenery. While, unobserved factors of neighbouring counties should affect second homes in a county just as observed factors do. For these reasons, spatial error effects are more suitable in our case, and we apply SEMs and CAR models.

[8] There are wide cross-county variations of second home counts, as shown by Figure B1 in Appendix B.

Several researchers have demonstrated that amenities are a key determinant of second home occurrence. Müller et al. (2004) believed second homes were often located in coastal areas, amenity-rich hinterlands, and areas with mountain landscapes. Barnett (2007) stated that an ideal second home location should include the following features: weather, infrastructure, views, history and nature. Kaltenborn et al. (2007) found that there has been a rapid expansion of second homes in sub-alpine areas in Norway, and they (2008) also indicated that some second homes were suited around mountain and coast tourism resorts, others were suited closely to historical areas. More recently, Back and Marjavaara (2017) indicated that more and more Norwegian second homes were located along coasts, around western mountain ranges and inland lakes. As such, we include the number of natural landscapes and of cultural landscapes. To facilitate interpretation, we use a logarithmic transformation on these variables. We also consider two dummy variables. They are the mountainous county and the coastal county. Positive signs are expected for these variables.

The second home occurrence also depends heavily upon local social, economic circumstances and policies (Brida et al., 2009; Hall, 2015). Some local governments passed tax incentive schemes to promote second home development, others used financial tools such as the levying of differential council taxes, tax penalties for second home owners, to control the development (Gallent and Tewdwr-Jones, 2000; Norris and Shiels, 2007). Furthermore, Barke (2007) showed that the number of second homes in a Spanish province was associated with the provincial population size, and depopulation was an important factor in creating second homes. Hence, we include two covariates, the local population and the council tax. With respect to accessibility, we include the distance to the nearest "gates", which involve all the ferry ports and airports that can connect Corsica to the French mainland. Table 2 provides the descriptive statistics for these covariates.

Table 2. Descriptive statistics for these explanatory variables.

Types	Statistic	N	Mean	SD	Min	Pctl(25)	Pctl(75)	Max	Expected signs
Environmental factors	$\log_2(\text{Natural landscapes})$	360	0.719	1.832	0	0	1	19	+
	$\log_2(\text{Cultural landscapes})$	360	0.747	2.422	0	0	1	28	+

	Mountainous county (0/1)	360	0.483		0			1	+
	Coastal county (0/1)	360	0.269		0			1	+
Social economic factors	$\log_2$(Population)	360	7.692	2.105	3.459	6.200	9.034	16.109	−
	$\log_2$(Council tax)	360	-2.460	0.525	-4.849	-2.684	-2.125	-1.206	−
Accessibility	$\log_2$(dis_gates)	360	3.915	0.928	-0.160	3.429	4.587	5.411	−

SEMs are widely used to deal with spatial error effects, and a CAR model is also considered here. As regards the spatial weight matrix in the model, a binary, relative graph neighbour structure is used (see Figure B2 in Appendix B).

Model comparison results

To begin with, we focus on model assessment, since it is helpful to identify a single best model. The classical criteria are employed again.

Table 3. DIC, WAIC, LCPO Statistics for SEMs and CAR Models.

	DIC	WAIC	LCPO
SEM	2766.55	2691.62	5.424
CAR	2762.51	2682.58	5.352

According to the DIC values in Table 3, the SEM has the worse fit (DIC=2766.55) The DIC value of the CAR model is 2762.51, suggesting the better fit and parsimony. Something similar happens with WAIC and LCPO. As a result, all of the model selection criteria undoubtedly point to the CAR model. It will be used to produce one-step

ahead, to demonstrate the interpretation of fixed covariate effects and more important latent random effects in the next section. Before doing this, we attempt to compare the posterior estimates of parameters in the two models briefly.

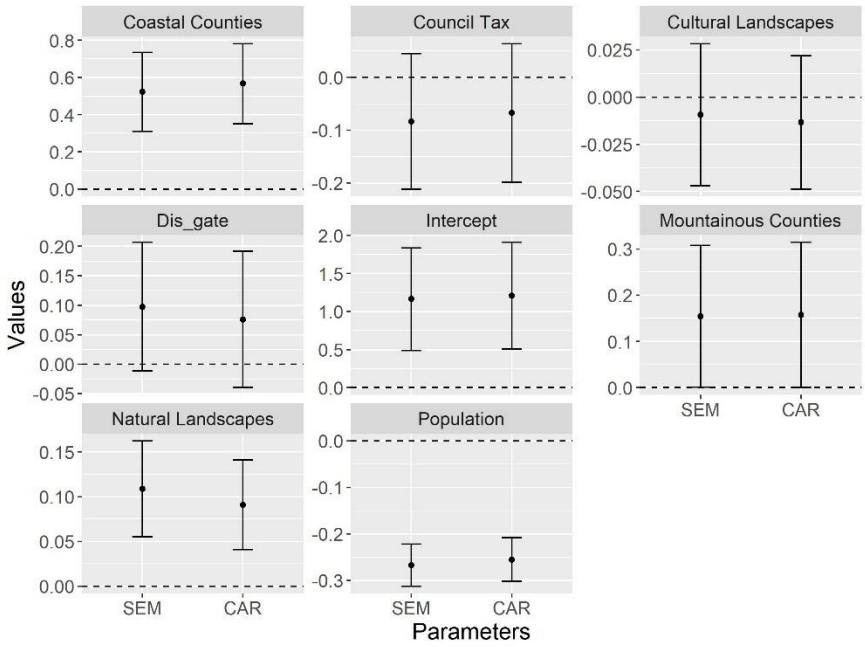


Figure 6. Posterior means (black dots) of fixed covariate effects with 95% credible intervals (vertical segments) based on the two model.

Figure 6 depicts the posterior distribution of the covariates. Except for the cultural landscape, the council tax and the distance to the nearest "gates" on the island, the majority of covariates are statistically significant. As such, the change in the specification does not change the statistical significance of covariates. Most coefficients and credible intervals are consistent between the SEM and CAR model. For example, the range of credible intervals of the coastal county and the population are likely stable, but the posterior mean for these covariates shows a slightly increased pattern as the model moves from the SEM to the CAR model. By contrast, the posterior mean for the natural landscape decreases.

The posterior estimates of hyperparameters are shown in Figure 7. The left panel reveals almost a complete overlap between the posterior marginals of variance for the CAR model and that for the SEM. The right panel indicates that

the CAR model yields higher posterior mean estimates (0.793) for the spatial autocorrelation parameter than the SEMs (0.451). As expected, these results correspond to the findings in the simulation study.

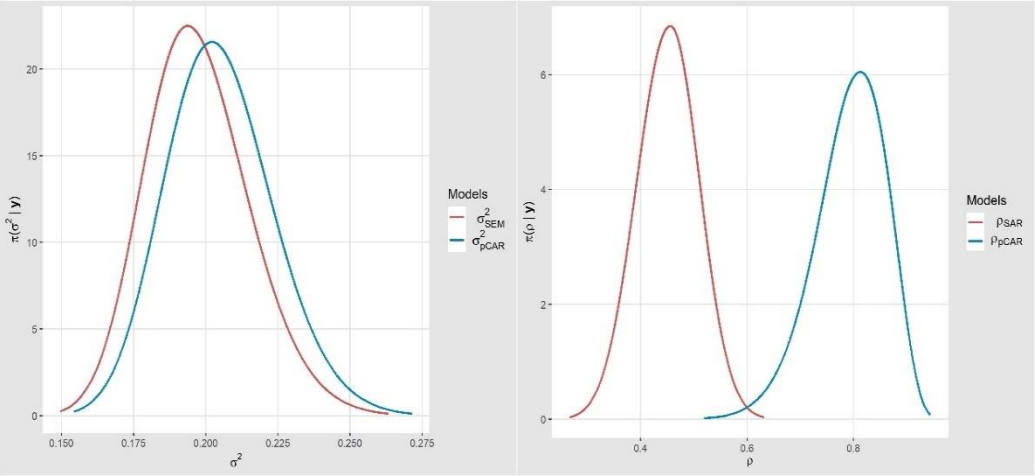


Figure 7. Posterior distribution of spatial autocorrelation parameters (the right panel) and of variance (the left panel)

Interpretation and discussion

Given that the CAR model is preferred according to its accuracy, we intend to interpret its fixed covariate effects and spatial random effects. To be more specific, once the covariate effects are taken into account, it enables us to know how location contributes to the second home incidence rate.

Table 4 displays the posterior mean for the fixed covariate effects and the corresponding 95% credible intervals.

Table 4. Posterior means, standard deviations (SD) and a 95% credible interval for all covariates.

	mean	SD	0.025quant	0.975quant
Intercept	1.211	0.357	0.506	1.910
Natural landscapes	0.091	0.025	0.041	0.141
Cultural	-0.013	0.018	-0.048	0.021

landscapes				
Mountainous county	0.158	0.080	0.0001	0.315
Coastal county	0.567	0.109	0.352	0.781
$\log_2$ Population	-0.255	0.024	-0.302	-0.209
$\log_2$ Council tax	-0.067	0.066	-0.198	0.063
$\log_2$ dis_gates	0.076	0.059	-0.040	0.191
ρ_{CAR}	0.792	0.057	0.645	0.902

For each covariate, the log odds of the second home incidence rate associated with a 1 unit or percentage increase. Only the number of cultural landscapes, the council tax and the distance to the nearest 'gates' are statistically insignificant.

We find that the relative log odds of the second home rate (π_i) increases 0.091 $(95\%CI, 0.041; 0.141)$ times with a 1-unit increase in the natural landscape count, given all else is equal. Our result conforms to many studies in the literature, where second home buyers prefer locations close to beautiful natural amenities and a peaceful living environment (Wong et al., 2017). Coastal and mountainous counties are positively associated with the second home incidence rate, and the posterior mean estimate for coastal counties (0.567 $(95\%CI, 0.352; 0.781)$) is greater than that for mountainous counties (0.158 $(95\%CI, 0.0001; 0.315)$). This indicates that second home buyers favour coastal areas.

A negative association of the rate with population size (-0.255 $(95\%CI, -0.302; -0.209)$) is obtained. Consequently, counties with high numbers of inhabitants tend to have low second home rates. This finding is consistent with Barke's results (2007). Regarding the Corsican context, most inhabitants live and work in the main urban area of the Corsica, such as Ajaccio (ID. 83), Bastia (ID. 360), and Corte (ID. 195). Housing supply in these cities is always limited. Furthermore, these cities lack high-quality natural amenities to attract second home buyers, and overcrowding can be an issue here (Milano et al., 2019). Thus, the imbalance between demand and supply restricted the development of the second homes, and low second home rates are observed.

The posterior distribution of spatial autocorrelation parameter ρ_{CAR} $(0.792, 95\%CI, 0.645; 0.902)$ reveals

significant spatial error effects and demonstrate positive spatial spillover in the error term.

The posterior mean estimates for some spatial random effects are shown in Table 5. These effects demonstrate how the log odd of county's incidence rate is influenced by neighbourhood shocks.

Table 5. Posterior means, standard deviations and a 95% credible interval for spatial random intercepts. Only selected counties are shown.

County ID	INSEE Code	County name	Mean	SD	interval
68	2B150	Lumio	1.648	0.165	(1.325 ; 1.975)
249	2A247	Porto-Vecchio	1.433	0.211	(1.019 ; 1.848)
248	2A139	Lecci	1.300	0.114	(1.076 ; 1.524)
51	2A130	Grosseto-Prugna	1.199	0.122	(0.957 ; 1.437)
33	2A276	Serra-di-Ferro	1.186	0.133	(0.925 ; 1.447)
73	2B050	Calvi	1.139	0.193	(0.757 ; 1.517)
279	2B010	Algajola	1.073	0.199	(0.685 ; 1.466)
28	2A065	Cargese	0.959	0.154	(0.657 ; 1.263)
23	2A070	Casaglione	0.914	0.125	(0.669 ; 1.16)
193	2A141	Letia	0.879	0.149	(0.591 ; 1.174)
207	2A348	Vico	0.874	0.116	(0.645 ; 1.102)
250	2A362	Zonza	0.839	0.165	(0.515 ; 1.164)
314	2B167	Montegrosso	0.717	0.141	(0.439 ; 0.992)
247	2A092	Conca	0.680	0.120	(0.445 ; 0.916)
42	2A269	Sari-Solenzara	0.607	0.135	(0.343 ; 0.872)
148	2B007	Albertacce	0.596	0.149	(0.306 ; 0.889)
78	2A228	Pietrosella	0.563	0.135	(0.299 ; 0.829)
197	2A284	Sollacaro	0.559	0.133	(0.296 ; 0.82)
175	2B023	Asco	0.553	0.161	(0.24 ; 0.873)

30	2A090	Coggia	0.516	0.120	(0.279 ; 0.752)
32	2A189	Olmeto	0.352	0.137	(0.085 ; 0.622)
83	2A004	Ajaccio	-0.626	0.246	(-1.11 ; -0.143)
117	2B309	Santa-Maria-di-Lota	-0.706	0.151	(-1.002 ; -0.411)
41	2A272	Sartène	-0.710	0.253	(-1.207 ; -0.213)
268	2A001	Afa	-0.728	0.169	(-1.062 ; -0.4)
119	2B305	San-Martino-di-Lota	-0.730	0.146	(-1.018 ; -0.445)
195	2B096	Corte	-0.852	0.244	(-1.332 ; -0.374)
355	2B353	Ville-di-Pietrabugno	-1.253	0.168	(-1.587 ; -0.925)
154	2A006	Alata	-1.433	0.146	(-1.723 ; -1.148)
360	2B033	Bastia	-2.027	0.268	(-2.555 ; -1.502)
299	2B120	Furiani	-2.290	0.162	(-2.612 ; -1.975)
218	2B037	Biguglia	-2.455	0.166	(-2.785 ; -2.132)

As previously indicated, SAR and CAR random effects can be interpreted as random intercepts. Significant “hot spots” in Corsica include Lumio (ID. 68), Porto-Vecchio (ID. 249), Lecci (ID. 248), Grosseto-Prugna (ID. 51), Calvi (ID. 73), Algajola (ID. 279), Cargese (ID. 28), Vico (ID.207), Montegrosso (ID. 314), Zonza (ID. 250), Conca (ID. 247), etc. By contrast, Alata (ID. 154), Biguglia (ID 218), Furiani (ID 299), Ville-di-Pietrabugno (ID. 355), Bastia (ID. 360) and Corte (ID. 195) belong to “cold spots”.

In Figure 8, we display the posterior mean estimates of the CAR-specified random field, which reveals some interesting spatial features.

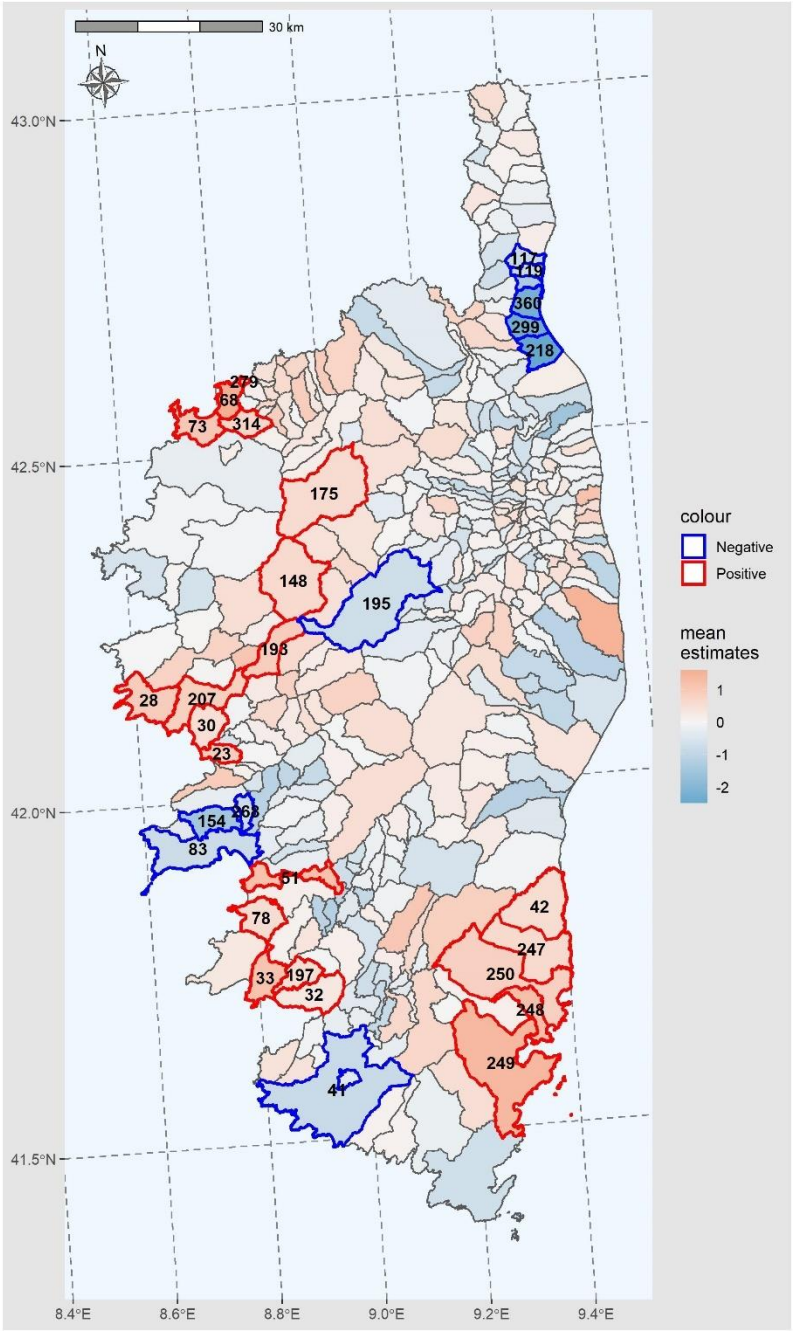


Figure 8. Posterior estimates of latent spatial random effects in LCAR model with “hot spots” and “cold spots”

Corsica has five significant clusters of "hot spots." Grosseto-Prugna (ID. 51, +1.199), Serra-di-Ferro (ID. 33, +1.186), Pietrosella (ID. 78, +0.563), Sollacaro (ID. 197, +0.559) and Olmeto (ID. 32, +0.352) form a cluster along Ajaccio's southern shore. A similar cluster of “hot spots” is identified in southeast Corsica with Porto-Vecchio (ID. 249, +1.433), Lecci (ID. 248, +1.300), Zonza (ID. 250, +0.839), Conca (ID. 247, +0.680), Sari-Solenzara

(ID. 42, $+0.607$). Vico (ID. 207, $+0.874$), Coggia (ID. 30, $+0.516$) and Casaglione (ID. 23, $+0.914$) are part of another cluster characterized by positive neighbourhood shocks around Cargese county (ID. 28, $+0.959$). The next cluster is located in the "Balagne" region, including Lumio (ID. 68, $+1.648$), Montegrosso (ID. 314, $+0.717$), Calvi (ID. 73, $+1.139$), Algajola (ID. 279, $+1.073$). Regarding the last cluster of "hot spots", the county of Asco (ID. 175, $+0.553$), Letia (ID. 193, $+0.879$) and Albertacce (ID. 148, $+0.596$) are sited at the centre of Corsica. All these suggest that benefits arise from a county for its neighbours. A possible explanation is that the five clusters are all tourist attractions. Calvi, Porto-Vecchio, Cargese and their neighbours are traditional tourist attractions. Asco is an important road junction on the GR20. Grosseto-Prugna and its neighbours are famous for beaches, seashore ponds and medieval landmarks. As such, we can expect that some unobserved factors due to scenery/recreation in a county spread over neighbouring counties.

On the other hand, we observe two clusters of "cold spots". One is in the Bastia area, another is in the Ajaccio area. For the Bastia area, negative shocks exist in counties such as Biguglia (ID. 218, -2.455), Furiani (ID. 299, -2.290), Bastia (ID. 360, -2.027), Ville-di-Pietrabugno (ID. 355, -1.253), Santa-Maria-di-Lota (ID. 117, -0.706), San-Martino-di-Lota (ID. 119, -0.730). The Ajaccio area involves Ajaccio (ID. 83, -0.626), Alata (ID. 154, -1.433) and Afa (ID. 268, -0.728). The two clusters correspond to the two major urban areas in Corsica. Further, ferry ports, airports and highways together make the two clusters the main hinge of the island, rather than the attraction due to its scenery. In addition, supply rigidities in housing may also limit the interest of home buyers.

Conclusion remarks

In this article, we reviewed the two common spatial "AR" models, known as SEMs and CAR models, to model spatial effects among data with discrete support in a space. The first model is often regarded as spatial econometrics, while the latter one is widely used in spatial statistics. To our knowledge, some research established the mathematical relationships and investigated the marginal correlation properties between SEM and CAR models, but no study investigates them from the hierarchical modeling perspective. Furthermore, some advanced Bayesian techniques, such as INLA, are rarely found in the economic literature to fit these models. The interpretation of the

two models is rarely discussed either.

Even though the two models are quite different according to econometricians, we argue that these two models can be unified through hierarchical models for several count data families (Poisson, binomial, negative binomial), while the fundamental difference between the two models is the covariance structure for specifying latent (Gaussian) random fields. Moreover, due to the separation of data models (i.e., likelihood) and process models (i.e., latent fields) in the hierarchical structure, we can avoid the complications arising in fitting SEMs using maximum likelihood methods. Lastly, since both models can be represented by a hierarchical model with Gaussian random effects, we can apply the INLA approach.

A simulation study is then carried out to compare these models. We find that CAR models often have better performance. The posterior mean value and credible intervals are quite consistent across both models in each scenario. Furthermore, the CAR model yields the higher posterior mean value of spatial autocorrelation parameter, which corresponds to the statement of Ver Hoef et al. (2018). This simulation study also demonstrates that INLAs can return accurate estimates with limited computing resources.

To provide an example for the two types of spatial autoregressive models, we used Corsican county-level data related to second homes, sceneries and social-economic indicators for the year of 2017. More importantly, our attention is paid to how the model estimates should be interpreted. In particular, how to interpret the latent random field. We therefore link the two types of models to latent factor model in econometrics and interpret the spatial latent field as neighbourhood shocks. Since the random field component is associated with random intercepts in a GLMM, we can determine the grouping pattern on the basis of posterior estimates of random intercepts. Subsequently, such group patterns can be plotted on the graph to identify several geographical “hot spots” with positive values and “cold spots” with negative values. We identify five “hot spots” and two “cold spots” in Corsica. We believe that these spots should be determined by some latent neighbourhood factor, such as the scenery.

Due to the flexibility of the INLA approach, it can be used to fit spatial, spatio-temporal and panel data models. We hope that this study triggers other research on applying the INLA approach to fit spatial econometric models when investigating regional economic issues.

There are several useful ways in which the current work can be extended. First, due to the availability of data, we include several social-economic indicators in the current study. Future work may address this limitation by

considering more social-economic covariates, such as poverty and unemployment rate at the county level. For practical use, it would be beneficial to have a suite of diagnostic metrics for evaluating the two types of models. Various metrics have been offered in the literature, and practitioners might benefit from a systematic effort to gather and implement them. Lastly, the binomial scenario can be extended to other count data families in order to examine regional foreign direct investment and international trade.

Reference

Andrews, D. (2005) 'Cross-Section Regression with Common Shocks', *Econometrica*, 73(5), pp. 1551–1585.

Anselin, L. (1988) Spatial Econometrics: Methods and Models, Operational Regional Science Series. Springer Netherlands.

Anselin, L. (2010) 'Thirty years of spatial econometrics', *Papers in Regional Science*, 89(1), pp. 3–25.

Back, A. and Marjavaara, R. (2017) 'Mapping an invisible population: the uneven geography of second-home tourism', *Tourism Geographies*, 19(4), pp. 595–611.

Bai, J. (2009) 'Panel Data Models With Interactive Fixed Effects', *Econometrica*, 77(4), pp. 1229–1279.

Baltagi, B. H., Egger, P. and Pfaffermayr, M. (2008) 'Estimating regional trade agreement effects on FDI in an interdependent world', *Journal of Econometrics*, 145(1), pp. 194–208.

Barke, M. (2007) 'Second Homes in Spain: An analysis of change at the provincial level, 1981-2001', *Geography*, pp. 195–207.

Barnett, R. (2007) 'Central and Eastern Europe: Real estate development within the second and holiday home markets', *Journal of Retail & Leisure Property*, 6(2), pp. 137–142.

Beguin, J. *et al.* (2012) 'Hierarchical analysis of spatially autocorrelated ecological data using integrated nested Laplace approximation', *Methods in Ecology and Evolution*, 3(5), pp. 921–929.

Besag, J. (1974) 'Spatial interaction and the statistical analysis of lattice systems (with discussion)', *Journal of the Royal Statistical Society Series B*, 36, pp. 192–236.

Besag, J., York, J. and Mollié, A. (1991) 'Bayesian image restoration, with two applications in spatial statistics', *Annals of the Institute of Statistical Mathematics*, 43(1), pp. 1–20.

Brida, J. G., Osti, L. and Santifaller, E. (2009) 'Second homes and the need for policy planning', *Tourismos: an international multidisciplinary journal of tourism*, 6(1), pp. 141–163.

Brueckner, J. K. (2003) 'Strategic Interaction Among Governments: An Overview of Empirical Studies', *International Regional Science Review*, 26(2), pp. 175–188.

Cressie, N. and Chan, N. H. (1989) 'Spatial modeling of regional variables', *Journal of the American Statistical Association*, 84(406), pp. 393–401.

Dubin, R., Pace, R. K. and Thibodeau, T. G. (1999) 'Spatial Autoregression Techniques for Real Estate Data', *Journal of Real Estate Literature*, 7(1), pp. 79–96.

Elhorst, J. P. (2010) 'Applied Spatial Econometrics: Raising the Bar', *Spatial Economic Analysis*, 5(1), pp. 9–28.

Ertur, C., Le Gallo, J. and Baumont, C. (2006) 'The European Regional Convergence Process, 1980-1995: Do Spatial Regimes and Spatial Dependence Matter?', *International Regional Science Review*, 29(1), pp. 3–34.

Gallent, N. and Tewdwr-Jones, M. (2000) Rural Second Homes in Europe: Examining housing supply and planning control. doi: 10.4324/9781315201979.

Le Gallo, J. *et al.* (2005) 'On the property of diffusion in the spatial error model', *Applied Economics Letters*, 12(9), pp. 533–536.

Le Gallo, J. and Ertur, C. (2003) 'Exploratory spatial data analysis of the distribution of regional per capita GDP in Europe, 1980−1995', *Papers in Regional Science*, 82(2), pp. 175–201.

Griffith, D. A. and Paelinck, J. H. P. (2007) 'An equation by any other name is still the same: on spatial econometrics and spatial statistics', *The Annals of Regional Science*, 41(1), pp. 209–227.

Haining, R. (2003) *Spatial data analysis: theory and practice*. Cambridge university press.

Hall, C. M. (2015) 'Second homes planning, policy and governance', *Journal of Policy Research in Tourism, Leisure and Events*, 7(1), pp. 1–14.

Ver Hoef, J. M. *et al.* (2018) 'Spatial autoregressive models for statistical inference from ecological data', *Ecological Monographs*, 88(1), pp. 36–59.

Kaltenborn, B. P., Andersen, O. and Nellemann, C. (2007) 'Second home development in the Norwegian mountains: Is it outgrowing the planning capability?', *The International Journal of Biodiversity Science and Management*, 3(1), pp. 1–11.

Kaltenborn, B. rn P. *et al.* (2008) 'Resident attitudes towards mountain second-home tourism development in Norway: The effects of environmental attitudes', *Journal of Sustainable Tourism*, 16(6), pp. 664–680.

Kauermann, G., Haupt, H. and Kaufmann, N. (2012) 'A hitchhiker's view on spatial statistics and spatial econometrics for lattice data', *Statistical Modelling*, 12(5), pp. 419–440.

Lee, D. (2011) 'A comparison of conditional autoregressive models used in Bayesian disease mapping', *Spatial and Spatio-temporal Epidemiology*, 2(2), pp. 79–89.

LeSage, J. (2000) ‘Bayesian Estimation of Limited Dependent Variable Spatial Autoregressive Models’, *Geographical Analysis*, 32(1), pp. 19–35.

LeSage, J. and Pace, R. K. (2009) *Introduction to spatial econometrics*. Chapman and Hall/CRC.

Mcculloch, C. E. and Neuhaus, J. M. (2014) ‘Generalized Linear Mixed Models’, *Wiley StatsRef: Statistics Reference Online*. (Major Reference Works). doi: https://doi.org/10.1002/9781118445112.stat07540.

Milano, C., Novelli, M. and Cheer, J. M. (2019) ‘Overtourism and Tourismphobia: A Journey Through Four Decades of Tourism Development, Planning and Local Concerns’, *Tourism Planning & Development*, 16(4), pp. 353–357.

Moraga, P. (2019) Geospatial health data: Modeling and visualization with R-INLA and shiny. CRC Press.

Mowl, G., Barke, M. and King, H. (2020) ‘Exploring the heterogeneity of second homes and the “residual” category’, *Journal of Rural Studies*, 79, pp. 74–87.

Müller, D. K., Hall, C. M. and Keen, D. (2004) ‘Second home tourism impact, planning and management’, in *Tourism, mobility and second homes: Between elite landscape and common ground*. Channel View Publications, pp. 15–32.

Müller, D. K. and Hoogendoorn, G. (2013) ‘Second homes: Curse or blessing? A review 36 years later’, *Scandinavian Journal of Hospitality and Tourism*, 13(4), pp. 353–369.

Norris, M. and Shiels, P. (2007) ‘Housing affordability in the Republic of Ireland: is planning part of the problem or part of the solution?’, *Housing Studies*, 22(1), pp. 45–62.

Norris, M. and Winston, N. (2009) ‘Rising second home numbers in rural Ireland: Distribution, drivers and implications’, *European Planning Studies*, 17(9), pp. 1303–1322.

De Oliveira, V. (2012) ‘Bayesian analysis of conditional autoregressive models’, *Annals of the Institute of Statistical Mathematics*, 64(1), pp. 107–133.

Paris, C. (2009) ‘Re-positioning second homes within housing studies: Household investment, gentrification, multiple residence, mobility and Hyper-consumption’, *Housing, Theory and Society*, 26(4), pp. 292–310.

Rönnegård, L., Shen, X. and Alam, M. (2010) ‘hglm: A package for fitting hierarchical generalized linear models’, *The R Journal*, 2(2), pp. 20–28.

Roos, M. and Held, L. (2011) ‘Sensitivity analysis in Bayesian generalized linear mixed models for binary data’, *Bayesian Analysis*, 6(2), pp. 259–278.

Rue, H. and Held, L. (2005) Gaussian Markov random fields : theory and applications. Chapman & Hall/CRC.

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

Spiegelhalter, D. J. *et al.* (2002) 'Bayesian measures of model complexity and fit', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4), pp. 583–639.

Tierney, L. and Kadane, J. B. (1986) 'Accurate approximations for posterior moments and marginal densities', *Journal of the american statistical association*, 81(393), pp. 82–86.

VerHoef, J., Hanks, E. and Hooten, M. (2018) 'On the relationship between conditional (CAR) and simultaneous (SAR) autoregressive models', *Spatial Statistics*, 25, pp. 68–85.

Wall, M. M. (2004) 'A close look at the spatial structure implied by the CAR and SAR models', *Journal of statistical planning and inference*, 121(2), pp. 311–324.

Watanabe, S. (2010) 'Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory', *Journal of Machine Learning Research*, 11(Dec), pp. 3571–3594.

Whittle, P. (1954) 'On stationary processes in the plane', *Biometrika*, pp. 434–449.

Wikle, C. K., Zammit-Mangion, A. and Cressie, N. (2019) *Spatio-temporal Statistics with R*. Chapman and Hall/CRC.

Wong, B. K. M., Musa, G. and Taha, A. Z. (2017) 'Malaysia my second home: The influence of push and pull motivations on satisfaction', *Tourism Management*, 61, pp. 394–410.

Wood, S. N. (2011) 'Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(1), pp. 3–36.