Modelling the presence of disease under spatial misalignment using Bayesian latent Gaussian models

Xavier Barber,[1] David Conesa,[2] Silvia Lladosa,[2] Antonio López-Quílez[2]

[1]Operational Research Centre, Miguel Hernández de Elche University, Elche; [2]Department of Statistics and Operational Research, University of Valencia, Valencia, Spain

Abstract

Modelling patterns of the spatial incidence of diseases using local environmental factors has been a growing problem in the last few years. Geostatistical models have become popular lately because they allow estimating and predicting the underlying disease risk and relating it with possible risk factors. Our approach to these models is based on the fact that the presence/absence of a disease can be expressed with a hierarchical Bayesian spatial model that incorporates the information provided by the geographical and environmental characteristics of the region of interest. Nevertheless, our main interest here is to tackle the misalignment problem arising when information about possible covariates are partially (or totally) different than those of the observed locations and those in which we want to predict. As a result, we present two different models depending on the fact that there is uncertainty on the covariates or not. In both cases, Bayesian inference on the parameters and prediction of presence/absence in new locations are made by considering the model as a latent Gaussian model, which allows the use of the integrated nested Laplace approximation. In particular, the spatial effect is implemented with the stochastic partial differential equation approach. The methodology is evaluated on the presence of the *Fasciola hepatica* in Galicia, a North-West region of Spain.

Correspondence: David Conesa, Department of Statistics and Operational Research, Faculty of Mathematics, University of Valencia, C/ Dr. Moliner 50, 46100 Burjassot, Valencia, Spain.
Tel: +34.963544362 - Fax: +34.963543238.
E-mail: conesa@uv.es

Key words: Bayesian Kriging; *Fasciola hepatica;* Geostatistics; INLA; Hierarchical Bayesian modelling.

Acknowledgements: this paper was mainly written while Xavier Barber was visiting the department of Statistics and Operational Research at the University of Valencia. Xavier Barber, David Conesa and Antonio Lopez-Quilez would like to thank the Ministerio de Economía y Competitividad (the Spanish Ministry of Economy and Finance) via research grant MTM2013-42323-P (jointly financed with the European Regional Development Fund). Authors would also like to thank the INLA-project group of researchers (in particular, Elias Kranski) for their prompt support with technical aspects in the usage of INLA and the SPDE module. Finally, authors would like to thank Marta Gonzalez-Warleta and Mercedes Mezo from the Centro de Investigaciones Agrarias de Mabegondo-INGACAL, Xunta de Galicia, for providing the dataset analysed and answering questions about the process for gathering data. Two anonymous referees must also be thanked for providing interesting comments that have resulted in a very improved version of this paper.

Received for publication: 24 September 2015.
Accepted for publication: 24 September 2015.

Geospatial Health 2016; 11:415
doi:10.4081/gh.2016.415

Introduction

Starting with the pioneering work by John Snow, who mapped out the spread of a cholera outbreak in London more than 150 years ago (Snow, 1857), researchers have been trying to identify disease causes by relating spatial disease patterns to geographic variation in health risks. The way to do so consists on building models that translate spatial data on health outcomes and possible related covariates (such as environmental, socio-economic, behavioral or demographic factors), into epidemiologically meaningful results. This presents, however, several methodological challenges arising from the fact that data can be aggregated at different scales.

Although individual humans would ideally represent the basic unit of spatial analysis in health research, publicly available data are often aggregated to a sufficient extent to prevent the disclosure or reconstruction of patient identity (Goovaerts, 2008). This kind of spatial data, usually know as areal or lattice data (Cressie, 1993), require methods that directly utilise the spatial setting and assume positive spatial correlation between observations, essentially borrowing more information from neighboring areas than from areas far away and smoothing local rates toward local, neighboring values.

In our case, we will focus in another kind of spatial data, usually known as geostatistical or point-referenced data, that come from those situations where the interest is to produce a continuous risk surface starting from point data [see Banerjee *et al.* (2014), Cressie (1993), Diggle and Ribeiro (2007) and references cited therein for a more detailed explanation about geostatistical data].

Geostatistical health data require methods that allow relating the disease data with potential related covariates (such as those above mentioned) by quantifying the spatial dependence. Nevertheless, one of the main interests in Geostatistics relies on predicting about the underlying process on those non-observed locations. When dealing with diseases, the interest is to create maps of disease prevalence. Kriging, so named by Matheron (1963) in honor of Krige's work

(Krige, 1951), is maybe the most known geostatistical technique. Basically, kriging takes into account the existing underlying spatial structure between observations to predict attribute values at unsampled locations using information related to one or several attributes. Kriging has been extensively used in disease epidemiology (including public health, plant pathology and veterinary). A very small sample of examples of its use include the mapping of influenza in France (Carrat and Valleron, 1992); the mapping of rotavirus in the United States; kriging of malaria risk (Kleinschmidt *et al.*, 2000); kriging of incidence rates of rare diseases in England (Webster *et al.*, 1994); the mapping of the Hepatitis B in China (Zhong *et al.*, 2005); the mapping of spatial patterns of infant mortality and birth defect rates in Iowa (Rushtong and Lolonis, 1996); the maps of many plant disease epidemics such as Phytophthora (Larkin *et al.*, 1995) and African cassava mosaic virus (Lecoustre *et al.*, 1989); and the map of rinderpest in Central and Southern Somalia (Ortiz-Pelaez *et al.*, 2010).

A usual extension to kriging arises when one is interested in including the effect of possible covariates in the modelling or either to apply it to situations in which the stochastic variation in the data is known to be non-Gaussian. The resulting models are usually named generalised linear geostatistical models used by Diggle *et al.* (1998) and further described in Diggle and Ribeiro (2007) under the generic term of model-based geostatistics. geoR and geoRglm (Ribeiro *et al.*, 2003) are two packages of the well-known statistical software R (R Core Team, 2015) that can be used to perform model-based geostatistical data-analysis, while Diggle *et al.* (2002) is a good example of the application of this approach in childhood malaria in the Gambia.

The combination of non-Gaussian data, the linear predictor and an unobserved latent variable usually makes estimation and prediction computationally difficult. Bayesian inference turns out to be a good option to deal with spatial hierarchical models analysis because it allows both the observed data and model parameters to be random variables (Banerjee *et al.*, 2014), resulting in a more realistic and accurate estimation of uncertainty [see for instance Haining *et al.* (2007), as an example of the advantages over conventional – non-Bayesian – modelling approaches]. Another advantage of the Bayesian approach is the ease with which prior information can be incorporated. Note that prior information can usually be very helpful in discriminating spatial autocorrelative effects from ordinary non-spatial linear effects (Gaudard *et al.*, 1999). This kind of approach is also known as Bayesian kriging (Handcock and Stein, 1993). Bayesian geostatistical models have been largely applied in the mapping of malaria (Craig *et al.*, 2007; Gemperli *et al.*, 2004; Gosoniu *et al.*, 2012; Haining *et al.*, 2007; Hay *et al.*, 2009; Kazembe *et al.*, 2006; Raso *et al.*, 2012; Schur *et al.*, 2011; Wardrop *et al.*, 2010), and neglected tropical diseases (Batchelor *et al.*, 2009; Clements *et al.*, 2009, 2010; González-Warleta *et al.*, 2013; Raso *et al.*, 2005; Schur *et al.*, 2011; Wardrop *et al.*, 2010); but also in veterinary epidemiology (Biggeri *et al.*, 2006).

Until recently, Markov Chain Monte Carlo (MCMC) algorithms (Gilks *et al.*, 1996) have been the most common method for making Bayesian statistical inference with generalised linear geostatistical models. Nevertheless, we use the integrated nested Laplace approximation (INLA) methodology (Rue *et al.*, 2009) and software (http://www.r-inla.org). This choice is mainly based on the speed of calculation and the ease with which model comparison can be performed (Rue *et al.*, 2009). Moreover, as geostatistical models are continuously indexed Gaussian Fields, we use the stochastic partial differential equation (SPDE) approach (Lindgren *et al.*, 2011) to deal with them.

Bayesian geostatistical analysis using INLA have been already applied for mapping diseases. Along with introducing the package geostatsinla for performing geostatistics with INLA in an easy way, Brown (2015) applies it in the context of mapping the Loa loa filariasis disease [a dataset previously cited in Diggle and Ribeiro (2007)]. Moreover, Karagiannis-Voules *et al.* (2013) have used Bayesian geostatistical negative binomial models to analyse reported incidence data of cutaneous and visceral leishmaniasis in Brazil covering a 10-year period, while González-Warleta *et al.* (2013) have also used Bayesian geostatistical binomial models to predict the probability of infection of paramphistomosis in Galicia (NW Spain).

The most basic format of geostatistical health data is composed by the spatial location, the measurement values about the disease prevalence observed at the location and the corresponding measurement values of the covariates of interest at the same location. Nevertheless, commonly the measurement values about the disease prevalence are not observed at the same locations that the covariates of interest. This problem, usually named misalignment, must be taken into account and so, it must be incorporated in the modelling, because it is important to note that uncertainty about the covariates could influence on the predictions. For the sake of simplicity, in many analyses, the values of the covariates of interest are previously predicted at the locations where the disease has been observed by using geostatistical methods, such as kriging. But this procedure clearly does not take into account the uncertainty of those predicted values. There have been many approaches to tackle the misalignment problem (Foster *et al.*, 2012; Gelfand, 2010; Haining, 2004; Miller *et al.*, 2007; Waller and Gotway, 2004). In our case, we use the INLA-SPDE module which allows us to incorporate the uncertainty by building a spatial model for the covariate and another for the response variable and then doing the estimation process jointly (Krainski and Lindgren, 2014).

Our interest here is twofold. One the one hand, we aim to show how to perform Bayesian geostatistical analysis for mapping diseases both taking and not taking into account the uncertainty of the covariates using INLA. But, on the other hand we also want to show how results can be influenced by not taking into account the misalignment problem. In particular, we present a Bayesian geostatistical analysis of the bovine fasciolosis in Galicia using information about the environmental and spatial features of each location. Finally, we would like to mention that both approaches here presented could also be employed in different settings with other diseases in order to improve knowledge about the spatial targeting of prevention and control.

Materials and Methods

In what follows, we firstly present the dataset used in this work. The remainder of the Section is devoted to present how to perform a model-based geostatistical approach for analysing the presence/absence of a disease using a Bayesian approach (in particular, the Integrated Nested Laplace approximation), firstly without taking into account spatial misalignment and then taking it into account.

A dataset for explaining bovine fasciolosis in Galicia

Fasciola hepatica is a parasitic trematode that infects a wide variety of domestic and wild mammals worldwide, including humans [see, for instance, Martínez-Valladares *et al.* (2013) for a description of the veterinary epidemiology of the disease]. The life-cycle of fascioliosis is complex. It involves a final host (where the adult worm lives), an intermediate host, mainly freshwater snails of the genus *Lymnaea* (where the larval stages of the worm develop) and a carrier (aquatic plants that will be ingested by animals). The fasciolosis infection depends to a great extent on the biological life cycle and is strictly linked to the envi-

ronmental and geographical conditions of the area where transmission occurs. It is an important parasitic disease of farm livestock, and a major cause of economic losses in farms due to decreased productivity and quality of milk and meat products. Clearly, a good knowledge of the spatial distribution of the disease could help animal health instances to increase awareness in high-risk areas and better target their monitoring and control efforts. Thus, it results a good example for showing the behavior of hierarchical Bayesian models in the context of predicting the occurrence of diseases, the plus being that data from the potential covariates were not observed at the same locations where the disease was observed (giving us also the possibility of describing the misalignment issue).

Data analysed in this work come from a study conducted during 2008 aiming to explore basic aspects of the epidemiology of fascioliasis in Galicia (NW Spain), the main cattle producing region in Spain. Galicia occupies a surface area of 29,575 km^2, administratively divided into 315 municipalities with very different cattle farming activity and stocking rate per surface unit. According to the 2008 livestock census, there were 339,530 dairy cows (99% on farms in the northern half of the region), and 221,917 beef cows (on farms spread over a larger area extending to the south-east of the region). Grasslands occupy approximately 60% of the useful agricultural land and cows usually graze throughout the year, mainly on beef cattle farms. The type of livestock husbandry and the climatic characteristics of the region favour grazing-linked transmission of helminthosis.

A slaughterhouse that processes cattle from the whole region was visited fortnightly during 2008. At each visit, 20 adult cattle slaughtered (over 2 years old) were selected at random to determine the existence of infections within the liver, as well as the occurrence of trematode eggs in the faeces. Species identification was carried out by conventional microscopy and subsequent confirmation by molecular techniques.

In total, 192 beef cows (each one from a different farm located across the region) were selected at random in the slaughterhouse for examination the presence of *Fasciola hepatica*. Figure 1 shows the observed presence/absence of the disease. Other covariates were analysed in the study. Nevertheless, taking into account that our main interest in here is to show the different results obtained when incorporating covariates with uncertainty, we will only use as covariate the annual mean temperature. This was calculated from the data recorded at the 67 official weather stations in Galicia during the period 2004-2008 (www.meteogalicia.es).

Modelling without incorporating misalignment

Our interest here is to present a methodology to produce maps of prevalence of diseases. As previously mentioned in the Introduction, the disease prevalence can be considered as a real-valued spatially continuous process, and so, geostatistical data can be used for produce the prevalence maps. The usual format for this kind of data is composed by the spatial location, the response variable (either the presence or absence of the disease, or the amount of people with the disease at particular locations, each one representing a finite area) and information gathered about the possible effect of some covariates. In our case, we will focus mainly in a Bernoulli response (presence or absence of the disease), although the following approach would also valid for situations with a Poisson response variable.

Taking into account that our interest is to analyse Non-Gaussian data and to include the effect of covariates, we have to deal within the model-based geostatistics approach (Diggle and Ribeiro, 2007). In particular, we model the presence/absence of diseases with a hierarchical spatial model by incorporating both the environmental and geographical features of each location and demographic characteristics of each observation, the final aim being to create maps of predicted probabilities of presence in unsampled areas. As previously mentioned, Bayesian inference turns out to be a good option to deal with this kind of models because it allows both the observed data and model parameters to be random variables, resulting in a more realistic and accurate estimation of uncertainty.

Until recently, Markov Chain Monte Carlo (MCMC) algorithms (Gilks *et al.*, 1996) has been the most common (and nearly the only) method for making Bayesian statistical inference with generalised linear geostatistical models. The R packages spbayes (Finley *et al.*, 2007) and the previously mentioned geoRglm (Christensen and Ribeiro, 2002) can be used to perform this approach, although WinBUGS (Lunn *et al.*, 2000) has been probably the most used software to perform Bayesian analysis. Nevertheless, we use the integrated nested Laplace approximation (INLA) methodology (Rue *et al.*, 2009) and software (http://www.r-inla.org) as an alternative to MCMC methods, the main reason being the speed of calculation. While MCMC simulations require much more time to run, and performing prediction has become practically unfeasible when a lot of locations are available, INLA produces almost immediately accurate approximations to posterior distributions even in complex models. Another advantage of this approach is its generality, which makes it possible to perform Bayesian analysis in a straightforward way and to compute model comparison criteria and various predictive measures so that models can be compared easily (Rue *et al.*, 2009).

In spite of its wide acceptance and its good behavior in many Latent Gaussian models, an additional development is needed to implement geostatistical models within INLA. The underlying reason is that in order to be computationally efficient and stable INLA works with latent Gaussian models admitting conditional independence, that is, latent Gaussian Markov random fields with a sparse precision matrix (Rue and Held, 2005), while geostatistical models are continuously indexed Gaussian Fields. Lindgren *et al.* (2011) have proposed an alternative approach by using an approximate stochastic weak solution to a stochastic partial differential equation (SPDE) as a Gaussian Markov random field approximation to continuous Gaussian Fields with Matern covariance structure.

Assuming that the probability of finding the disease is related to its prevalence, the presence/absence can be modelled by using a point-referenced spatial hierarchical model. Specifically, if represents presence (1) or absence (0) at location i ($i=1,\ldots,n$), the full model can be stated as follows:

$$\begin{aligned}
Y_i \mid \pi_i &\overset{iid}{\sim} Ber(\pi_i), i = 1,\ldots,n \\
\text{logit}(\pi_i) &= \beta_0 + X_i\boldsymbol{\beta} + \theta_i \\
p(\beta_0) &\propto 1 \\
\beta_i &\sim \text{N}\left(0,0.1\right) \\
\theta_i &\sim \text{N}\left(\mathbf{0},\text{Q}\left(\kappa,\tau\right)\right) \\
2\log\kappa &\sim \text{N}\left(m_\kappa,p_\kappa\right) \\
\log\tau &\sim \text{N}\left(m_\tau,p_\tau\right)
\end{aligned} \qquad \text{eq. 1}$$

where π_i is the probability of occurrence at location i, $X_i\beta$ is the linear predictor for observation i and θ represents the spatially structured random effect. This modelling is based on Lindgren *et al.* (2011), who proposed an approach that avoids the computational issues arising when using INLA with continuously indexed Gaussian Fields. It is worth noting that when using this approach the correlation function is

not modelled directly. Instead, the Gaussian field θ is found numerically as a (weak) solution of a stochastic partial differential equation (the one that relates the continuous Gaussian field with Matérn covariance structure as a Gaussian Markov random field), and it depends on two parameters k and t which determine the range of the effect and the total variance, respectively. More precisely, the range is approximately $\phi \approx \sqrt{8}/\kappa$ while the variance is $\sigma_W^2 = 1/\left(4\pi\kappa^2\tau^2\right)$. Note also, that in the model we have specified the prior distributions for the parameters by setting a flat improper prior on the intercept (the default in INLA), and independent zero-mean Gaussian priors with a fixed vague precision (0.1). The priors for and are specified over the reparameterisations log and 2logk as independent Gaussian distributions. We also use the default values for their parameters. Specifically, m_k is chosen automatically such that the range of the field is about 20% of the diameter of the region, while m_t is chosen so that the corresponding variance of the field is 1. For more information about the practical implementation of the SPDE approach in R see Krainski and Lindgren (2015).

Once the model is determined, the next step is estimate its parameters, but more importantly to make prediction in unsampled locations. INLA performs both the inference and prediction simultaneously. To do so, we need to construct a lattice over the unsampled locations enabling us to get a point estimation probability of occurrence. In contrast to ordinary kriging, an irregular grid is used in prediction process. The INLA-SPDE module allows us to create a Delaunay triangulation (Hjelle and Dæhlen, 2006) around the sampled points in the region. Observations are treated as initial vertices for the triangulation, and extra vertices are added heuristically to minimise the number of triangles needed to cover the region subject to the triangulation constraints. These extra vertices are used as prediction locations. This partition is usually called *mesh* (see Figure 2 for an example of this kind of triangulation in the data analysed in this work), and a good reason for its use instead of a regular lattice is that it is denser in regions where there are more observations and consequently brings more information. Another advantage is that it saves computing time, because prediction locations are typically much lower in number than those in a regular grid.

As mentioned above, along with the inferential results about the parameters, INLA-SPDE module can be used simultaneously to perform prediction in unobserved locations (considering the prediction locations as points where the response is missing), which constitutes the real interest in this problem. The basic idea is to deal with the disease presence at a new location as a random variable with a certain probability of *success* and to calculate a point estimation of this probability, and even its full predictive density. Once the prediction is performed in the selected location, there are additional functions that linearly interpolate the results within each triangle into a finer regular grid. As a result of the process, a faceted surface prediction is obtained which approximates to the true predictive surface.

Modelling under uncertainty in the covariates

In order to perform the above predictive procedure at unobserved locations, the measurement values of the covariates should be known both at the observed locations of the response variable and at those locations where we are going to make predictions. Nevertheless, as Banerjee *et al.* (2014) stand *with the explosion in spatial data collection, it is increasingly common to find that different spatial data layers are collected at different scales*. In the particular case of disease prevalence, information about possible covariates of interest usually come

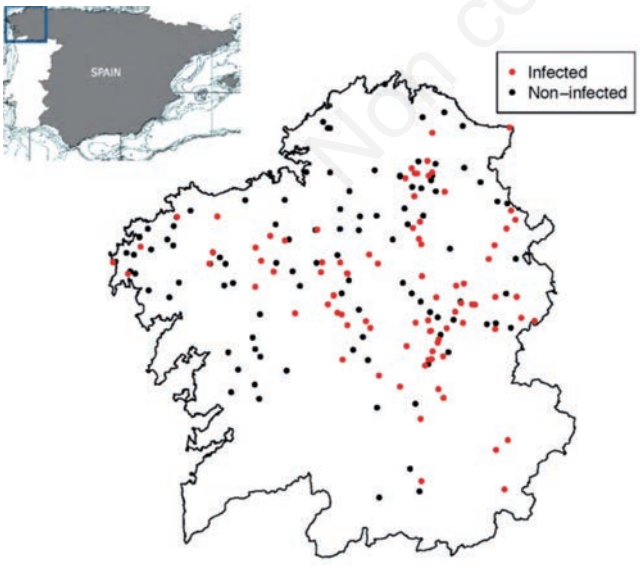

Figure 1. Sampling locations for the occurrence of *Fasciola hepatica* in Galicia (North-West of Spain), where red dots are infected locations, while black dots are non-infected locations.

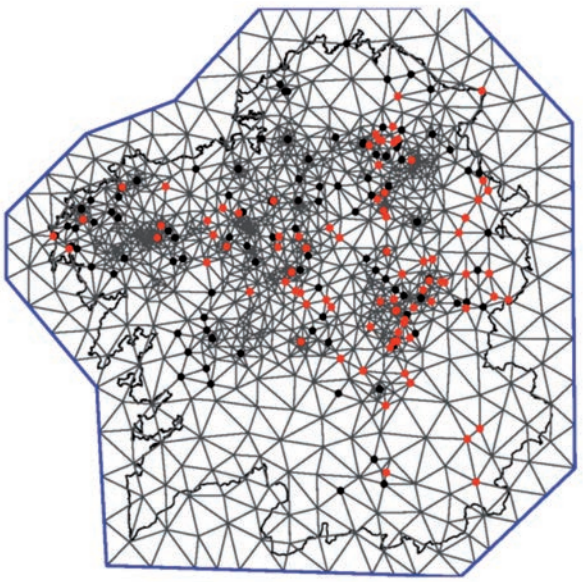

Figure 2. Delaunay triangulation (*mesh*) of Galicia with the presence (red dots) and absence (black dots) of the parasites in beef cows. Each *mesh* vertex is either an observed point or a prediction point.

from other studies, and so, locations where we have information about covariates are partially (or totally) different than those of the response and those in which we want to predict. This problem, usually named misalignment, has gained a lot of attention in the literature [see Gotway and Young (2002), Gelfand (2010) and Chapter 7 of Banerjee *et al.* (2014) and references cited therein for very good and detailed descriptions of the problem], the most important reason behind being that not taking it into account could clearly influence results. Gryparis *et al.* (2009) and Wannemuehler *et al.* (2009) are good examples of epidemiological studies presenting misalignment in which the usual models cannot be applied.The naïve solution to this problem would be to predict the value of the covariates at those locations on which we want to predict the occurrence of the disease by using geostatistical methods (for instance, using kriging), and then, plug-in these values in the prediction process, using them instead of the *true* (but unknown) values. This two-stage analysis is used in preference to forming a joint geostatistical model for the covariates and the response variable. But, as Foster *et al.* (2012) stands *it is not immediately clear what effect ignoring this extra level of variation will have on the validity of the ecological models*. Among the solutions proposed to overcome this problem, Miller *et al.* (2007) propose to predict those *true* (but unknown) values using nearest neighbor interpolation. In the same line, Gryparis *et al.* (2007) use semi-parametric smoothing to solve the issue. Other approaches for the treatment of unknown information avoiding the naïve solution above mentioned are Waller and Gotway (2004), Haining (2004) or Pfeiffer and Robinson (2008).

Nevertheless, all these approaches are based on finding ways to approximate the values of the predicted values, but not to try to include the uncertainty of the covariates in the statistical model. But, as Stein (1999) stands measurement errors are unavoidable, and so, a good approximation for the problem would be to obtain optimal prediction by using similar methodologies than those used for the non-spatial missing data problems (Buonaccorsi, 2010). In this line, Foster *et al.* (2012) include the extra variability in the statistical model by specifying a Berkson error model instead of classical measurement error (ME) models (Carroll *et al.*, 2006) and compare it with the commonly performed analysis. Szpiro and Paciorek (2012) is an example of an epidemiological study presenting misalignment analysed specifying a Berkson error model. Bayesian methodology provides a flexible framework to account for ME, because expert knowledge can be incorporated in the prior. Muff *et al.* (2014) show how the most common approaches to adjust for ME (the classical and Berkson ME) can be reformulated as latent Gaussian fields, allowing them to use the INLA approach to perform Bayesian inference on them.

In our case, we use another approach to tackle the misalignment problem, the one presented in Chapter 6 of the SPDE Manual (Krainski and Lindgren, 2014), which is based on the assumption (which turns out into a restriction) that there is only one covariate influencing the response variable and the covariate has spatial dependency. It is worth noting that in many cases this assumption is easily achieved as many covariates have indeed this spatial dependency, the big restriction being that usually we have more than one covariate. Based on this, we can build a spatial model for the covariate and another spatial model for the response variable, and then perform the estimation and prediction processes jointly using the INLA-SPDE module.

In particular, if Y_i represents the presence (1) or absence (0) at location i (i=1,…,n), the joint model for the spatial covariate and the response variable can be stated as follows:

Modelling Response			Modellig Covariate			
$Y_i \mid \pi_i$	$\overset{iid}{\sim}$	$Ber(\pi_i),\ i=1,\ldots,n$				
$\text{logit}(\pi_i)$	$=$	$\beta_0 + X_i\beta + \theta_i$				
$p(\beta_0)$	$\propto$	1				
β_i	$\sim$	$\text{N}(0,0.1)$	X_i	$\overset{iid}{\sim}$	$\text{N}\left(\phi_i, \sigma_x^2 = 0\right)$	
θ_i	$\sim$	$\text{N}\left(\mathbf{0}, Q(\kappa,\tau)\right)$	ϕ	$\sim$	$\text{N}\left(\mathbf{0}, Q(\gamma,\delta)\right)$	
$2\log\kappa$	$\sim$	$\text{N}\left(m_\kappa, p_\kappa^2\right)$	$2\log\gamma$	$\sim$	$\text{N}\left(m_\gamma, q_\gamma^2\right)$	
$\log\tau$	$\sim$	$\text{N}\left(m_\tau, p_\tau^2\right)$	$\log\delta$	$\sim$	$\text{N}\left(m_\delta, q_\delta^2\right)$	eq. 2

where p_i is the probability of occurrence at location i, X is the covariate of interest with spatial dependency being $b\beta$ its corresponding parameter, and q and ϕf are the corresponding Gaussian Markov random fields with sparse precision structure associated with the *mesh* [as in the previous subsection, this modelling is based on Lindgren *et al.* (2011)] respectively for the response variable and the covariate. Note that we choose X_i to be equal to f_i, that is, we consider that the covariate is a realisation of the random field. Note also, that in the model we have again specified the default prior distributions for the intercept and βb_1. As in the previous subsection, priors for k, t, $g\gamma$ and d are specified over the reparameterisations log$t$, 2log$k$, log$g$ and 2log$d$ as independent Gaussian distributions with the default values for their parameters. For more information about the practical implementation of this approach in R see Krainski and Lindgren (2014).

As above mentioned, once the model is determined, the next step is estimate its parameters and to make prediction in unsampled locations and INLA performs both simultaneously. Again, the idea could be to construct a *mesh* and perform the prediction in its vertices. Another option is to do a prediction in a 100x100 grid as in Cameletti *et al.* (2012). Along with the inferential results about the parameters, with INLA-SPDE module we can calculate a point estimation of the probability of presence at each location, and even its full predictive density, and then obtain a faceted surface prediction, which is the final aim of our work.

Results

In this section, we present the results obtained when applying both methodologies (not taking and taking the misalignment into account) to model the occurrence pattern of the *Fasciola hepatica* in the Galician livestock by examining the data collected above introduced.

Obtaining probability maps of presence of fasciolosis not taking into account the misalignment

As previously mentioned the final model selected for fitting to the data of *Fasciola hepatica* infection in beef cows is the one that includes the annual mean temperature as covariate, and a stochastic spatial component that accounts for the residual spatial autocorrelation.

Note that this is a typical example of misalignment, as information about both climatic variables included in the study (annual mean temperature and total rainfall) has been recorded at the 67 official weather stations in Galicia during the period 2004-2008 (www.meteogalicia.es), but the locations of the farms are not the same that the weather stations, neither the locations where we want to predict, specifically, each vertex of the *mesh.* Figure 3 shows this situation.

A first (naïve) approach would consist in using the average annual values obtained at the 67 official weather stations to make a kriging

grid (1 km^2 grid), and then interpolate the value of the mean temperature at the farms and at the prediction locations (the vertices of the *mesh*). Figure 4 represents the annual mean temperature in Galicia obtained doing so. Clearly, using the values obtained at Figure 4 to do prediction does not take into account the uncertainty about them. In order to learn about the behaviour of this (naïve) approach (and so, to be able to compare with a better approach), we present in Table 1 the numerical summary of the final model obtained using INLA, in particular the mean, standard deviation, median and 95% credible interval of the two parameters (intercept and the coefficient of the covariate). It is worth noting that the annual mean temperature has a negative effect on the response variable, with a 0.81 posterior probability of being lower than zero. The remaining term of the model is the spatial component, which can be observed at Figure 5. This component shows a strong effect, with positive values in the southern half of province of Lugo and the eastern part of Ourense, and values around zero in the rest of study region.Finally, Figure 6 shows the mean predicted distribution of the probability of infection of beef cows across Galicia and the variability of this estimation with the first and third quartiles of the posterior distribution of occurrence. It can be seen that the highest values of the probability of occurrence of fasciolosis infection covers the southeast area, in a similar way to the spatial effect. But note that this map combines both the covariate effect (in Figure 4 we can appreciate that lower temperatures are located in the eastern part of Galicia) and the spatial component, with a marked influence of the latter.

Incorporating the uncertainty

In what follows we present the results obtained when instead of perform a kriging for the mean temperature, we build a spatial model for the mean temperature and another spatial model for the presence of the fasciolosis, and then perform the estimation and prediction processes jointly using the INLA-SPDE module. As previously mentioned, in order to apply this methodology, the covariate must have a spatial dependency, something we can clearly assume for the mean temperature. The other restriction is that only one covariate can be included in the model, as it is our case in this example.

Following the ideas and code presented in the INLA-SPDE manual, we construct an observation matrix that extracts the values of the spatial field at the measurement locations. We then build a joint model including the two likelihoods, one for the response variable and another for the covariate, including respectively information about the presence and about the mean temperature. As the mean temperature is Gaussian and the presence/absence of the bacteria is clearly binomial, we need to specify the identity and logit links for each distribution when elicitating the likelihood. Table 2, and Figures 7 and 8 contain respectively the numerical summary of the parameters of the model, the spatial component (its mean and standard deviation) and the pre-

Table 1. Numerical summary of the posterior distributions of the fixed effects for the infection of beef cows with *Fasciola hepatica*.

	Mean	SD	*Q*0.025	*Q*0.5	*Q*0.975
Intercept	2.89	2.64	-2.46	2.96	7.88
Temperature	-0.34	0.22	-0.77	-0.34	0.11

SD, standard deviation.

Table 2. Numerical summary of the posterior distributions of the fixed effects for the infection of beef cows with *Fasciola hepatica* taking into account misalignment.

	Mean	SD	*Q*0.025	*Q*0.5	*Q*0.975
Intercept	-2.88	6.65	-17.70	-3.14	12.92
Temperature	0.15	0.15	-0.14	0.16	0.43

SD, standard deviation.

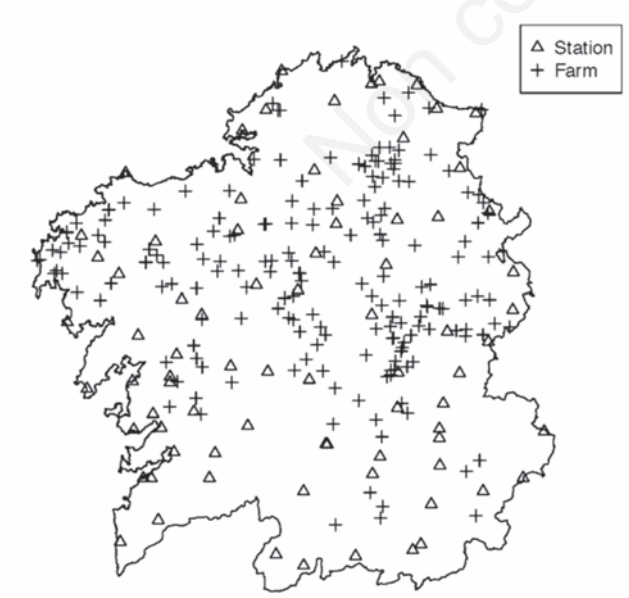

Figure 3. Misalignment problem: the sixty-seven official weather stations in Galicia do not coincide with the farms where data were observed.

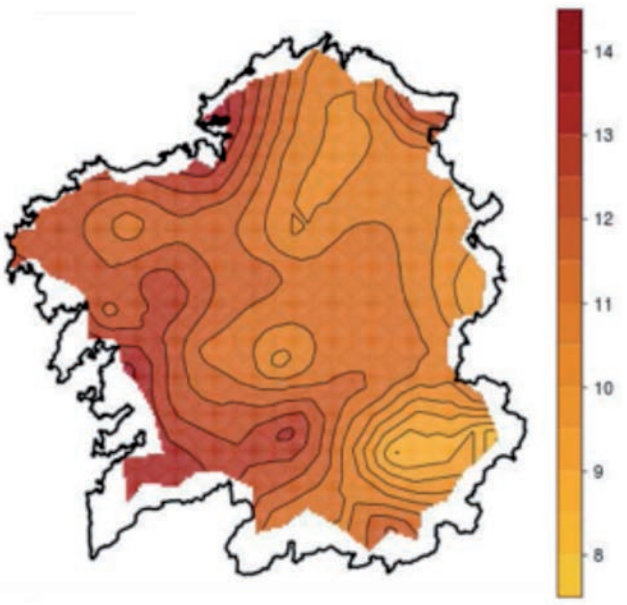

Figure 4. Annual mean temperature in Galicia. It was calculated from the data recorded at the sixty-seven official weather stations in Galicia during the period 2004-2008.

dicted distribution (expressed with the median, and the first and third quartiles) of the probability of infection of beef cows across Galicia. As it can be seen there are some differences with the results in the previous subsection. The annual mean temperature has a positive effect on the response variable (there is a 0.76 posterior probability of the corresponding coefficient being greater than zero). Although this is the opposite as in the previous subsection it is worth to note that the spatial component shows a stronger effect, giving less importance to the covariate effect. This behaviour is mainly due to the fact that we have less information about the temperature, *i.e.*, there is a lot of uncertainty about the temperature in the places of prediction. But more importantly, this is caused by the strong correlation between the covariate and the underlying spatial structure, which gives more importance to the effect of the spatial effect. Interestingly, Figure 8 shows a more reliable situation (in terms of the presence/absence observed) about the probability of occurrence of fasciolosis infection. The prediction is a combination of the covariate effect and the spatial component, but now it seems that both are expressing the opposite and so their effects are cancelled.

Discussion

The main advantage of using the Integrated nested Laplace approximation in order to perform inference and prediction is the computational ease in model fit and prediction compared to classical geostatistical methods. In classical geostatistical applications, the full range of

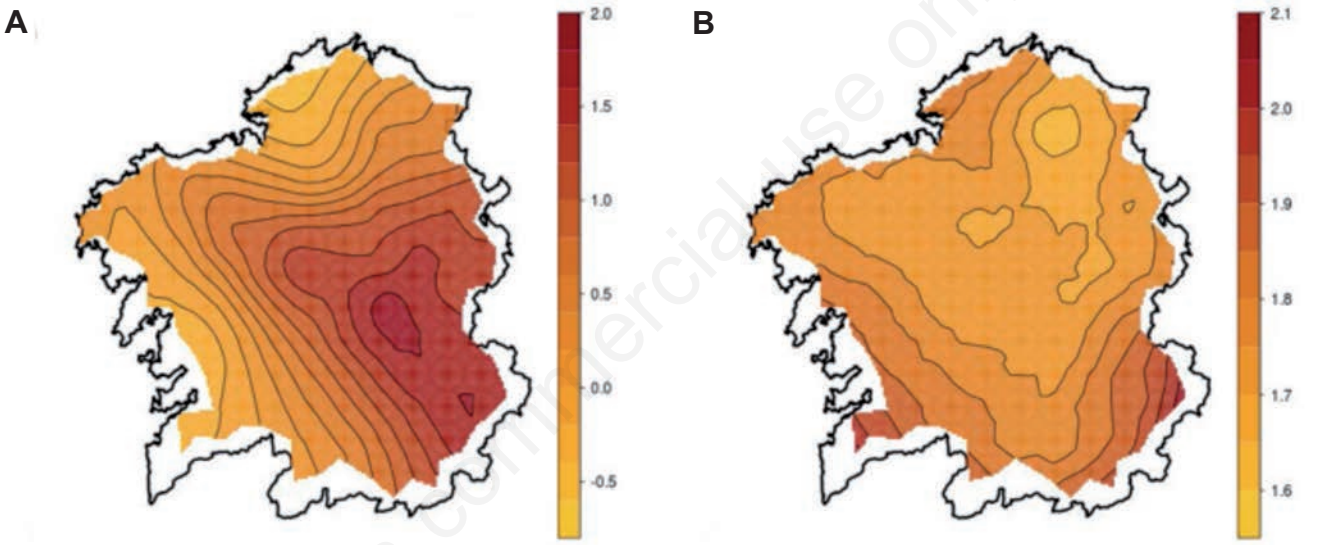

Figure 5. Posterior mean (A) and the standard deviation (B) of the spatial effect.

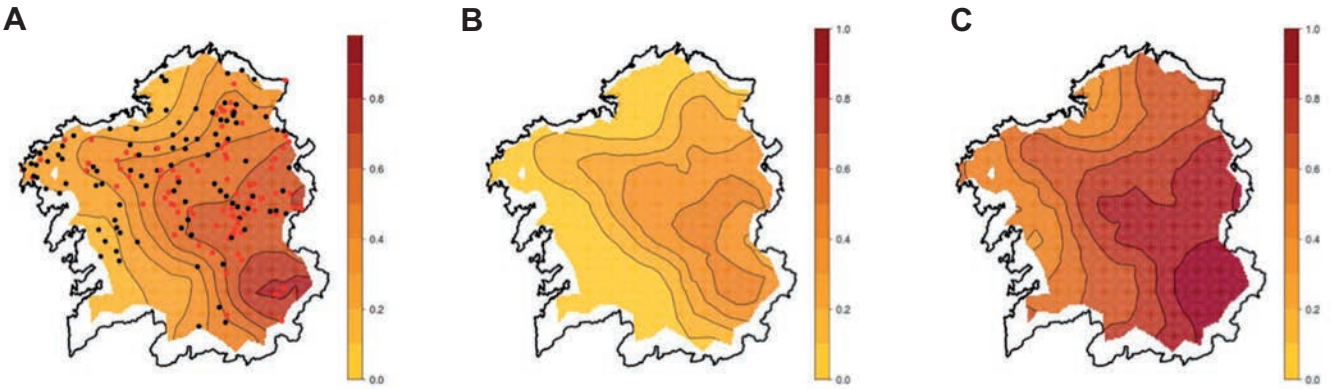

Figure 6. Mean (A), and the first (B) and third (C) quantiles of the posterior distribution of the probability of occurrence. Red dots are infected locations, while black dots are non-infected locations.

uncertainties that are always associated with species distribution models is not correctly measured, as many parameters that are considered to be *known* are actually estimated through the statistical model (Diggle and Ribeiro, 2007), a potential cause of optimistic assessments of predictive accuracy. Using the approach here presented, parameter uncertainty can be incorporated into the prediction process. But not only that, we can also incorporate the uncertainty about the covariates involved in the model by means of the Stochastic Partial Differential Equation approach (Lindgren *et al.*, 2011), which provides an explicit link between Gaussian Fields and Gaussian Markov Random Fields. Thanks to the R-INLA library, the SPDE approach can be easily implemented providing results in reasonable computing time (in contrast to MCMC algorithms).

Extensions to this modelling arise in two lines of research. On the one hand, one could include into the analysis the possibility that the sample design (which results in the observed locations) could be stochastically dependent of the process, which generates the measurements. This type of sampling is usually named preferential sampling (Diggle *et al.*, 2010) and could cause a big influence on the results if not taken into account. On the second hand, a natural extension would be to expand this modelling to the spatiotemporal domain by incorporating an extra term for the temporal effect, using parametric or semi-

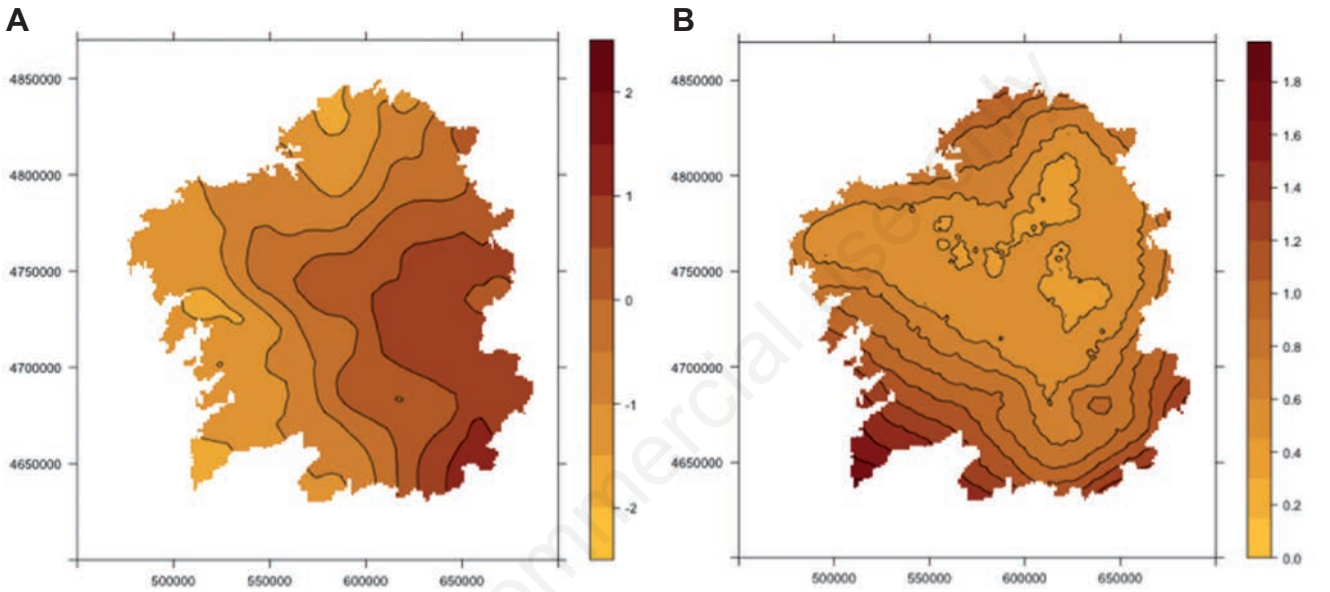

Figure 7. Posterior mean (A) and the standard deviation (B) of the spatial effect taking into account the misalignment.

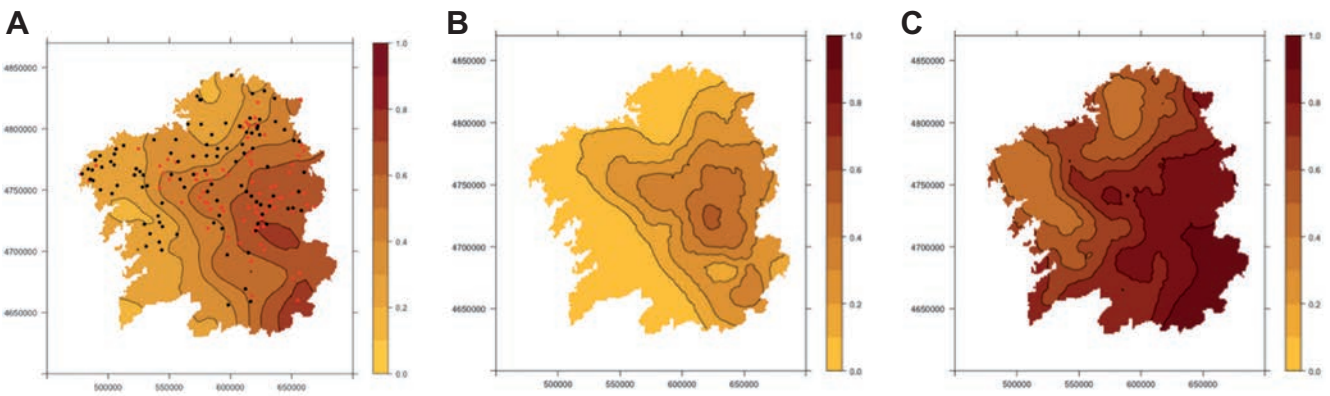

Figure 8. Mean (A), and the first (B) and third (C) quantiles of the posterior distribution of the probability of occurrence taking into account the misalignment. Red dots are infected locations, while black dots are non-infected locations.

parametric constructions to reflect linear, nonlinear, autoregressive or more complex behaviors [see for instance Blangiardo *et al.* (2013) for examples of how to include these effects]. Nevertheless, in our case, the information available did not include a reasonable enough number of years for performing any temporal analyses.

Conclusions

To conclude, we would like to mention that our results clearly show that the misalignment problem is an important issue that we must incorporate in our analysis. Results obtained when not taking into account can be misleading. In our case, results could indicate that there is an important region of Galicia with a high prevalence of fasciolosis, while this could be an artefact of the uncertainty about the mean temperature. Results of the modelling incorporating uncertainty about the mean temperature show that the high prevalence areas are in closer places but not exactly the same ones. Of course, there is still room for improvement in our conclusions, and experts should analyse results with care in order to check about possible reasons underneath.Finally, we would also like to mention that the analytical approach we used here to document the spatial patterns in the prevalence of diseases can be applied similarly in many fields of research like Environmental science, hydrogeology, mining, remote sensing, *etc*. Muñoz *et al.* (2013) and Pennino *et al.* (2013, 2014) are examples of this approach in the fisheries context.

References

Banerjee S, Carlin BP, Gelfand AE, 2014. Hierarchical modeling and analysis for spatial data. CRC-Press, Boca Raton, FL, USA.

Batchelor NA, Atkinson PM, Gething PW, Picozzi K, Fèvre EM, Kakembo ASL, Welburn SC, 2009. Spatial predictions of Rhodesian human African trypanosomiasis (Sleeping Sickness) prevalence in Kaberamaido and Dokolo, two newly affected districts of Uganda. PLoS Neglect Trop D 3:e563.

Biggeri A, Dreassi E, Catelan D, Rinaldi L, Lagazio C, Cringoli G, 2006. Disease mapping in veterinary epidemiology: a Bayesian geostatistical approach. Stat Methods Med Res 15:337-52.

Blangiardo M, Cameletti M, Baio G, Rue H, 2013. Spatial and spatio-temporal models with R-INLA. Spat Spatiotemporal Epidemiol 7:39-55.

Brown PE, 2015. Model-based geostatistics the easy way. J Stat Softw 63:1-24.

Buonaccorsi JP, 2010. Measurement error. Models, methods and applications. CRC-Press, Boca Raton, FL, USA.

Cameletti M, Lindgren F, Simpson D, Rue H, 2012. Spatio-temporal modeling of particulate matter concentration through the SPDE approach. Adv Stat Anal 97:109-31.

Carrat F, Valleron AJ, 1992. Epidemiologic mapping using the "kriging" method: application to an influenza-like epidemic in France. American J Epidemiol 135:1293-300.

Carroll RJ, Ruppert D, Stefanski LA, Crainiceanu C, 2006. Measurement error in nonlinear models: a modern perspective. Chapman and Hall/CRC Press, oca Raton, FL, USA.

Christensen OF, Ribeiro PJ, 2002. GeoRglm-a package for generalised linear spatial models. R News 2:26-8.

Clements ACA, Deville MA, Ndayishimiye O, Brooker S, Fenwick A, 2010. Spatial co-distribution of neglected tropical diseases in the East African great lakes region: revisiting the justification for integrated control. Trop Med Int Health 15:198-207.

Clements ACA, Firth S, Dembelé R, Garba A, Touré S, Al E, 2009. Use of Bayesian geostatistical prediction to estimate local variations in *Schistosoma haematobium* infection in western Africa. B World Health Organ 87:921-9.

Craig MH, Sharp BL, Mabaso MLH, Kleinschmidt I, 2007. Developing a spatial-statistical model and map of historical malaria prevalence in Botswana using a staged variable selection procedure. Int J Health Geogr 6:44.

Cressie N, 1993. Statistics for spatial data. Wiley, Koboke, NJ, USA.

Diggle PJ, Menezes R, Su T, 2010. Geostatistical inference under preferential sampling. J Roy Stat Soc C-App 52:191-232.

Diggle PJ, Moyeed R, Rowlingson B, Thomson M, 2002. Childhood malaria in the Gambia: a case-study in model-based geostatistics. J Roy Stat Soc C-App 51:493-506.

Diggle PJ, Ribeiro PJ, 2007. Model based geostatistics. Springer, New York, NY, USA.

Diggle PJ, Tawn JA, Moyeed RA, 1998. Model-based geostatistics. J Roy Stat Soc C-App 47:299-350.

Finley AO, Banerjee S, Carlin BP, 2007. SpBayes: an R package for univariate and multivariate hierarchical point-referenced spatial models. J Stat Softw 19:1-24.

Foster S, Shimadzu H, Darnell R, 2012. Uncertainty in spatially predicted covariates: is it ignorable? Appl Stat 61:637-52.

Gaudard M, Karson M, Linder E, Sinha D, 1999. Bayesian spatial prediction. Environ Ecol Stat 6:147-71.

Gelfand AE, 2010. Handbook of spatial statistics. CRC press, Boca Raton, FL, USA.

Gemperli A, Vounatsou P, Kleinschmidt I, Bagayoko M, Lengeler C, Smith T, 2004. Spatial patterns of infant mortality in Mali: the effect of malaria endemicity. Am J Epidemiol 159:64-72.

Gilks WR, Richardson S, Spiegelhalter DJ, 1996. Introducing Markov chain Monte Carlo, Markov chain Monte Carlo in practice. Springer, New York, NY, USA.

González-Warleta M, Lladosa S, Castro-Hermida JA, Martínez-Ibeas AM, Conesa D, Muñoz F, López-Quílez A, Manga-González Y, Mezo M, 2013. Bovine paramphistomosis in Galicia (Spain): prevalence, intensity, aetiology and geospatial distribution of the infection. Vet Parasitol 191:252-63.

Goovaerts P, 2008. Geostatistical analysis of health data: state-of-the-art and perspectives. In: Soares A, Pereira MJ, Dimitrakopoulos R, eds. geoENV VI. Geostatistics for environmental applications. Springer-Verlag, New York, NY, USA, pp 3-22.

Gosoniu L, Msengwa A, Lengeler C, Vounatsou P, 2012. Spatially explicit burden estimates of malaria in Tanzania: Bayesian geostatistical modeling of the malaria indicator survey data. PLoS One 7:e23966.

Gotway CA, Young LJ, 2002. Combining incompatible spatial data. J Am Stat Assoc 97:632-48.

Gryparis A, Coull BA, Schwartz J, Suh HH, 2007. Semiparametric latent variable regression models for spatiotemporal modelling of mobile source particles in the greater Boston area. J Roy Stat Soc C-App 56:183-209.

Gryparis A, Paciorek CJ, Zeka A, Schwartz J, Coull BA, 2009. Measurement error caused by spatial misalignment in environmental epidemiology. Biostatistics 10:258-74.

Haining R, 2004. Spatial data analysis. Theory and practice. Cambridge University Press, Cambridge, UK.

Haining R, Law J, Maheswaran R, Pearson T, Brindley P, 2007. Bayesian modelling of environmental risk: example using a small area ecological study of coronary heart disease mortality in relation to modelled outdoor nitrogen oxide levels. Stoch Env Res Risk A 21:501-9.

Handcock MS, Stein ML, 1993. A Bayesian analysis of kriging. Technometrics 35:403-10.
Hay SI, Guerra CA, Gething PW, Patil AP, Tatem AJ, Noor AM, Kabaria CW, Manh BH, Elyazar IRF, Brooker S, 2009. A world malaria map: *Plasmodium falciparum* endemicity in 2007. PLoS Med 6:e1000048.
Hjelle Ø, Dæhlen M, 2006. Triangulations and applications. Springer, New York, NY, USA.
Karagiannis-Voules D-A, Scholte RGC, Guimaraes LH, Utzinger J, Vounatsou P, 2013. Bayesian geostatistical modeling of leishmaniasis incidence in Brazil. PLoS Neglect Trop D 7:e2213.
Kazembe L, Kleinschmidt I, Holtz T, Sharp B, 2006. Spatial analysis and mapping of malaria risk in Malawi using point-referenced prevalence of infection data. Int J Health Geogr 5:41.
Kleinschmidt I, Bagayoko M, Clarke GPY, Craig M, Le Sueur D, 2000. A spatial statistical approach to malaria mapping. Int J Epidemiol 29:355-61.
Krainski ET, Lindgren F, 2014. The R-INLA tutorial: SPDE models. Norwegian University of Science and Technology, Trondheim, Norway.
Krainski ET, Lindgren F, 2015. The R-INLA tutorial: SPDE models. Norwegian University of Science and Technology, Trondheim, Norway.
Krige DG, 1951. A statistical approach to some basic mine valuation problems on the Witwatersrand. J Chem Metal Min Soc 52:119-39.
Larkin RP, Gumpertz ML, Ristaino JB, 1995. Geostatistical analysis of Phytophthora epidemic development in commercial bell pepper fields. Phytopathology 85:191-202.
Lecoustre R, Fargette D, Fauquet C, De Reffye P, 1989. Analysis and mapping of the spatial spread of African cassava mosaic virus using geostatistics and the kriging technique. Phytopathology 79:913-20.
Lindgren F, Rue H, Lindström J, 2011. An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. J Roy Stat Soc B 73:423-98.
Lunn DJ, Thomas A, Best N, Spiegelhalter D, 2000. WinBUGS. A Bayesian modelling framework: concepts, structure, and extensibility. Stat Comput 10:325-37.
Martínez-Valladares M, Robles-Pérez D, Martínez-Pérez JM, Cordero-Pérez C, Famularo MR, Fernández-Pato N, González-Lanza C, Castañón-Ordónez L, Rojo-Vázquez FA, 2013. Prevalence of gastrointestinal nematodes and *Fasciola hepatica* in sheep in the northwest of Spain: relation to climatic conditions and/or man-made environmental modifications. Parasite Vector 6:282.
Matheron G, 1963. Principles of geostatistics. Econ Geol 58:1246-66.
Miller KA, Siscovick DS, Sheppard L, Shepherd K, Sullivan JH, Anderson GL, Kaufman JD, 2007. Long-term exposure to air pollution and incidence of cardiovascular events in women. New Engl J Med 356:447-58.
Muff S, Riebler A, Held L, Rue H, Saner P, 2014. Bayesian analysis of measurement error models using integrated nested Laplace approximations. J Roy Stat Soc C-App 64:231-52.
Muñoz F, Pennino MG, Conesa D, López-Quílez A, Bellido JM, 2013. Estimation and prediction of the spatial occurrence of fish species using Bayesian latent Gaussian models. Stoch Env Res Risk A 27:1171-80.
Ortiz-Pelaez A, Pfeiffer DU, Tempia S, Otieno FT, Aden HH, Costagli R, 2010. Risk mapping of Rinderpest sero-prevalence in Central and Southern Somalia based on spatial and network risk factors. BMC Vet Res 6:22.
Pennino MG, Muñoz F, Conesa D, López-Quílez A, Bellido JM, 2013. Modelling sensitive elasmobranch habitats. J Sea Res 83:209-18.
Pennino MG, Muñoz F, Conesa D, López-Quílez A, Bellido JM, 2014. Bayesian spatio-temporal discard model in a demersal trawl fishery. J Sea Res 90:44-53.
Pfeiffer DU, Robinson T, 2008. Spatial analysis in epidemiology. Oxford University Press, Oxford, UK.
R Core Team, 2015. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
Raso G, Matthys B, N'goran EK, Tanner M, Vounatsou P, Al E, 2005. Spatial risk prediction and mapping of *Schistosoma mansoni* infections among schoolchildren living in western Côte d'Ivoire. Parasitology 131:97-108.
Raso G, Schur N, Utzinger J, Koudou BG, Tchicaya ES, Rohner F, N'goran EK, Silué KD, Matthys B, Assi S, 2012. Mapping malaria risk among children in Côte d'Ivoire using Bayesian geo-statistical models. Malaria J 11:160.
Ribeiro PJ, Christensen OF, Diggle PJ, 2003. Geor and Georglm: software for model-based geostatistics. Available from: https://www.r-project.org/conferences/DSC-2003/Proceedings/RibeiroEtAl.pdf
Rue H, Held L, 2005. Gaussian Markov random fields: theory and applications. CRC Press, Boca Raton, FL, USA.
Rue H, Martino S, Chopin N, 2009. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. J Roy Stat Soc B 71:319-92.
Rushtong G, Lolonis P, 1996. Exploratory spatial analysis of birth defect rates in an urban population. Stat Med 15:717-26.
Schur N, Hurlimann E, Garba A, Traoré MS, Ndir O, Al E, 2011. Geostatistical model-based estimates of schistosomiasis prevalence among individuals aged ≤20 years in West Africa. PLoS Neglect Trop D 5:e1194.
Snow J, 1857. On the adulteration of bread as a cause of rickets. Lancet 70:4-5.
Stein ML, 1999. Interpolation of spatial data. Some theory for Kriging. Springer, New York, NY, USA.
Szpiro AA, Paciorek CJ, 2012. Measurement error in two-stage analyses, with application to air pollution epidemiology. Environ Health Persp 24:501-17.
Waller LA, Gotway CA, 2004. Applied spatial statistics for public health data. Wiley-Interscience, Koboke, NJ, USA.
Wannemuehler KA, Lyles RH, Waller LA, Hoekstra RM, Klein M, Tolbert P, 2009. A conditional expectation approach for associating ambient air pollutant exposures with health outcomes. Environmetrics 20:877-94.
Wardrop NA, Atkinson PM, Gething PW, Fèvre EM, Picozzi K, Al E, 2010. Bayesian geostatistical analysis and prediction of Rhodesian human African trypanosomiasis. PLoS Neglect Trop D 4:e914.
Webster R, Oliver MA, Muir KR, Mann JR, 1994. Kriging the local risk of a rare disease from a register of diagnoses. Geogr Anal 26:168-85.
Zhong S, Xue Y, Cao C, Cao W, Li X, Guo J, Fang L, 2005. Explore disease mapping of hepatitis B using geostatistical analysis techniques. Lect Notes Comput Sci 3516:464-71.