Helena Baptista*, Peter Congdon, Jorge M. Mendes, Ana M. Rodrigues, Helena Canhão and Sara S. Dias

Disease mapping models for data with weak spatial dependence or spatial discontinuities

https://doi.org/10.1515/em-2019-0025
Received November 28, 2019; accepted October 26, 2020; published online November 11, 2020

Abstract: Recent advances in the spatial epidemiology literature have extended traditional approaches by including determinant disease factors that allow for non-local smoothing and/or non-spatial smoothing. In this article, two of those approaches are compared and are further extended to areas of high interest from the public health perspective. These are a conditionally specified Gaussian random field model, using a similarity-based non-spatial weight matrix to facilitate non-spatial smoothing in Bayesian disease mapping; and a spatially adaptive conditional autoregressive prior model. The methods are specially design to handle cases when there is no evidence of positive spatial correlation or the appropriate mix between local and global smoothing is not constant across the region being study. Both approaches proposed in this article are producing results consistent with the published knowledge, and are increasing the accuracy to clearly determine areas of high- or low-risk.

Keywords: bayesian modelling; body mass index(BMI); limiting health problems; spatial epidemiology; similarity-based and adaptive models.

Background

To allocate the scarce health resources to the spatial units that need them the most is of paramount importance nowadays. Methods to identify excess risk in particular areas should ideally acknowledge and examine the extent of potential spatial clustering in health outcomes (Tosetti et al. 2018).Identification of risk may also be based on relatively rare area health outcomes, and model based methods are required for spatial smoothing, typically using Bayesian principles (Best, Richardson, and Thomson 2005).Where it exists, spatial clustering is the basis for local smoothing, or spatial borrowing of strength. Conditionally specified Gaussian Markov random field models with adjacency-based neighborhood weight matrix, have been the mainstream approach to spatial smoothing in Bayesian disease mapping. However, there are cases when there is no evidence of positive spatial correlation or the appropriate mix between local and global smoothing is not constant across the region being studied. Two models are considered for such situations, a conditionally specified Gaussian random field model (GRF) using a similarity-based non-spatial weight matrix to facilitate non-spatial

***Corresponding author: Helena Baptista**, NOVA Information Management School, Universidade Nova de Lisboa, Lisboa, Portugal, E-mail: mhbaptista@novaims.unl.pt
Peter Congdon, School of Geography and Life Sciences Institute, Queen Mary, University of London, London, United Kingdom of Great Britain
Jorge M. Mendes, NOVA Information Management School, Universidade Nova de Lisboa, Lisboa, Portugal
Ana M. Rodrigues, Nova Medical School, Lisbon, Portugal; Comprehensive Health Research Center, Lisbon, Portugal
Helena Canhão, Comprehensive Health Research Center, Lisbon, Portugal; EpiDoC Unit, CEDOC, NOVA Medical School, Universidade Nova de Lisboa, Lisboa, Portugal
Sara S. Dias, Comprehensive Health Research Center, Lisbon, Portugal; EpiDoC Unit, CEDOC, NOVA Medical School, Universidade Nova de Lisboa, Lisboa, Portugal; and Center for Innovative Care and Health Technology (ciTechCare), School of Health Sciences, Polytechnic of Leiria, Leiria, Portugal

smoothing in Bayesian disease mapping, and a spatially adaptive conditional autoregressive prior model. The former model, named similarity-based GRF, is motivated for modelling disease mapping data in situations where the underlying small area relative risks and the associated determinant factors do not varying systematically in space, and the similarity is defined by similarity with respect to the associated disease determinant factors. The latter model considers a spatially adaptive extension of Leroux, Lei, and Breslow (2000) prior to reflect the fact that the appropriate mix between local and global smoothing may not be constant across the region being studied. Local smoothing will not be indicated when an area is disparate from its neighbours (in terms of social or environmental risk factors). A large epidemiological study run in Portugal and data for London areas (long term illness, and breast cancer) are used to test the ability to improve on the smoothing process. Results are presented and we can conclude that both models behave as expected, producing results that are more consistent with the published knowledge of the studied patterns.

Case studies

Spatial epidemiology models are being extensively used to describe geographical patterns of mortality and morbidity rates. Information provided by these models is considered invaluable by health researchers and policy-makers as it allows, for example, to effectively allocate funds in high risk areas, and/or to plan for localised prevention/intervention programmes.

To illustrate the important role that socioeconomic determinants play in spatial structuring of disease we will use two datasets.

Rodrigues et al. (2015) detail the design, methodology and planned analyses of the study EpiDoC1. EpiDoC1 is a national population-based survey designed by the Portuguese Society of Rheumatology. Along with information on rheumatic and musculoskeletal diseases, information on the Body Mass Index (BMI) was also collected. Nowadays, obesity has become a civilisation disease and the proportion of people suffering from it continues to grow, especially in the developed countries. There are plenty of published evidence of the relationship between income and poverty on the probability of being obese. The paradox of obesity and poverty relationship is observed especially in the developed and developing countries. Some recent work from Zukiewicz-Sobczak et al (2014), presents the reasons for the growing obesity in the population of poor people as potentially being the higher unemployment, lower education level, and irregular meals. Salmasi and Celidoni (2017) studied the effect of income- and wealth-based poverty on the probability of being obese for the elderly in Europe and concluded that poor individuals are more likely to be obese than non-poor individuals. The risk factor for the BMI data collected by EpiDoC1 used in this paper is the *per capita purchasing power index* (PcPp), used as a proxy of income (INE 2013).

We also use data on two health outcomes for London small areas. One dataset, drawing from the UK 2011 Census, concerns the prevalence of long-term health problems or disabilities that limit day-to-day activity. For these data the global spatial shrinkage principle may need to be modified when areas are distinct socioeconomically from their neighbours. The other dataset concerns breast cancer incidence. Here we see how deprivation and ethnic mix affect incidence itself, and the extent to which spatial shrinkage is appropriate.

In cases of epidemiological studies with relatively small sample sizes in some (almost all) of the areas, the classical estimators of the morbidity rates show high variability, and spatial disease mapping models overcome that by borrowing strength from spatial *neighbours*. One rationale is that the spatial random effects used to implement such borrowing of strength are proxies for unobserved risk factors that vary smoothly in space. Models used in disease mapping are usually generalized linear mixed models (GLMM) formulated within a hierarchical Bayesian framework, and Poisson or Binomial likelihood is often assumed for data in the form of counts of cases for each areal unit. Neighbourhood information is explicitly incorporated into the model by means of an appropriate prior specification.

The seminal work of Besag, York, and Mollié (1991) provides a pair of area-specific random effects to model unstructured heterogeneity (extra-Poisson variation) and spatial similarity. The Besag-York-Mollié (BYM) model is an extension of the intrinsic conditional autocorrelation (CAR) model, a well known Gaussian Markov

random field (GMRF) prior in disease mapping. In the same field, Leroux, Lei, and Breslow (2000) (LLB) proposed a conditional autoregressive prior incorporating a spatial correlation parameter, with its extreme values corresponding to pure spatial and pure unstructured residual variation. One important aspect of the CAR modelling is the definition of the so-called neighbourhood matrix, which characterises the spatial structure of the data at hand, and is based on the concept of *neighbours*. Griffith (1996) highlights the importance of the selected specification of the neighbourhood in spatial analysis of areal data.

The debate on the definition of *neighbours* can be traced back to Besag (1974). Others have worked in defining *neighbours* in several different ways, Besag et al. (1991) defines *neighbours* as those regions sharing a common boundary, Best et al. (1999) uses the distance between the centroids of local areas to define neighbourhoods, while Lee and Mitchell (2013) work on cases in which one area is disparate from its *neighbours* and implement local adaptive spatial smoothing.

Recent work from Etxeberria, Goicoa, and Ugarte (2018) has approached this problem from a different angle, by using two or more related diseases. The bibliography therein shows clearly that this is also an area with important recent developments. For a more detailed explanation of multivariate Gaussian Markov random fields refer to MacNab (2018).

Most of the research in disease mapping is related with diseases resulting from environmental exposures, such as respiratory complications and cancer. Those extrinsic disease determinant factors vary smoothly in space, and using some kind of spatial proximity, either by adjacency or by distance, between areas in the definition of *neighbours* has therefore provided good results. In cases in which no spatial positive autocorrelation is displayed by the data, the neighbourhood matrix as it exists today may not be adequate. The similarity-based GRF approach, proposed by Baptista et al. (2016), replaces the neighbourhood-based GMRF approach. The structure of the conditionals is maintained, but the smoothing and borrowing strength mechanisms are now based on the similarity of the areas, regardless of their relative location in space.

Another approach to the same aspect is proposed by Congdon (2008), where it is argued that uniform borrowing of strength based simply on proximity or contiguity may not be appropriate when there are discontinuities in the spatial pattern of health events or risk factors; for instance, a low mortality area surrounded by high mortality areas. Such discontinuity may often reflect spatial discontinuities in risk factors, whether observed or unobserved. An area showing such discontinuity may have a distorted smoothed rate when smoothing is towards the local mean.

The datasets mentioned before (BMI in Portugal and health outcomes problems in London) illustrate the implementation of these techniques. We will present results of the implementation of the above two mentioned models. The data will provide an introduction to the data, and will introduce the basic model as proposed by Leroux, Lei, and Breslow (2000). A similarity-based Gaussian random field model will provide an overview of the similarity model while A Spatially Adaptive Conditional Autoregressive Prior - Modifying uniform association and borrowing of strength will provide an overview of the adaptive model. Results will provide the results of the application of both models to the BMI data collected by the EpiDoC1 study (Rodrigues et al. 2015) and to health outcomes data for London. Conclusions will end with a summary discussion.

The data and the basic model

The data

Rodrigues et al. (2015) provide the proportion of adults (non-institutionalised people with ≥18 years old, living in private households in Portugal), with a prevalence rate of overweight people to be 35.1%, as measured by the BMI (calculated based on the World Health Organization guidelines). In this paper the study region consider is mainland Portugal, excluding Islands (Madeira and Azores), which is partitioned into 28 units (areas) called NUTS 3 (Nomenclature of territorial units for statistics, as defined by Eurostat, the European statistics authority, corresponding to the third level territorial units aggregation).

As in other developed countries, an association with income and BMI was found in Portugal. Higher income people tend to have a lower BMI. However, at the aggregated level, the PcPp does not show a correlation with the number of cases in each area, as it will be shown in Results. PcPp has the value 100 at the national level; areas with values below 100 have a lower than national per capita purchasing power. We may be in the presence of the well-known "ecological fallacy" (Wakefield and Lyons 2010), which refers to the difference between estimated associations on ecological- and individual-level data.

Two health outcomes are considered for London. The first, for all London, concerns limiting health problems among males aged 65–69, taken from the 2011 UK census data. The data are at a small area level: for 983 Census areas called middle super output areas, or MSOAs for short. The observed data consists of cases of long term illness, the population aged 65–69, and information on an index of multiple deprivation (IMD). Prevalence of limiting health problems is closely related to deprivation: areas with an above median IMD have a higher mean prevalence of 43%, compared to 29% in areas with lower IMD scores. The other outcome, for a London subregion with 83 MSOAs, concerns breast cancer incidence. Evidence from other sources indicates potential associations between incidence and deprivation, and between incidence and ethnic mix.

The basic model

A general formulation of the likelihood of a Bayesian hierarchical model can be the following. Suppose y_i are counts, and that N_i are populations at risk in n small areas labelled as i=1, …, n, with $y_i \sim Bin(N_i, \pi_i)$. Then, when the event is relatively frequent, one may specify

$$\text{logit}(\pi_i) = \mathbf{X}_i \boldsymbol{\beta} + s_i,$$

where π_i are latent probabilities of the event, $\mathbf{X}_i$ are covariates including an intercept term, $\boldsymbol{\beta}$ is a vector of regression parameters, and the s_i are latent random effects that may be spatially dependant.

The random effects are commonly modelled by the class of conditional autoregressive (CAR) prior distributions, which are a type of Markov random field model. These models can be specified in two equivalent ways: by a single multivariate joint distribution $f(s)$, or by a set of n univariate full conditional distributions $f(s_i|\mathbf{s}_{-i})$ where $\mathbf{s}_{-i} = (s_1, \ldots, s_{i-1}, s_{i+1}, \ldots, s_n)$, for i=1, …, n. Conditions for equivalence are discussed by Besag and Kooperberg (1995) and Brook (1964), including symmetry constraints which ensure the conditional distributions yield a valid joint distribution. The specification of s_i involves an $n \times n$ spatial interaction matrix W=$[w_{ij}]$ indexing areas i and j. The w_{ij} can be based on inter-area distances, but are most commonly defined by adjacency: w_{ij} =1 if areas i and j share a border, w_{ij} =0 otherwise. In this case, when two areas share a common border, they are considered *neighbours*, a property denoted in this paper by $i \sim j$, and d_i denotes the number of *neighbours*. If two areas are *neighbours* their random effects are correlated, while random effects in *non-neighbouring* areas are modelled as being conditionally independent given the remaining elements of s.

The s_i are generally taken to represent unmeasured risk factors, assumed to be positively correlated in space and so produce smoothly varying disease risk; they "encode the belief that the residual spatial random effects of nearby areas have similar values" (Smith,Wakefield, and Dobra, 2015).This framework may however lead to over-smoothing, masking discontinuities in disease risk.

One attempt to modify the uniform smoothing principle is provided in the LLB model (Leroux, Lei, and Breslow 2000). Here the random effects s=(s_1, …, s_n) have a joint density which is a multivariate Gaussian distribution.

$$s \Big| W, \tau^2, \lambda, \mu \sim N\left(\boldsymbol{\mu}, \tau^2 [\lambda W + (1-\lambda) I_n]^{-1}\right). \quad (1)$$

In this density the mean is $\boldsymbol{\mu}$, often taken as a zero constant, while τ^2 is a scaling parameter. Define $\omega = 1/\tau^2$, and $Q = [\lambda W + (1 - \lambda) I_n]$ Then the precision matrix (inverse covariance matrix) in the joint prior is ωQ. When W is defined by adjacency, the matrix Q in the joint prior has elements

$$q_{ij} = \begin{cases} 1 - \lambda + \lambda d_i, & \text{if } i = j \\ -\lambda, & \text{if } i \sim j \\ 0, & \text{otherwise,} \end{cases}$$

The precision matrix is hence a weighted average of spatially dependent (represented by W) and independent (represented by I_n) correlation structures, where the weight is equal to λ.

This spatial prior can be expressed equivalently by a set of conditional densities (Banerjee, Carlin, and Gelfand, 2014):

$$s_i | s_{-i} \sim N\left(\frac{\lambda}{1 - \lambda + \lambda d_i} \sum_{j \sim i} s_j, \frac{\tau^2}{1 - \lambda + \lambda d_i}\right).$$

This model can represent a range of weak and strong spatial correlation structures, with the special case of λ=0 simplifying to a model with independent random effects. When there is strong spatial correlation in the data, λ will be close to one and the conditional variance is approximately τ^2/d_i. In contrast, if the random effects are independent, the conditional variance is a constant, τ^2. For non-zero λ, the terms $a_i = 1 - \lambda + \lambda d_i$ scale the variance in the conditional prior: the larger the number of neighbours d_i, the more precisely is the spatial effect for area i defined. However, a non-zero λ still acts to produce uniform spatial smoothing without local adaptivity. In coming sections the interaction matrix W, the precision matrix Q and consequent conditional densities are central in defining proposed models, which allow more flexibility in modelling spatial risk surfaces for disease.

A similarity-based Gaussian random field model

The GRF model proposed by Baptista et al. (2016) no longer retains the Markovian properties as those based on the neighbourhood weights, like the LLB model (Leroux, Lei, and Breslow 2000). Instead of using spatial distance or spatial adjacency, a measure reflecting similarity between areas is introduced. This requires a deep knowledge of the disease data at hand, and therefore cannot be governed by convenience and/or convention. Data used should come from: a) a disease determinant factor or a combination of factors, b) a source external to the survey that collected the disease data. The main objective of the proposed model is the provision for borrowing strength between areas with similar disease determinant factors.

Firstly, regions exhibiting the *same or close* level of risk in a determinant factor will be regions with the *same or close* risk of the disease. Secondly, if disease data need to be *strengthened*, using disease determinant factor information collected by the same survey might inflate or not remediate possible *weaknesses* of the disease data. Therefore, an external source for the disease determinant factor is critical.

The rationale of our approach is the following: in cases of diseases with no environmental determinant factors, use of a positive spatial correlation based on physical distance or adjacency, in the GRF/GMRF model, may not be the best way to reflect similarity between areas. By using the GRF model reflecting *how similar* each area is to one another, in terms of a disease determinant factor that was collected by an external source, the disease risk distribution can be better assessed.

Based on a matrix definition proposed by Best et al. (1999), the new similarity matrix (the W matrix mentioned previously in The basic mode), with elements h_{ij} for each region j, has the following structure:

$$h_{ij} = \begin{cases} e^{-p_{ij}/\delta}, & \text{if } i \neq j \\ \dfrac{1}{n-1} \sum h_{(-i)}, & \text{otherwise,} \end{cases}$$

where p_{ij} is the absolute gap between region i and region j,

$$p_{ij} = |p_i - p_j|, \tag{2}$$

in terms of the disease determinant factor, and δ is equal to a value that gives a relative weight of 1% (h_{ij}=0.01)

to an area i whose difference from an area j is the mean inter-region difference for the country. Elements h_{ii} need a specific definition, otherwise their value would be the one contributing the most to the prior, as e^0=1 and all other h_{ij} elements have values between 0 and 1. Therefore, p_{ii} values are equal to the average value of all elements except the ith area value.

Usually, p values are taken as fixed, assumed measured without error, one single value per region, based on official published statistics and are not subject to any type of inference. When there is only one p, meaning using only one disease determinant factor the absolute gap can be used, as in Eq. (2). However, p_{ij}, as the similarity between regions i and j, can be defined in broader terms. The similarity could correspond to the Euclidean distance in $\Re$ for p_i determinant factors, with i=1, …, n:

$$p_{ij} = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)'(\mathbf{x}_i - \mathbf{x}_j)},$$

or even the multivariate version of the statistical distance, the Mahalanobis distance:

$$p_{ij} = \sqrt{(\mathbf{x}_i - \overline{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x}_j - \overline{\mathbf{x}})}, \tag{3}$$

where $\mathbf{x}_i' = (x_{1i}, x_{2i}, \ldots, x_{pi})$, $\mathbf{x}_j' = (x_{1j}, x_{2j}, \ldots, x_{pj})$, $\overline{\mathbf{x}}' = (\overline{x_1}, \overline{x_2}, \ldots, \overline{x_p})$ and $\mathbf{S}^{-1}$ is the inverse of the sample covariance matrix of the disease determinant factors $\mathbf{x}$. Of course, other possibilities are acceptable like a multivariate linear model.

The proposed methodology has proven to gain efficiency when compared with the BYM model (Besag et al. 1991) by simulation studies (Baptista et al. 2016). In this paper is going to be applied with the LLB (Leroux, Lei, and Breslow 2000) prior. However, results should still apply as both priors (BYM and LLB) use GMRF.

A spatially adaptive conditional autoregressive prior - modifying uniform association and borrowing of strength

The model definition

Borrowing of strength based simply on proximity or contiguity may not be appropriate when there are clear discontinuities in the spatial pattern of health events; for instance, a low mortality area surrounded by high mortality areas. Such discontinuity may often reflect spatial discontinuities in risk factors, whether observed or unobserved. An area showing such discontinuity may have a distorted smoothed rate when smoothing is towards the local mean. This provides one reason why a model with spatial adaptivity in the association between health risks in adjacent areas may be beneficial. Another reason is that a model with a single parameter representing spatial association - as is the case in the Leroux, Lei, and Breslow (2000) prior and the proper CAR prior (Smith et al. 2015) - may oversimplify association in large regions.

The similarity based GRF prior (A similarity-based Gaussian random field model) replaces spatial proximity as a basis for borrowing strength by similarity in one or more risk factors, and so takes explicit account of the actual spatial pattern of risk factors. By contrast, here we discuss a spatially adaptive approach, which retains the broad principle of spatial borrowing of strength, but modifies it to better represent discontinuities in the outcome and/or observed risk factors. It may also be relevant when a single association parameter is an over-simplification, even in the absence of major discontinuities. The degree of spatial correlation is allowed to vary between sub-regions of the region under consideration, with one possible scheme linking varying spatial correlation to spatial similarity (or dissimilarity) in risk factors. For example, mortality is commonly linked to socio-economic deprivation, and spatial correlation in mortality may be weaker when socio-economically distinct areas are adjacent, such that is there is local dissimilarity in risk factors.

Let us return to the already mentioned s_i random effects. While spatially correlated random effects s=(s_1, …, s_n), assuming uniform spatial correlation, and hence uniform smoothing to the local mean, may be postulated, this can amount to an informative prior assumption. There may be spatially disparate areas

defined by risk factors (e.g. deprived areas with mainly social renting surrounded by affluent areas with mainly owned housing). There may, in such situations, be a gain from a prior that instead allows local downweighting of the principle of uniform spatially based borrowing of strength. We consider developing such a prior based on the principles for specifying spatial priors jointly and conditionally.

Here we propose a spatially adaptive version of the LLB model (Leroux, Lei, and Breslow 2000) based on area specific λ_i. The λ_i are varying indicators of spatial association (in disease risk between area i and its neighbours that adapt the principle of global smoothing (inherent in a uniform parameter λ) to allow locally adaptive smoothing and represent discontinuities in the disease risk surface. Distinctly low λ_i correspond to spatially disparate areas, unlike their neighbours in health risk and/or risk factors, so that there may be benefit in downweighting the principle of uniform pooling to the neighbourhood mean.

Let W=[w_{ij}] (for n areas i and j) denote spatial interactions as discussed in The basic model. The precision matrix in the joint prior is ωQ, where $\omega = 1/\tau^2$, and Q has diagonal elements $1 - \lambda_i + \lambda_i \sum_j w_{ij}$, and off-diagonal elements $Q_{ij} = -\lambda_i \lambda_j w_{ij}$.

The conditional prior is

$$s_i | s_{-i} \sim N\left(\frac{\lambda_i}{1 - \lambda_i + \lambda_i \sum_j w_{ij}} \sum_{j \neq i} \lambda_j s_j, \frac{\delta}{1 - \lambda_i + \lambda_i \sum_j w_{ij}} \right). \tag{4}$$

Typically the w_{ij} are binary indicators of adjacency such that $\sum_j w_{ij} = d_i$, as discussed in The basic model.

As λ_i tends to zero the conditional prior reduces to an iid normal density, with mean zero and variance δ, so that the random term s_i is independent of the neighbourhood. As λ_i tends to 1, the conditional prior tends to that in the specification of Besag et al. (1991), so that the spatial effect of area i is entirely determined by the average of spatial effects in neighbouring areas.

We need to specify a prior density for the set of adaptive association parameters λ_i. Just as λ is between 0 and 1, so are these varying indicators.

Possible priors for the λ_i include beta priors, or probit-normal or logit-normal priors, such as

$$logit(\lambda_i) \sim N(\mu_\lambda, 1/\tau_\lambda)$$

where the average and precision $\{\mu_\lambda, \tau_\lambda\}$ are extra unknowns. However, if predictors R_i measuring dissimilarity in observed risk factors are available, and so potentially relevant to whether local pooling be modified, one can use the scheme

$$logit(\lambda_i) \sim N(R_i \gamma, 1/\tau_\lambda),$$

where γ are regression parameters and R are covariates measuring spatial dissimilarity in risk factors between area i and surrounding areas. For example, supposing there is a single dissimilarity index, then

$$logit(\lambda_i) \sim N(\gamma_1 + \gamma_2 R_i, 1/\tau_\lambda), \tag{5}$$

where γ=(γ_1, γ_2) are regression parameters. One would expect lower λ_i for areas dissimilar from their neighbours on the risk factor; that is, γ_2 is anticipated to be negative. It may be relevant to transform R_i in the event of skewness in the dissimilarity index. A nonlinear regression

$$logit(\lambda_i) \sim N(\gamma_1 + g(R_i), 1/\tau_\lambda)$$

where $g(R_i)$ is a smooth function may also be considered. In Congdon (2008), the dissimilarity measure is based on a measure z_i of socioeconomic deprivation, and dissimilarity measured as

$$R_i = \left| z_i - \overline{Z}_i \right| \tag{6}$$

with $\overline{Z}_i$ being the average deprivation level in the locality L_i around area i, namely $\overline{Z}_i = \sum_{j \in L_i} z_j / d_i$.

Table 1: Simulated datasets, adaptive LLB model, and estimated parameters.

	Number of low λ	mean(λ_i)	π_{max}/π_{min}	γ_2	β_2	σ_λ
Simulated datasets (Preset or Mean of simulations)	34.15	0.576	2.60	−1.5	0.03	1
Results from adaptive LLB model Applied to simulated datasets						
Mean	32.67	0.594	2.57	−1.47	0.0297	1.20
2.5%	29.24	0.557	2.52	−1.70	0.0291	1.16
97.5%	36.10	0.632	2.63	−1.24	0.0304	1.25
Median	30.97	0.606	2.54	−1.60	0.0297	1.16

A simulation study

We demonstrate validation of the adaptive LLB by simulating long term illness data y in a London subregion (the three boroughs of Redbridge, Havering, and Barking and Dagenham in outer NE London). The R code used is in the Appendix. There are 83 MSOAs in this region, and we define W by binary adjacency. The simulated data y_i are assumed binomial, as in The basic model. Known data in the simulations are male populations (N_i) and a single risk factor X_i, an index of multiple deprivation (IMD). This index also defines a single dissimilarity measure R_i, which is used to generate λ_i via logit regression, as in Eq. (5). Spatial random effects s_i are simulated from the joint multivariate prior. The simulated spatial effects depend on also simulated λ_i. The simulation specifications are for (γ_1, γ_2), β_1 (intercept) and β_2(impact of deprivation on illness probabilities π_i), and for specified $\sigma_\lambda = 1/\tau_\lambda^{0.5}$ and $\omega = 1/\tau^2$. Specifically γ= (0, −1.5), β=(−0.5, 0.03) ,σ_λ=1, and ω=20. The precision matrix in the joint prior Q in the code is ωQ, where $Q_{ii} = 1 - \lambda_i + \lambda_i d_i$, and the off-diagonal elements are $Q_{ij} = -\lambda_i\lambda_j$ for neighbouring areas, and Q_{ij}= 0 otherwise. The Appendix also contains the BUGS code used in actual estimation from data; this code uses the conditional prior for the s_i.

We simulate 100 datasets, and define criterion indices which measure the correspondence between the parameters (and risk patterns) estimated from the simulated data, and the parameters assumed for (and consequent risk patterns present in) the simulated data. These criteria are the (a) the number of low λ_i, namely below 0.5; (b) the mean of the simulated λ_i; (c) the ratio of the maximum simulated illness probability π_i (over the 83 areas) to the minimum simulated probability; (d) the mean of the estimated γ_2; (e) the mean of the estimated β_2; and (f) the mean of the estimated σ_λ. For example, Table 1 shows that the average number of low λ_i in the 100 simulated datasets is 34.15, while the estimated mean number of low λ_i from applying the adaptive LLB model to these datasets is 32.67 with 95% interval (29.24, 36.10). The 95% intervals for the estimated parameters also contain the other criteria, except for σ_λ which is slightly overestimated.

Results

Portuguese BMI data

Figure 1 shows the overweight prevalence rate calculated with the data collected by the epidemiological study (Rodrigues et al. 2015) for each of the 28 areas in Portugal. The prevalence rate values do not seem to have spatial correlation and show some discontinuities.

Portuguese overweight data are taken as binomial $y_i \sim Bin(P_i, \pi_i)$, with π_i being overweight probabilities.

The two first models are run for 100 000 McMC (Markov chain Monte Carlo) samples with convergence obtained according to the criteria of Brooks and Gelman (1998). Inferences are based on the last 50 000 iterations. The uniform association Leroux model (Leroux, Lei, and Breslow 2000) is applied first, with relative risks modelled as

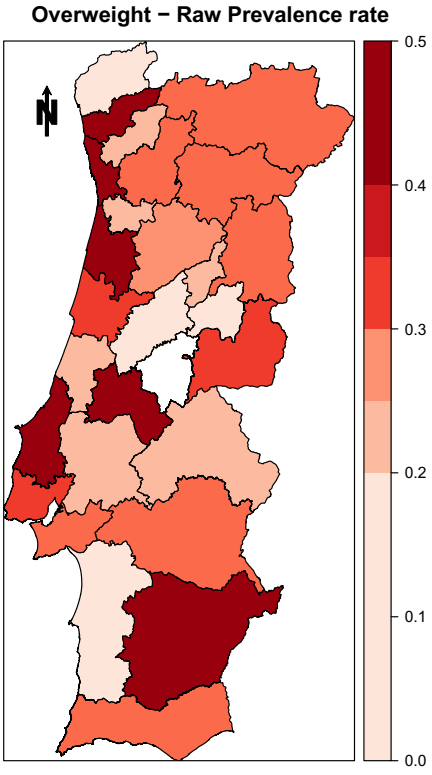

Figure 1: Raw prevalence rate for overweight cases in Portugal.

$$logit\,(\pi_i) = \beta_0 + s_i,$$

where β_0 is an intercept, and the s_i are as in Eq. (1). The WAIC (Vehtari, Gelman, and Gabry, 2016) (and s.d.) is 394.3 (5). A low association parameter λ with posterior median (95% Credible Region Interval - CRI) of 0.35 (0.01.0.91) is obtained.

We then apply the adaptive model, with s_i as in Eq. (4), and the λ_i having a prior as in Eq. (5). Risk factor dissimilarity, measured as in Eq. (6), is based on a measure of the PcPp (INE 2013). The W_i values are centred, so that

$$\lambda_\gamma = exp\,(\gamma_1)/\,(1 + \exp\,(\gamma_1))$$

provides a measure of the overall spatial association, analogous to λ in Leroux, Lei, and Breslow (2000).

A similar fit to the first model is obtained, with an unchanged WAIC (s.d.) of 394.4 (5). However significant variation in the λ_i is evident, with $\sigma_\lambda = 1/\tau_\lambda^{0.5}$ having posterior median (95% CRI) of 0.12 (0.05, 0.58). The posterior 95% CRI of γ_2 in Eq. (5) is biassed to negative values, namely (–1.17.0.14). The posterior medians of the area-specific λ_i vary from 0.00 to 0.10, with posterior median (95% CRI) of the overall measure λ_γ being 0.02 (0.00.0.16). This approach emphasizes the lack of spatial association in the data.

Lastly, we apply the similarity model, using the similarity matrix proposed in A similarity-based Gaussian random field model. To calculate the similarities between the areas we also use the PcPp (INE 2013), using Eq. (2). Posterior inference is based on 9 000 McMC samples, which are obtained by running one chain for 100 000 samples, by which convergence is assumed to have occurred. We ignore the first 10 000 samples as burn-in, and use the remaining 90 000 subsequent samples to obtain the posterior distributions of the parameters of interest (a thin of 10 is used to avoid autocorrelation). The WAIC (s.d.) is 394.8 (5) . The s_i are as in Eq. (1).

The whole posterior distribution can be usefully exploited in an effort to detect true raised- and diminished-risk areas. We calculated the standardised morbidity ratio (SMR), using an indirect standardisation based on the size

and demographic structure of the population living in each area. Areas with elevated risk will display an SMR above one. Table 2 shows the SMRs calculated by the three different models.

It is important to mention the effect of the similarity model and the adaptive model when compared with the results of the LLB model. The first case is in the area of "Grande Porto". This area has a PcPp value above the country average, but the raw and the LLB posterior SMR shows a value above one. The similarity model is able to recognise the fact and the posterior credible interval includes the value one, making it no longer a high risk area. The inverse happens in two other areas, "Tâmega" and "Alto Trás-os-Montes". Both areas have a low value of PcPp, which would indicate a high risk for the disease, but LLB model posterior SMR consider those as low-risk areas. The similarity model is able to include uncertainty and the posterior credible interval includes the value one now, which means those areas are no longer low risk areas. There is one example when the effect is not the expected one. The area "Península de Setúbal" has an PcPp above 100 and the LLB model is keeping it as a low risk area. The similarity model includes uncertainty on that value and the credible interval includes the value one now. It may be due to the fact that the PcPp value is close to 100 (101.09). The adaptive model, in the area of "Grande Lisboa" produces a change that is also expected. Both the LLB model and the similarity model produce results that include the value one in the posterior credible interval. The adaptive model is able to define it as a low risk area, as it was expected given the very high PcPp.

Underlining the importance of the choice of the disease determinant factors, a second similarity model was run. We apply the similarity model, using the similarity matrix proposed in A similarity-based Gaussian random field model. From other analysis conducted with subject level data, a relationship between overweight and gender was found. Men have an higher probability of being overweight in Portugal. Therefore, to calculate

Table 2: SMR results. SMR posterior median (95% CRI).

Area	PcPp	Raw SMR	LLB	Adaptive	Similarity
M-Lima	77.57	0.49	0.49 (0.45, 0.54)	0.49 (0.45, 0.54)	0.50 (0.45, 0.55)
Cávado	85.88	1.15	1.15 (1.10, 1.20)	1.15 (1.10, 1.20)	1.15 (1.03, 1.29)
Ave	81.15	0.70	0.70 (0.66, 0.74)	0.70 (0.66, 0.74)	0.70 (0.61, 0.80)
G Porto	111.28	1.21	1.21 (1.18, 1.23)	1.21 (1.18, 1.23)	1.05 (0.97, 1.13)
Tâmega	67.15	0.86	0.86 (0.82, 0.90)	0.86 (0.82, 0.90)	0.87 (0.73, 1.12)
E Douro Vouga	84.42	0.67	0.67 (0.62, 0.72)	0.67 (0.62, 0.72)	0.67 (0.61, 0.74)
Douro	74.08	0.94	0.94 (0.88, 1.00)	0.94 (0.88, 1.00)	0.94 (0.86, 1.03)
Alto T-os-M	72.35	0.92	0.92 (0.86.0.98)	0.92 (0.86, 0.98)	0.92 (0.84, 1.01)
Algarve	96.74	0.84	0.84 (0.80.0.89)	0.84 (0.80, 0.89)	0.85 (0.72, 0.99)
B Vouga	90.87	1.23	1.23 (1.18, 1.28)	1.23 (1.18, 1.28)	1.23 (1.11, 1.43)
B Mondego	102.82	1.02	1.02 (0.97, 1.07)	1.02 (0.97, 1.07)	1.02 (0.88, 1.21)
Pinhal Litoral	94.27	0.59	0.59 (0.54, 0.64)	0.59 (0.54, 0.64)	0.59 (0.53, 0.65)
Pinhal I Norte	67.42	0.48	0.49 (0.43, 0.55)	0.49 (0.43, 0.55)	0.49 (0.42, 0.55)
Dão-Lafões	78.05	0.82	0.82 (0.77.0.87)	0.82 (0.77, 0.87)	0.82 (0.75, 0.89)
Pinhal I Sul	64.44	0.00	0.79 (0.38, 1.36)	0.79 (0.39, 1.36)	0.77 (0.35, 1.38)
S da Estrela	69.82	0.65	0.65 (0.55, 0.77)	0.66 (0.55, 0.77)	0.66 (0.55, 0.78)
Beira I Norte	76.78	0.83	0.83 (0.75, 0.91)	0.83 (0.75, 0.91)	0.83 (0.75, 0.92)
Beira I Sul	86.75	1.03	1.02 (0.92, 1.12)	1.02 (0.92, 1.12)	1.02 (0.92, 1.12)
C da Beira	80.37	0.54	0.55 (0.47, 0.63)	0.55 (0.47, 0.63)	0.55 (0.48, 0.63)
Oeste	89.51	1.19	1.19 (1.14, 1.24)	1.19 (1.14, 1.24)	1.19 (1.09, 1.32)
Médio Tejo	86.66	1.32	1.32 (1.26, 1.38)	1.32 (1.26, 1.38)	1.32 (1.24, 1.40)
G Lisboa	142.41	0.97	0.97 (0.95, 1.00)	0.97 (0.95, 0.99)	1.05 (0.97, 1.13)
P de Setubal	101.09	0.91	0.91 (0.88, 0.95)	0.91 (0.88, 0.95)	0.92 (0.82, 1.05)
Alentejo Lit	92.86	0.43	0.44 (0.38, 0.51)	0.44 (0.38, 0.52)	0.45 (0.38, 0.52)
Alto Alentejo	82.03	0.63	0.63 (0.57, 0.70)	0.63 (0.57, 0.70)	0.63 (0.56, 0.71)
Alentejo C	89.62	0.87	0.87 (0.80, 0.93)	0.87 (0.80, 0.93)	0.87 (0.79, 0.94)
Baixo Alentejo	81.18	1.21	1.20 (1.12, 1.28)	1.20 (1.12, 1.28)	1.20 (1.11, 1.29)
Lezíria do Tejo	91.26	0.63	0.64 (0.59, 0.69)	0.64 (0.59, 0.69)	0.64 (0.58, 0.70)

the similarity matrix between the areas we use the PcPp (INE 2013), and the proportion of women in each region. Using Eq. (3) a value was calculated by region and using Eq. (2) the distance between regions were calculated. Posterior inference is based on 9 000 McMC samples, which are obtained by running one chain for 100 000 samples, by which convergence is assumed to have occurred. We ignore the first 10 000 samples as burn-in, and use the remaining 90 000 subsequent samples to obtain the posterior distributions of the parameters of interest (a thin of 10 is used to avoid autocorrelation). The WAIC for this model is at the same level of the uniform association Leroux model. The s_i are as in Eq. (1). Results are not shown because, in terms of SMR (see Table 2) the results obtained by this model are essentially equal to the results obtained with uniform association Leroux model.

Limiting health problems

A second application is to binomial data on limiting health problems among males aged 65–69, based on the 2011 UK Census. The spatial framework is provided by 983 small areas in London (middle level super output areas, or MSOAs). Under the uniform association LLB, the λ parameter has posterior median (95% CRI) of 0.92 (0.40, 0.79), and the WAIC (s.d.) is 6547.8 (35.2). The spatially adaptive LLB, based on dissimilarity model 5 uses a deprivation index (the index of multiple deprivation, IMD) as the basis of the dissimilarity index W_i, and then applies a log transform, so that $logit\,(\lambda_i) \sim N\,(\gamma_1 + \gamma_2 \log\,(W_i), 1/\tau_\lambda)$.

This model has a WAIC (s.d.) of 6549.5 (34.9). Taking account of variability in the fit measure, the two models have essentially comparable fit. In line with expectations, γ_2 has a posterior median (95% CRI) of –2.39 (–3.64,-0.46). The average association measure is 0.937. However, 36 of the posterior median λ_i are under 0.9. Figure 2 shows the spread of λ_i and Figure 3 maps out local variations in association as represented by posterior median λ_i, with clustering of higher λ_i and lower λ_i apparent.

The WAIC (s.d) is 6541.7 (35) on the similarity approach. The similarity matrix is as proposed in A similarity-based Gaussian random field model, based on IMD values, using Eq. (2). The s_i are as in Eq. (1).

We again analysed the whole posterior distribution and compared results. In this case as we have 983 small areas we cannot present the results for all of them. We started by comparing the results from the similarity and adaptive models with the results of the LLB. On a second step we compare the posterior median values of the prevalence rate produce by the three models (LLB, adaptive and similarity) with the raw values.

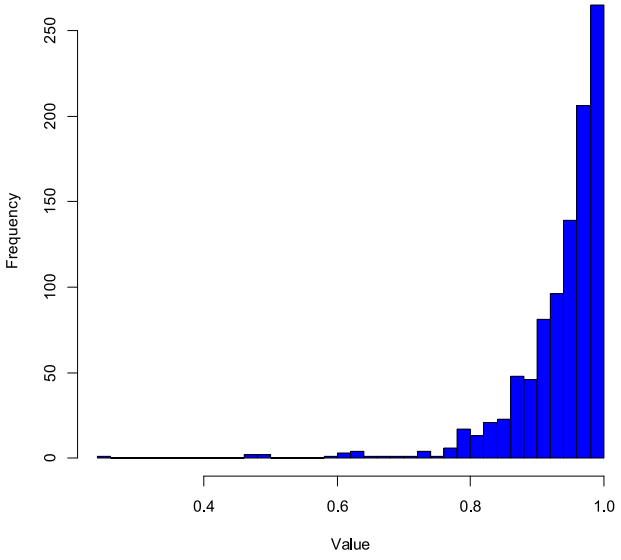

Figure 2: Spread of local lambda statistics, λ_i.

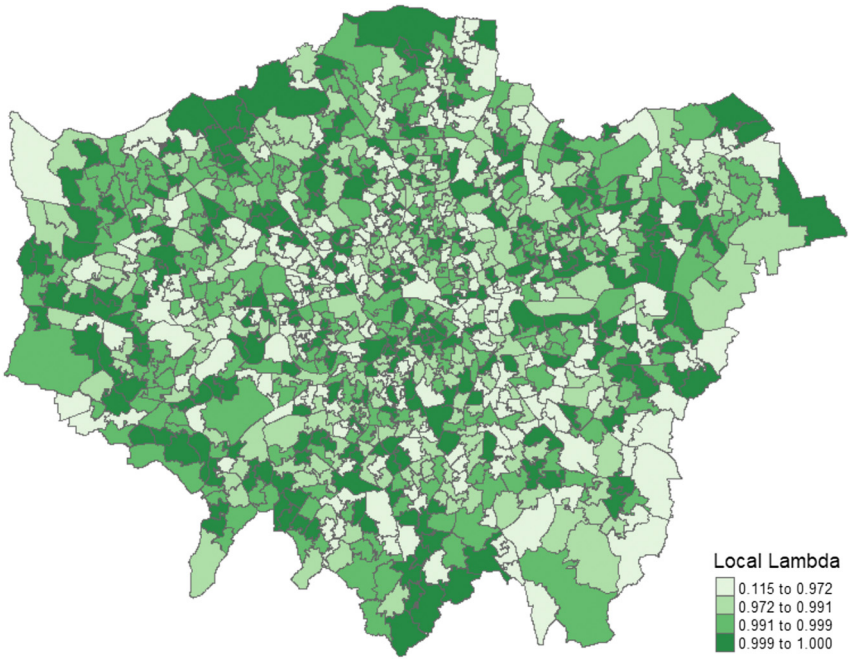

Figure 3: Local variations in association as represented by posterior median λ_i.

Given the fact that the limiting health problems data show positive spatial correlation, the LLB model produces what can be considered good smoothing results. From the 489 areas that have an above the median IMD, all but three (Camden 001, Westminster 015 and Westminster 023) have a posterior LLB SMR above one (i.e. 95% CRI is entirely above 1). From those three, the similarity model further increases the posterior SMR of Westminster 023 to be above one. In the opposite side, from the 494 areas that show a below the median value of IMD, the LLB posterior SMR is below one (i.e. 95% CRI is entirely below 1) for all areas but two (Barking and Dagenham 011 and Hounslow 010). From those two, both the similarity and the adaptive model have posterior SMRs below one for Barking and Dagenham 011.

Table 3 shows the number of small areas for which the respective model mentioned increases the raw prevalence rate. Column two shows from those how many are areas of high risk, as defined by the IMD, meaning those areas with a higher level of deprivation. There are a total of 489 small areas with an higher than the median IMD. All models increase the prevalence rate on about the same number of areas (irrespectively if those areas are low on high prevalence areas). If we analyse where those increases happen, we can see that, for the similarity model, as expected, almost half of those increases happen in areas where due to its level of IMD the prevalence rates were expected to be high. The same effects happens for the low risk areas, those with lower levels of deprivation. In this case (see Table 4) the similarity model is also, as expected, the one having almost half of the decreases happening on areas with a low IMD. There are a total of 494 small areas with a lower than the median IMD.

Table 3: Prevalence rate posterior median comparison. Notes: 1. Number of areas where the respective model increases the prevalence rate. 2. Number of areas which have an above the median value for IMD.

Model	Increases[1]	High IMD[2]	%
Leroux	481	159	33.1%
Adaptive	475	153	32.2%
Similarity	471	210	44.6%

Table 4: Prevalence rate posterior median comparison. Notes: 1. Number of areas where the respective model decreases or maintains the prevalence rate. 2. Number of areas which have a below the median value for IMD.

Model	Decreases[1]	Low IMD[2]	%
Leroux	501	172	34.3%
Adaptive	508	172	33.9%
Similarity	512	233	45.5%

The action of the spatial effects

The action of the spatial effects s_i may be affected to some extent by the choice of covariates X_i in the model for the prevalence rate. Such covariates may reduce the extent of shrinkage in estimated prevalence towards the neighbourhood average. However, in many applications there is still substantial unexplained variation after including relevant covariates, and still the need to avoid over smoothing and acknowledge spatial discontinuities in disease risk.

To illustrate a multivariate application (with multiple risk factors and dissimilarity indicators) we consider the same subregion as in the simulation example of A spatially adaptive conditional autoregressive prior - modifying uniform association and borrowing of strength, namely 83 small areas of outer NE London. The three boroughs in the subregion have distinct ethnic and socioeconomic structures: Redbridge and Barking and Dagenham are socioeconomically and ethnically mixed, whereas Havering is more homogenous: mostly affluent with a majority (over 90%) white adult female population. The health outcome is the standardized incidence ratio for breast cancer for 2012–16 (with known variance), which is treated as normally distributed. Risk factors X_{i1} and X_{i2} are the IMD and the percent of women in each MSOA with Asian or black ethnicity. There is evidence that breast cancer is in fact negatively related to deprivation (Cancer Research UK and National Cancer Intelligence Network 2014), and lower for women of non-white ethnicity (Gathani et al. 2014).The IMD and ethnicity variables are also used to form dissimilarity index R_{i1} and R_{i2}.

We find insignificant risk factor effects: β_2, the coefficient for the impact of IMD on breast cancer incidence has posterior mean (sd) of –0.065 (0.221), while the ethnic variable has a posterior mean (sd) of –0.103 (0.108). The effect of dissimilarity in social deprivation (R_{i1}) is also inconclusive with mean (s.d.) of 0.947 (0.622). However, the coefficient γ_3 (the effect of dissimilarity in ethnic mix on shrinkage towards the neighbourhood incidence) is significantly negative with mean (sd) of –1.58 (0.58) and 95% interval (–2.42, –0.42). Hence spatial pooling towards the neighbourhood average for incidence is contra-indicated for areas with dissimilar

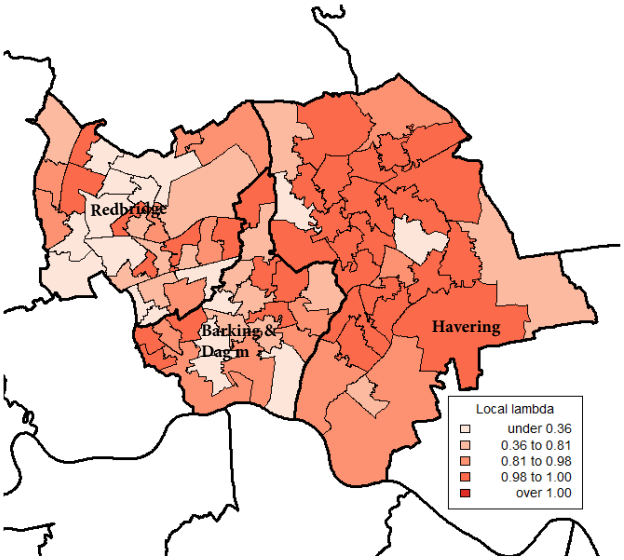

Figure 4: Posterior mean λ_i, North east london, breast cancer incidence.

ethnic structures in their female populations. Figure 4 maps out the posterior mean λ_i. These are highest in the east of the region, namely the borough of Havering, with relatively homogenous ethnic structure (majority white) in different MSOAs.

Conclusions

In the cases of diseases that have not spatially smoothed disease determinant factors, the classical borrowing strength from spatial *neighbours* mechanism can no longer be used.

In this paper we have used two different methods to overcome that circumstance in two datasets. In one case, in the BMI data from Portugal, the disease determinant factor is not correlated with the disease prevalence rate at the aggregated level. In the second case, in the limiting health problems in the UK, the disease determinant factor is correlated with the disease prevalence rate. In the first case, a low association parameter λ with posterior median (95% CRI) of 0.35(0.01, 0.91) indicates the low spatial correlation, while in the second case, a very high association parameter λ with posterior median (95% CRI) of 0.92(0.40, 0.79) clearly indicates spatial correlation.

In the first case, the similarity model is able to identify some of those areas which were being "wrongly" (accordingly with the corresponding PcPp value) considered high- or low-risk areas and transform those into areas of uncertainty. When more disease determinant factors were included, that "adjustment" was lost. This underlines the importance of choosing carefully the disease determinant factors which should be used in the similarity matrix. This increases the complexity of the models, but should be a relatively easy task for epidemiologists.

In the second case, even with a high spatial correlation, the similarity and the adaptive model are also able to identify more areas, than the LLB for which prevalence rates should be increased/decreased depending on the IMD value.

Both models are performing as expected in both cases. In the second case, the adaptive model is producing a negative γ_2, as there are areas of discontinuity, and by doing that is breaking the global smoothing of the LLB model. The similarity model is modelling the prevalence rate accordingly with the similarity between the small areas at the disease determinant factor level.

The unsolved problem is, of course, that both diseases prevalence rates are not the result of a single determinant factor. Including more information will add on the complexity of the models but may also contribute to better smoothing or not depending on the relevance and/or quality of the data included.

Both diseases here modelled are of high relevance for the society, and pose significant challenges to the health-care systems worldwide (Barnett et al. 2012; Wang et al. 2011). Identifying the spatial distribution of those diseases may significantly help in the allocation of the scarce society health resources. However, more important than those two specific cases is the development of methodologies that can help identifying with greater accuracy the areas where more help is needed. As can be seen in this work, using methodologies that incorporate the disease determinant factors can improve the accuracy of the model results, indicating more clearly areas of high- or low-risk.

Acknowledgments: The authors acknowledge the reviewers, whose comments and suggestions helped to improve the presentation of the paper.
Research funding: None declared.
Author contributions: All authors have accepted responsibility for the entire content of this manuscript and approved its submission.
Competing interests: Authors state no conflict of interest.

Appendix

Estimation and simulation programs - 83 areas in NE London

```
# MODEL ESTIMATION CODE (BUGS CODE)
model {for (i in 1:N) {# model for observed data
y[i] ~ dbin(p[i],T[i])
logit(p[i]) <- alpha+beta*(IMD[i]-mean(IMD[]))+s[i]
# replicate data and predictive checks
yrep[i] ~ dbin(p[i],T[i]);
check[i] <- step(yrep[i]-y[i]-0.001)+0.5*equals(yrep[i],y[i])
# log-likelihood
loglik[i] <-  logfact(T[i])-logfact(y[i])-logfact(T[i]-y[i])
+y[i]*log(p[i])+(T[i]-y[i])*log(1-p[i])
# Absolute difference in risk factor between area i and its locality
 absdif.IMD[i]   <- abs(IMD[i]-sum(WIMD[C[i]+1:C[i+1] ])/d[i])}
# Error vector and deprivation over neighbours (listed in map vector)
for (i in 1:NN) { Ws[i] <- s[map[i]]*lam[map[i]]; WIMD[i] <- IMD[map[i]]}
# Priors
tau.s ~ dgamma(1,0.001); tau.lam ~ dgamma(1,0.001)
alpha ~ dnorm(0,0.001); beta ~ dnorm(0,0.001);
for (j in 1:2) {gamma[j] ~ dnorm(0,1)}
# Validation Criteria
p.ext <- ranked(p[],N)/ranked(p[],1)
lam.mu <- mean(lam[])
sd.lam <- sqrt(1/tau.lam)
sum.low <- sum(low.lam[])
# Find lambda values under 0.5
for (i in 1:N) {low.lam[i] <- step(0.5-lam[i])}
# Adaptive spatial prior
for (i in 1:N) { # Regression using risk factor discrepancy
lam[i] <- exp(lgt.lam[i])/(1+exp(lgt.lam[i]))
lgt.lam[i] ~ dnorm(m.lam[i],tau.lam)
m.lam[i] <- gamma[1]+gamma[2]*(absdif.IMD[i]-mean(absdif.IMD[]))
# Spatial effects conditional prior
s[i]  ~ dnorm(S[i],tau[i])
tau[i]   <- tau.s * (1-lam[i]+lam[i]*d[i])
S[i]   <- (lam[i]/(1-lam[i]+lam[i]*d[i]))*sum(Ws[C[i]+1:C[i+1] ])}}

# SIMULATION CODE (R CODE)
library(maptools); library(spdep); library(mvtnorm)
setwd("C:/users/p congdon/documents/qgis maps/msoas 2011 outer ne london")
MSOAmap <- readShapePoly("onel")
MSOAnb <- poly2nb(MSOAmap)
# spatial files: adj, d (numbers of neighbours), N, NN
MSOAspat  <- nb2WB(MSOAnb)
d=MSOAspat$num
adj=MSOAspat$adj
N=length(d); NN=sum(d)
# Interaction matrix
W=matrix(0,N,N);  tadj=0
```

Continued.

```
for (i in 1:N) { for (j in 1:d[i]) {tadj=tadj+1
W[i,adj[tadj]]=1}}
# Deprivation Scores
IMD=c(38.3,25.4,24.9,32.9,41.1,35.8,35.8,37.6,25.7,20,38.5,39.3,42.9,37.2,28.6,
38.4,35, 33.7,33.4,41.7,41.3,33.6,31.8,24.5,41.5,21.2,30.2,14.3,11.2,13.5,11.9,
20.8,10.4,26.3,8.8,7.5,19, 22.1,9.3,3.7,15,12.3,11.4,11.2,4.7,8.7,18.7,17.9,28,
14,25.2,7.5,31.6,25.7,17,19.1,19.3, 11.1,17.9,17.8,23,11.8,20.5,12.8,14.3,
17.3,11.1,20.8,18.8,23.6,18.1,25.4,20.3,19.7,31.3, 28.6,20.5,34.3,34.2,
36.1,25.2,18.6,12.1)
# Populations (males 65-69)
T=c(107,151,103,128,99,142,90,94,122,150,98,139,98,37,96,103,89,118,109,63,42,
148,127,163,134,169,128,205,200,219,194,134,201,130,224,200,131,163,189,167,
172,160,162,178, 240,188,169,177,177,193,92,269,129,88,114,100,263,185,121,
195,102,121,115,214,88, 123,98,149,121,80,116,106,83,114,96,140,205,80,68,
133,149,237,251)
# Compare area to its locality on Deprivation Score (W[i] in eqn 5)
lw <- nb2listw(MSOAnb, style="W", zero.policy=TRUE)
# Average IMD in locality surrounding each area
IMD.lag <- lag.listw(lw, IMD)
absdif.IMD=abs(IMD-IMD.lag)
# Specifications
m.lambda=lgt.lambda=lambda=stats=c()
Qdiag =numeric(N); Q = COV = matrix(,N,N)
# preset parameters
gamma=c(0,-1.5); alpha=-0.5; beta=0.03; tau.s = 20; sd.lambda =1
# Sampling Loop
NITER=100
y=stats=matrix(,N,NITER)
for (iter in 1:NITER) {
for (i in 1:N) {# regression for lambda using risk factor discrepancy
m.lambda[i] = gamma[1]+gamma[2]*(absdif.IMD[i]-mean(absdif.IMD[]))
lgt.lambda[i] =rnorm(1,m.lambda[i],sd.lambda)
lambda[i] = exp(lgt.lambda[i])/(1+exp(lgt.lambda[i]))}
# simulate spatial effects, Q is precision matrix
for(i in 1:N) { Qdiag[i] = tau.s*(1-lambda[i]+lambda[i]*d[i])}
for(i in 1:N) { for (j in 1:N) {
Q[i,j] = (i==j)*Qdiag[i]-(1-(i==j))*tau.s*lambda[i]*lambda[j]*W[i,j]}}
COV = solve(Q)
s = rmvnorm(1, mean = rep(0, nrow(COV)), sigma=COV, method=c("svd"))
# regression model including effect of Deprivation Score (IMD)
eta = alpha+beta*(IMD-mean(IMD))+s
# MSOA Illness probability
p = exp(eta)/(1+exp(eta))
# simulated illness counts
y[,iter] =rbinom(N,T,p)
# assessment criteria
stats[1,iter] =mean(lambda)
stats[2,iter] =sum(lambda < 0.5)
stats[3,iter] = max(p)/min(p)}
```

Overweight model, similarity matrix

```
#Create similarity matrix
n<-length(base_models$PcPc)
matrix<-matrix(NA,nrow=n,ncol=n)
for (i in 1:n){
  for (k in i:n){
    matrix[i,k]<-abs(base_models$PcPc[i]
                     -base_models$PcPc[k])
  }
}
Sim<-forceSymmetric(matrix)
Sim<-as.matrix(Sim)
zeta<- -mean(Sim)/log(0.01) # Best et al (1999), p. 136
S<-exp(-Sim/zeta)
S1<-S
S0<-ifelse(S==1,0,S)

for (i in 1:n){
  S[i,i]<-mean(S[i,-i])
}
head(S)
rowSums(S)

#run model

model_overweight<- S.CARleroux(formula = ceiling(overweigh_cases/100)~1,
family = "binomial", trials = ceiling(population/100), data = base_models, W=S,
burnin=10000, n.sample=100000, thin=10)
```

References

Banerjee, S., B. P. Carlin, and A. E. Gelfand. 2014. Hierarchical Modeling and Analysis for Spatial Data. , 2nd ed. Boca Raton: Chapman&Hall/CRC.

Baptista, H., J. M. Mendes, Y. C. MacNab, M. Xavier, and J. M. C. de Almeida. 2016. "A Guassian Random Field Model for Similarity-Based Smoothing in Bayesian Disease Mapping." *Statistical Methods in Medical Research* 25: 1166–1184.

Barnett, K., S. W. Mercer, M. Norbury, G. Watt, S. Wyke, and B. Guthrie. 2012. "Epidemiology of Multimorbidity and Implications for Health Care, Research, and Medical Education: A Cross-Sectional Study." *The Lancet* 380: 37–43.

Besag, J. 1974. "Spatial Interaction and the Statistical Analysis of Lattice Systems." *Journal of the Royal Statistical Society* 36: 192–236.

Besag, J., and C. Kooperberg. 1995. "On Conditional and Intrinsic Autoregressions." *Biometrika* 82: 733–746.

Besag, J., J. York, and A. Mollié. 1991. "Bayesian Image Restoration, with Two Applications in Spatial Statistics (With Discussion)." *Annals of the Institute of Statistical Mathematics* 43: 1–59.

Best, N., R. Arnold, A. Thomas, L. Waller, and E. Conlon. 1999. "Bayesiann Models for Spatially Correlated Disease and Exposure Data." In Bayesian Statistics 6, edited by Bernardo, J., Berger, J., Dawid, A. and Smith, A., pp. 131–147. Oxford: Oxford Science Publications.

Best, N., S. Richardson, and A. Thomson. 2005. "A Comparison of Bayesian Spatial Models for Disease Mapping." *Statistical Methods in Medical Research* 14: 35–59.

Brook, D. 1964. "On the Distinction between the Conditional Probability and the Joint Probability Approaches in the Specification of Nearest-Neighbour Systems." *Biometrika* 51: 481–483.

Brooks, S., and A. Gelman. 1998. "General Methods for Monitoring Convergence of Iterative Simulations." *Journal of Computational & Graphical Statistics* 7: 434–455.

Cancer Research UK and National Cancer Intelligence Network. 2014. *Cancer by Deprivation in England: Incidence, 1996-2010, Mortality, 1997-2011*. Technical report. London, UK: NCIN.

Congdon, P. 2008. "A Spatially Adaptive Conditional Autoregressive Prior for Area Health Data." *Statistical Methodology* 5: 552–563.

Etxeberria, J., T. Goicoa, and M. D. Ugarte. 2018. "Joint Modelling of Brain Cancer Incidence and Mortality Using Bayesian Age- and Gender-specific Shared Component Models." *Stochastic Environmental Research and Risk Assessment* 32: 2951–1969.

Gathani, T., R. Ali, A. Balkwill, J. Green, G. Reeves, V. Beral, and K. A. Moser. 2014. "Ethnic Differences in Breast Cancer Incidence in England Are Due to Differences in Known Risk Factors for the Disease: Prospective Study." *British Journal of Cancer* 110: 224–229. URL.

Griffith, D. A. 1996. "Some Guidelines for Specifying the Geographic Weights Matrix Contained in Spatial Statistical Models." In Practical Handbook of Spatial Statistics, edited by Arlinghaus, S. L., pp. 65–82. Boca Raton: CRC Press.

INE. 2013. Estudo sobre o Poder de Compra Concelhio. Technical report. Lisbon: INE.

Lee, D., and R. Mitchell. 2013. "Locally Adaptive Spatial Smoothing Using Conditional Auto-Regressive Models." *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 62: 593–608.

Leroux, B. G., X. Lei, and N. Breslow (2000): ''Estimation of Disease Rates in Small Areas: A New Mixed Model for Spatial Dependence,'' In *Statistical Models in Epidemiology, the Environment, and Clinical Trials, The IMA Volumes in Mathematics and its Applications*, edited by M. E. Halloran and D. Berry, Vol. 116, New York, NY: Springer New York, pp. 179–191.

MacNab, Y. C. 2018. "Some Recent Work on Multivariate Gaussian Markov Random Fields." *Test* 27: 554–569.

Rodrigues, A., A. Sepriano, S. P. Gonçalves, A. M. Rodrigues, N. Gouveia, L. Pereira, M. Eusébio, and S. Ramiro. 2015. "EpiReumaPt-the Study of Rheumatic and Musculoskeletal Diseases in Portugal : A Detailed View of the Methodology EpiReumaPt – the Study of Rheumatic and Musculoskeletal Diseases in Portugal : a Detailed View of the Methodology," *Acta reumatologica portuguesa* 40: 110–124.

Salmasi, L., and M. Celidoni. 2017. "Investigating the Poverty-Obesity Paradox in Europe." *Economics and Human Biology* 26: 70–85.

Smith, T., J. Wakefield, and A. Dobra. 2015. "Restricted Covariance Priors with Applications in Spatial Statistics." *Bayesian Analysis* 10: 965.

Tosetti, E., R. Santos, F. Moscone, and G. Arbia. 2018. "The Spatial Dimension of Health Systems." In *Oxford Research Encyclopedia of Economics and Finance*. Oxford: Oxford University Press.

Vehtari, A., A. Gelman, and J. Gabry. 2016. "Practical Bayesian Model Evaluation Using Leave-One-Out Cross-Validation and WAIC." *Statistics and Computing*: 1–20.

Wakefield, J., and H. Lyons. 2010. "Spatial Aggregation and the Ecological Fallacy." In Handbook of Spatial Statistics, edited by Gelfand, A. E., Diggle, P. J., Fuentes, M. and Guttorp, P., pp. 541–58. Boca Raton: Taylor & Francis Group. chapter 30.

Wang, Y. C., K. McPherson, T. Marsh, S. L. Gortmaker, and M. Brown. 2011. "Health and Economic Burden of the Projected Obesity Trends in the USA and the UK." *The Lancet* 378: 815–825.

Żukiewicz-Sobczak, W., P. Wróblewska, J. Zwoliński, J. Chmielewska-Badora, P. Adamczuk, E. Krasowska, J. Zagórski, A. Oniszczuk, J. Piątek, and W. Silny. 2014. "Obesity and Poverty Paradox in Developed Countries." *Annals of Agricultural and Environmental Medicine* 21: 590–594.