REVSTAT – Statistical Journal
Volume 13, Number 1, March 2015, 79–101

ALCOHOL ABUSE DISORDER PREVALENCE AND ITS DISTRIBUTION ACROSS PORTUGAL. A DISEASE MAPPING APPROACH

Authors: Helena Baptista
– NOVA IMS, Universidade Nova de Lisboa, Lisboa, Portugal
d2011073@novaims.unl.pt

Jorge M. Mendes
– NOVA IMS, Universidade Nova de Lisboa, Lisboa, Portugal
jmm@novaims.unl.pt

José Caldas de Almeida
– Department of Mental Health, NOVA Medical School, Portugal
jcaldasalmeida@gmail.com

Miguel Xavier
– Department of Mental Health, NOVA Medical School, Portugal
migxavier@gmail.com

Received: October 2014 Revised: November 2014 Accepted: December 2014

Abstract:

- Disease mapping is linked to two other scientific areas: small area estimation and ecological-spatial regression. This paper reviews similarities and differences among them. Bayesian hierarchical models are typically used in this context, using a combination of covariate data and a set of spatial random effects to represent the risk surface. The random effects are typically modeled by a conditional autoregressive prior distribution, and a number of alternative specifications have been proposed in the literature. The four models assessed here are applied to a study on alcohol abuse in Portugal, using data collected by the World Mental Health Survey Initiative.

Key-Words:

- *alcohol abuse; Bayesian hierarchical models; disease mapping; generalized linear models; small area estimation.*

AMS Subject Classification:

- 62F15, 62J12, 62M30, 62P10.

1. INTRODUCTION

The availability of disease data in sets of non-overlapping and contiguous spatial areal units has increased over the last few decades. Concepts such as small area estimation (SAE), disease mapping (DM) and ecological-spatial regression (ESR) are linked and are used in the context of the analysis of this type of data.

The purpose of this work is two-fold; first we clarify those concepts, and second, after focusing on DM, we apply several models to Portuguese alcohol abuse disorder (AAD) data, collected by the World Mental Health Survey Initiative (WMHSI), as specified in [39]. Harmful use of alcohol was considered by the World Health Organization (WHO) as one of the world's leading risk factors for disease and disability ([38]).

The goal of DM is to estimate the spatial pattern in disease risk over a geographical region, so that small areas with elevated risk can be identified. This term was first used in [5]. It uses the spatial setting and assumes positive spatial correlation between observations, essentially 'borrowing' more information from neighboring areas than from areas far away, smoothing local rates toward local neighboring values ([37]).

The remainder of this paper is organized as follows. Section 2 introduces the Portuguese data, as well as some background information on AAD. Section 3 provides the DM definition highlighting the differences and common aspects among DM, SAE and ESR. Section 4 deals with the most common and widely used models for DM, providing some basic information on those, as well as some challenges and recent methodological advances. Section 5 contains the results of the models, reviewed in Section 4, applied to the data defined in Section 2. Finally, Section 6 contains a concluding discussion and areas of future work.

2. DATA

The WMHSI was administered at the households of a nationally representative sample of respondents, between October 2008 and December 2009. The target population for the survey was defined as the resident, non-institutionalized, Portuguese-speaking population of the Portuguese mainland, aged 18 or above, residing in permanent private dwellings. Details regarding the design, target population, sampling, tools, measures, fieldwork organization, procedures, and weighting are reported in detail elsewhere ([39]). This is a cross-sectional study, meaning that both disease cases and possible risk factors are collected at the same time. As reported in [36] that restricts the conclusions that can be drawn from

the models. It is not possible to establish causal relationships between disease cases and possible covariates.

Data collected by cross-sectional studies may have several types of biases. In the present case the possibility of selection bias is particularly evident, as only the non-institutionalized population and the population above 18 years of age was selected, and accordingly to the WHO ([38]) the alcohol consumption is rising between adolescents (13–18 years of age) and young adults. Therefore inferences can only be made on the study population and not on the global Portuguese population. Another possible common bias is the misclassification bias, *i.e.*, the incorrect assignment of a disease to the study participants. This type of bias may occur in studies like this one, because there is no intervention of a medical doctor during the questionnaire's self-administration. This problem seems to have been solved in France, Italy, Spain and United States of America, as [14] and [10] provide evidence that the diagnoses of substance abuse disorders identified by the questionnaire used in this initiative, the CIDI 3.0 (Composite International Diagnoses Interview) have generally good concordance with diagnoses based on blinded clinical reappraisal interviews. Unfortunately those tests have not been conducted in Portugal. Although the alcohol consumption and related disorders are very much connected with cultural aspects ([28]), we think that the performance while identifying the actual presence of disease has not been seriously affected.

According to the WHO ([38]) approximately 5.1% of the global burden of disease, and 5.9% of all deaths worldwide are attributable to alcohol consumption. Furthermore, harmful use of alcohol inflicts significant social and economic losses on individuals and society at large.

In accordance with the DSM-IV ([8]) criteria there are two possible diagnoses of alcohol disorders, the alcohol abuse disorder and the alcohol dependence disorder. In the six European countries ([1]) covered by the ESEMeD project[1], 5.2% of the respondents report a lifetime history of alcohol abuse and/or dependence disorders. In the WMHSI, lifetime and 12-month alcohol disorder diagnoses are provided. From the data collected in Portugal the prevalence rate of a lifetime history of alcohol abuse and/or dependence disorders is 10.0%, while the last 12-month prevalence rate is 1.6%; the lifetime prevalence rate of alcohol dependence disorder is 1.3%, while the last 12-month prevalence rate is 0.26%; the lifetime prevalence rate of alcohol abuse disorder is 8.7%, whereas the last 12-month prevalence rate is 1.3%. The high prevalence of alcohol abuse disorder found in Portugal reiterates the need to maintain alcohol abuse as a public health priority in the country, and therefore more detailed studies are needed.

[1]The ESEMeD Project was created to fully study the results of the WMHSI on the following countries: Belgium, France, Germany, Italy, the Netherlands and Spain. As Portugal joined the WMHSI later than others, most of the publications, including the [1], do not include Portuguese results.

The study region is mainland Portugal partitioned into 28 units called NUTS 3[2], corresponding to the 3rd level territorial units aggregation. There are 30 NUTS 3 in Portugal, from which 28 are in mainland Portugal and 2 are in the Islands. The *response variable* is the number of lifetime AAD cases per NUTS 3. Differences in the size and demographic structure of the population living in each NUTS 3 are accounted for by computing the expected number of AAD cases using indirect internal standardization, based on age specific AAD rates for the whole study region.

The age standardization process, as defined in [36], can be direct or indirect. The choice between direct and indirect standardization is usually defined by the type of data available. Age-specific rates for the disease at each NUTS 3 are not available and therefore the indirect method is used, by applying the age-specific disease rate for the global population to the NUTS 3 age-specific population, provided by the Portuguese Statistics organization for the year of 2008. As this standardization is done using the age-specific disease rate for the global population, as it was collected by the survey itself, the standardization is internal (external standardization only occurs when standard tables of age-specific rates for the disease are available). As mentioned in [2] internal standardization is 'cheating' in some sense, since 'a degree of freedom is lost' by estimating the age-specific disease rate from the current data.

Accordingly, the following notations and/or definitions are introduced:

a) Y_k the random variable representing the number of observed cases (y_k) in each k age group;

b) n_k representing the number of people at risk in each k age group;

c) $r_k = \frac{y_k}{n_k}$ representing the observed prevalence proportion for each k age group;

d) n_{ik} representing the number of people at risk in each k age group in the i^{th} NUTS 3;

e) E_{ik} and y_{ik} representing the expected and observed number of cases for the k age group in the i^{th} NUTS 3, respectively, where $E_{ik} = r_k n_{ik}$;

f) $E_i = \sum_k r_k n_{ik}$ and $y_i^\star = \sum_k y_{ik}$ representing the total number of expected and observed cases in the i^{th} NUTS 3, respectively;

g) $SMR_i = \frac{Y_i^\star}{E_i}$, the *standardized morbidity ratio*, representing the risk of each i^{th} NUTS 3. A value of SMR greater (less) than one indicates that the area i has a higher (lower) than average disease risk. If the $SMR_i = 1.15$, it can be said that area i has a 15% increased risk of the disease.

[2] *Nomenclatura Comum das Unidades Territoriais Estatísticas*, in Portuguese language as defined by Eurostat, the European statistical organization.

Figure 1 shows the raw SMR values for the 28 NUTS 3.

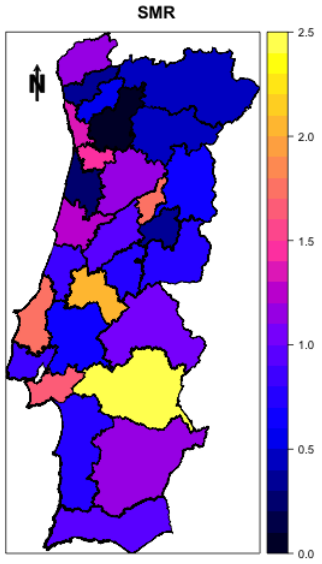

Figure 1: AAD Raw SMRs per NUTS 3. The four regions, which had originally missing values, are shown already with the imputed mean values resulting from the GLM (see Section 5).

Our illustrative example also considers two ecological covariates that are widely known as being associated with the AAD ([13, 18, 27, 33]), which are (a) proportion of population aged 18 to 34, (b) proportion of males. AAD is more prevalent in younger men. These data are only available per NUTS 3, for the year of 2011, as provided by the latest census conducted in Portugal, which we find to be temporally misaligned with WMHSI data used in this work. However as population age and gender structures do not significantly change in 3 years, no corrective measures have been implemented.

3. SMALL AREA ESTIMATION, DISEASE MAPPING AND ECOLOGICAL-SPATIAL REGRESSION

DM joins together three different disciplines: statistics/biostatistics, epidemiology and geography. DM focuses on the challenge of obtaining reliable statistical estimates (statistics/biostatistics) of local disease risk based on counts of observed cases (epidemiology) within small administrative districts or regions (geography) coupled with potentially relevant background information. DM goals are twofold: obtain statistically precise local estimates of disease risk for each region and maintain the regions 'small' in order to keep the geographic resolution.

The areas are not only small in size (relative to the area of the full spatial domain of interest), but are also small in terms of local sample size, resulting in deteriorated local statistical precision. To solve this problem the classical design-based solutions are often infeasible since the local sample sizes within each region, required for the desired level of statistical precision, are often unavailable or unattainable. The model-based approaches can help overcome this problem by the mechanism of 'borrowing strength' across small areas to improve local estimates.

3.1. DM as a special case of SAE

Nowadays sample survey data are extensively used to provide reliable direct estimates of parameters of interest for the whole population. When it comes to getting the same estimates for domains of that population, and due to the small sample sizes in those domains, direct survey estimates are likely to yield unacceptably large standard errors. This makes it necessary to combine survey data collected from the small areas with auxiliary information from sources external to the survey. In this context, named as SAE, several indirect estimators have been extensively used. Some of the most common are the traditional indirect estimators based on implicit models, which include synthetic and composite estimators, and the Empirical Best Linear Unbiased Prediction approach. Most of these approaches also consider a contiguity matrix that describes the neighborhood structure between small areas 'borrowing strength' from related areas to find more accurate estimates for a given area. The works [29] and [6] provide respectively an overview of the foundations of SAE and a comparison between several traditional estimators and some proposed estimators using a Monte Carlo simulation.

DM is a special case of SAE, since the goal is to find reliable statistical estimates of local disease risk. As mentioned by [37] DM refers to a collection of methods extending SAE to directly utilize the spatial setting and assumed positive spatial correlation between observations. The data used are aggregated or averaged values at the small area level, representing disease incidence, prevalence or mortality rates, frequently not coming from surveys but coming from counts of disease cases from hospital admissions ([21, 24]), counts of cancer cases or cancer deaths ([3, 16, 34]), and mortality data ([7, 24, 25]). In the present work we use counts of disease cases from a survey.

3.2. DM and ESR apply the same methodologies to reach different goals

By combining data from administrative registries and/or surveys with auxiliary data, DM goal is to predict area-level outcome summaries, to identify areas

of elevated risk. ESR uses the same type of data and the same methodologies but its objective is the estimation of associations between covariates and the disease cases.

Therefore, two common problems found in ESR are not of a concern in DM: (a) ecological bias and (b) the inclusion of spatially correlated errors changing the association between disease cases and fixed effects.

Ecological bias is the difference between estimated associations on ecological- and individual-level data ([35]). Data used in DM and ESR, both for the number of cases and for the covariates are found rarely at individual-level, mainly due to confidentiality reasons, and therefore the association found at the aggregated level might not be the same if we would have used individual-level data. Aggregated data is usually designated as areal data ([2]). The objective of DM is not to estimate the associations between the cases and the covariates or to improve predictions, and therefore ecological bias is not a concern (for more details on the subject see [35]).

The inclusion of spatially correlated errors, changing the association between disease cases and fixed effects, has been studied by [34] and [12]. Often the study of ESR has provided estimates of the fixed-effect coefficients substantially different from those of ecological regressions. ESR is an ecological regression augmented with the inclusion of random effects modeled by a globally smooth conditional autoregressive model. If the covariates are also globally smooth, collinearity problems might change dramatically the coefficients of the fixed-effects. As before, the coefficients of association are not of direct interest in DM, and therefore this aspect is not a concern.

4. DISEASE MAPPING MODELS

DM methodologies are explained in [37] and [2]. DM methodologies for areal data are usually divided in frequentist methods and hierarchical Bayesian models [2]. To provide a wide comparison of methods, [15] presents some preliminary results concerning the goodness-of-fit of a variety of DM models applied to simulated disease incidence data. These simulated models cover simple risk gradients and more complex true risk structures, including spatial correlation. Authors conclude that full Bayesian hierarchical models are the most robust across a range of diverse models. A number of hierarchical Bayesian models have been proposed in the literature, including the following two, which have been widely used: a) the model developed by Besag, York and Mollié ([3]), from now on designated as BYM model and b) the model developed by Leroux, Lei and Breslow ([22]), from now on designated as LLB model. These two models will be used in Section 5.

Authors of [4] review the main classes of Bayesian models, among which the BYM model is included (but not the LLB model) and conclude that the BYM model has good properties for modeling a single disease and 'appears to be the only fully Bayesian spatial model to have been used in published applications of disease mapping outside of the statistical literature' (page 57). Recently, [24] and [16] published comparisons between hierarchical Bayesian models and both conclude that the LLB model is the best overall, because it produces consistently good results across a range of spatial correlation scenarios, is more parsimonious on parameters, and has less undesirable features (this subject will be further developed in Subsection 4.1).

One of the challenges posed at the DM level arises from its basic goal, the smoothing of local rates toward local neighboring values. When real discontinuities exist between neighboring areas, the models will lead to oversmoothing blurring the edges, which may not be appropriate. If the goal is to identify boundaries or regions of rapid change, the methods of *boundary analysis* or *wombling* need to be applied. For more detail see the recent works of [19] and [20].

A general formulation for the first level of the hierarchical Bayesian models used in DM is given by

$$Y_i|E_i, R_i \sim Poisson(E_i R_i) \qquad \text{for} \quad i = 1, ..., n \ ,$$

$$\ln(R_i) = \mu + \mathbf{x_i^T}\boldsymbol{\beta} + \phi_i \ . \tag{4.1}$$

If E_i is not too large (as it is the case of rare diseases) or the regions i are sufficiently small, the usual model for the Y_i is a Poisson model ([2]). In the model, R_i denotes the risk of disease in area i, which is modeled by an intercept term μ, a set of p covariates $\mathbf{x_i^T} = (x_{i1}, ..., x_{ip})$ and a random effect ϕ_i. The random effects are included to model any overdispersion and/or spatial correlation that might remain in the data after have being accounted for by the included covariate information. Most studies of this type show overdispersion, meaning that $Var[Y_i] > \mathbb{E}[Y_i]$, which has several possible causes: subject heterogeneity; correlation between individual responses; omitted unobserved variables; and/or excess zero counts. Inference for this type of model is based on Markov chain Monte-Carlo (MCMC) simulation, using a combination of Gibbs sampling and Metropolis-Hastings steps and more recently using Integrated nested Laplace approximations ([31]).

The random effects $\boldsymbol{\phi} = (\phi_1, ..., \phi_n)$ are usually modeled by the class ([2]) of conditional autoregressive (CAR) prior distributions, which are a type of Markov random field model ([11]). Instead of a specification of a single multivariate distribution $f(\boldsymbol{\phi})$, the above models are specified by a set of univariate full conditional distributions $f(\phi_i|\boldsymbol{\phi}_{-i})$, where $\boldsymbol{\phi}_{-i} = (\phi_1, ..., \phi_{i-1}, \phi_{i+1}, ..., \phi_n)$. To determine the spatial correlation between the random effects, we use the neighborhood matrix

W, which is a binary $n \times n$ matrix, with elements w_{ji}:

$$w_{ji} = \begin{cases} 1, & \text{if } j \sim i\,, \\ 0, & \text{otherwise}\,, \end{cases}$$

where $j \sim i$ represents contiguous areas, and therefore j and i are considered neighbors. Other *adjacency-based* weights are available but are much less widely applied ([37]). If two areas are neighbors we believe their random effects are correlated, while non-neighboring areas are modeled as being conditionally independent given the remaining elements of $\boldsymbol{\phi}$.

4.1. BYM model

The BYM model combines the intrinsic CAR (ICAR) with an additional set of independent random effects.

The full conditional distributions of ICAR, as proposed by [3] are given by

$$u_i|\boldsymbol{u}_{-i}, \sigma^2 \sim N\left(\frac{1}{n_i}\sum_{j\sim i} u_j, \frac{\sigma^2}{n_i}\right). \tag{4.2}$$

The conditional expectation of u_i is equal to the mean of the random effects in neighborhood areas, while the conditional variance is inversely proportional to the number of neighbors n_i. The variance parameter σ^2 controls the amount of variation between the random effects. The ICAR model has three main drawbacks:

1) Its simplicity turns it into a very restrictive prior. Its single parameter does not determine the strength of the spatial correlation (for example multiplying each u_i by 10, will only increase σ^2 leaving the spatial correlation unchanged). If data are weakly correlated, the ICAR is not the most appropriate model ([16]).

2) The joint distribution for $f(\mathbf{u})$ corresponding to (4.2) is improper (it does not determine a legitimate probability distribution, one that integrates to 1). Nevertheless, this is easily solved by enforcing a constrain such as, $\sum_{j=1}^{n} u_j = 0$, which can be *numerically* imposed by recentering each sampled **u** vector around its own mean following each Gibbs iteration ([2]).

3) According to [24] the ICAR has an undesirable *global* (*i.e.* large-scale) property of tending to a negative pair-wise risk dependence as the 'spatial proximity' of the two regions is further apart.

The BYM model defines ϕ in (4.1) by

$$\phi_i = \theta_i + \psi_i \ , \tag{4.3}$$

$$\theta_i|\sigma_\theta^2 \sim N(0, \sigma_\theta^2) \ ,$$

$$\boldsymbol{\psi} = (\psi_1, ..., \psi_n)|\mathbf{W} \ , \qquad \sigma_\psi^2 \sim ICAR(\mathbf{W}, \sigma_\psi^2) \ ,$$

where $\mathbf{W}$ is defined in Section 4). More details on the BYM model are provided by [21] and [3].

The set of random effects $\boldsymbol{\theta} = (\theta_1, ..., \theta_n)$ is independent between areas. Different strengths of spatial correlation can be represented by varying the relative sizes of the two components $(\boldsymbol{\theta}, \boldsymbol{\psi})$. In practice, it will often be the case that either $\boldsymbol{\theta}$ or $\boldsymbol{\psi}$ dominates the other depending upon the strength of the spatial structure and the relative sizes of σ_θ^2 and the σ_ψ^2. This flexibility is also a disadvantage, as each data point is represented by two random effects while only their sum $(\theta_i + \psi_i)$ is identifiable. In order to attain model identification and achieve convergence when MCMC is used, at least one considerably informative hyper prior has to be assumed either for σ_θ^2 or σ_ψ^2. Several authors have studied this aspect ([37, 2]), and ([24]) implemented a model that can 'attain model identifiability, allow the data to inform risk decomposition, and facilitate principled attribution of the relative risk variability to spatially varying *clustering* effects and randomly varying *heterogeneity* effects based on the *given data*' (page 66), hereafter called Modified BYM (MBYM). This model replaces (4.3) by

$$\boldsymbol{\phi} = \sqrt{\lambda}\boldsymbol{\psi} + \sqrt{1-\lambda}\boldsymbol{\theta} \ , \qquad \boldsymbol{\psi} \perp \boldsymbol{\theta}, \quad \lambda \in (0, 1) \ . \tag{4.4}$$

One interpretation of the above is that it represents a re-parameterized BYM prior with $\sigma_\psi^2 = \lambda\sigma^2$ and $\sigma_\theta^2 = (1-\lambda)\sigma^2$. The new prior interpolates between the ICAR prior and the Gaussian prior for θ. λ serves as a spatial smoothing parameter and determines the proportion of the spatially structured risk variability over the total risk variability.

4.2. LLB model

The LLB model is based on a single set of random effects $\boldsymbol{\phi} = (\phi_1, ..., \phi_n)$, represented by a multivariate Gaussian distribution

$$\boldsymbol{\phi}|\mathbf{W}, \sigma^2, \rho, \boldsymbol{\mu} \sim N(\boldsymbol{\mu}, \sigma^2[\rho\mathbf{W}^* + (1-\rho)I_n]^{-1}) \ . \tag{4.5}$$

The prior above has a constant non-zero mean $\boldsymbol{\mu} = (\mu, ..., \mu)$, avoiding the use of the intercept term in (4.1). In the matrix, $\sigma^2[\rho\mathbf{W}^* + (1-\rho)I_n]^{-1}$, I_n is an

$n \times n$ identity matrix and the elements of $\mathbf{W}^*$ are equal to

$$w_{ji}^* = \begin{cases} n_i, & \text{if } j = i\,, \\ -1, & \text{if } j \sim i\,, \\ 0, & \text{otherwise}\,. \end{cases}$$

The precision matrix is a weighted average of the spatially dependent correlation structures, represented by the matrix $\mathbf{W}^*$, the independent correlation structures, represented by the identity matrix, and the weight represented by the parameter ρ. When $\rho = 0$ the model becomes a simple independent random effects model and when $\rho = 1$ the model becomes the ICAR as in (4.1). When $0 \leqslant \rho < 1$ the joint distribution (4.5) is proper. The full conditional distributions corresponding to (4.5) are given by

$$\phi_i|\boldsymbol{\phi}_{-i}, \mathbf{W}, \sigma^2, \rho, \mu \sim N\left(\frac{\rho\sum_{j\sim i}\phi_j + (1-\rho)\mu}{n_i\rho + 1 - \rho}, \frac{\sigma^2}{n_i\rho + 1 - \rho}\right). \tag{4.6}$$

The conditional expectation is the weighted average of the random effects in the neighboring areas and the overall mean μ. The conditional variance, in the presence of strong spatial correlation is approximately σ^2/n_i, the same as the ICAR, but if the random effects are independent then it is a constant (σ^2).

4.3. Localized conditional autoregressive model

All three models defined above use CAR priors that are globally smooth. The random effects are forced to exhibit a single global level of spatial smoothness determined only by geographical adjacency. With real data such a uniform level of smoothness for the entire region is unrealistic. It is more realistic to think that sub-areas of spatial autocorrelation co-exist with areas of discontinuity. As an example, areas of wealth and poverty, sharing boundaries, are very common in the biggest cities of the world, showing different patterns in the disease risk. A possible solution to this problem is presented by [21], and is called Bayesian localized conditional autoregressive model, LCAR from now on. This model was initially applied to a ESR, but as explained in Subsection 3.2 the same methodology can be applied in the DM field.

The LCAR treats the elements in the neighborhood matrix, representing contiguous areas, as a set of binary random quantities and not as fixed values. The elements of this new neighborhood matrix, $\tilde{\boldsymbol{W}}$, continue to be set to zero for non adjacent areas but adjacency is no longer the only reason for those elements to be set to one. When all adjacencies are kept, the model simplifies to the ICAR, while if all adjacencies are removed the random effects are independent.

The model defines $\boldsymbol{\phi}$ in (4.2) as $\tilde{\boldsymbol{\phi}} = (\boldsymbol{\phi}, \phi_{\circledast})$ where $\phi_{\circledast}$ is a global random effect that is potentially common to all areas and prevents any unit from having no information to 'borrow strength' from. Based on the extended matrix, the proposal is to model $\tilde{\boldsymbol{\phi}}$ as $\tilde{\boldsymbol{\phi}} \sim N(\mathbf{0}, \sigma^2 \mathbf{Q}(\tilde{\mathbf{W}}, \epsilon)^{-1})$, with the precision matrix given by

$$\mathbf{Q}(\tilde{\boldsymbol{W}}, \epsilon) = \text{diag}(\tilde{\boldsymbol{W}}\boldsymbol{I}) - \tilde{\boldsymbol{W}} + \epsilon \boldsymbol{I} \ , \tag{4.7}$$

The component $\text{diag}(\tilde{\boldsymbol{W}}I) - \tilde{\boldsymbol{W}}$ corresponds to the ICAR model applied to the extended random effects vector $\tilde{\boldsymbol{\phi}}$ and the component ϵ ensures that the matrix is diagonally invertible. This restriction is now needed because $\mathbf{Q}(\tilde{\boldsymbol{W}})$ is no longer fixed. The parameter ϵ is recommended to be set as $\epsilon = 0.001$. The full conditional distributions corresponding to the LCAR model are given by

$$\phi_j | \boldsymbol{\phi}_{-j} \sim N\left(\frac{\sum_{i=1}^{n} w_{ij}\phi_i + w_{i\circledast}\phi_{\circledast}}{\sum_{i=1}^{n} w_{ij} + w_{i\circledast} + \epsilon}, \frac{\sigma^2}{\sum_{i=1}^{n} w_{ij} + w_{i\circledast} + \epsilon} \right), \quad j = 1, ..., n \ , \tag{4.8}$$

$$\phi_{\circledast} | \boldsymbol{\phi}_{-\circledast} \sim N\left(\frac{\sum_{i=1}^{n} w_{i\circledast}\phi_{\circledast}}{\sum_{i=1}^{n} w_{i\circledast} + \epsilon}, \frac{\sigma^2}{\sum_{i=1}^{n} w_{i\circledast} + \epsilon} \right).$$

In (4.8) the conditional expectation is a weighted average of the global random effect $\phi_{\circledast}$ and the random effects in the neighboring areas, with the binary weights depending on the current value of $\tilde{\mathbf{W}}$. The conditional variance is approximately (due to ϵ) inversely proportional to the number of neighbors remaining in the model, including the global random effect $\phi_{\circledast}$.

The matrix $\tilde{\boldsymbol{W}}$ is treated by the LCAR model as a single random quantity, which avoids several problems identified by other authors (for more details see [21], Subsection 3.2). The authors propose eliciting the set of candidate values of $\tilde{\boldsymbol{W}}$ from data having a similar spatial structure as the response variable.

The increased flexibility provided by the LCAR model inevitably means that it is more computationally demanding than the common BYM model.

5. ALCOHOL ABUSE DISORDER DISTRIBUTION ACROSS PORTUGAL

The number of lifetime AAD cases vary between 2679 (16A – Cova da Beira) and 136789 (171 – Grande Lisboa). There are four NUTS 3 (164 – Pinhal Interior Norte, 166 – Pinhal Interior Sul, 169 – Beira Interior Sul and 181 – Alentejo Litoral) where no cases were identified. The national nature of the survey sampling design creates situations where very small or even zero samples at the NUTS 3 level occur. In this situation it might happen that no cases are estimated, which does not mean that no disease diagnoses exist. Therefore, these

areas are treated as having missing values and not as having a null number of cases. The first level of the Bayesian hierarchical model, as seen in (4.1), involves complex calculations, very difficult to run on such numbers, therefore numbers of cases per 100 inhabitants, as well as expected number of cases per 100 inhabitants are used (this change does not eliminate the need of using the expected number of cases because only the size of the population is accounted for, not the structure).

The R software (version 3.1.1), with the package **CARBayes** ([17]) is used to fit the hierarchical models. The main advantages of this package are: (1) the spatial adjacency information is easy to specify as a binary neighborhood matrix; (2) given the neighborhood matrix the models can be implemented by a single function call in R; (3) maps with the disease risk estimates can easily be produced. The package has predefined the following models that will be used: BYM, LLB and LCAR. By running the same model on R and on the BUGS software ([23]) the package's author shows that there is good agreement between the two sets of point estimates, as we confirm in the present work. One disadvantage of the package is that it cannot handle missing values at the response variable level. To overcome this, a Generalized Linear model (GLM), Poisson (quasi-likelihood) model ([26]), is fitted using as response variable the number of lifetime observed cases per NUTS 3 and as covariates the ecological variables defined before, namely the proportion of men and the proportion of population aged 18 to 34. The mean estimated number of lifetime observed cases achieved for the four areas with missing data are incorporated in the response variable vector $\mathbf{Y}$. This methodology is debatable and more work needs to be done, in order to evaluate all possible consequences of this approach.

The MBYM model is fitted using the OpenBUGS software ([23]). Even though the Bayesian methodology could handle the missing values, for comparison purposes the missing values are also replaced by the mean estimated values.

As mentioned in Subsection 4.3, the authors of LCAR propose that, for the elicitation of $\tilde{\boldsymbol{W}}$, data having a similar spatial structure as the response variable should be used. In their case, the prior elicitation was based on response variable data from previous years. Our decision was to use the number of cases of four other related mental disorders, chosen as follows. Disorders considered in the Portuguese version of the WMHSI include a comprehensive range of mental disorders, and a GLM is fitted (Binomial model) to understand which mental disorders are most commonly present with AAD. The response variable is two-level categorical, taking value one if the individual suffers from AAD and taking value zero otherwise, and the covariates are of the same type, for all other disorders. At a lower than 5% significance level, the following disorders have an Odds Ratio larger than one: Alcohol Dependence, Oppositional Defiant Disorder, Hypomania, and Intermittent Explosive Disorder. In the cases where values are missing the procedure followed is the one defined before, using as covariates the remaining disorders. For example, for alcohol dependence disorder as response variable, the

covariates are: alcohol abuse disorder, oppositional defiant disorder, hypomania and intermittent explosive disorder. The mean estimated number of cases are imputed in the response variable vectors. There are two reasons to use a different approach in the present case. First, in Portugal, data on ADD from previous surveys is not available. Second, this work is on DM and not on ESR, therefore the decision is to use data from related mental disorders.

5.1. Hyperpriors

Table 1 shows the prior distributions implemented in the four models. In the LCAR model, on top of the already mentioned information for the $\tilde{\boldsymbol{W}}$ matrix, the parameter ϵ is set to 0.001.

Table 1: Prior distributions for the models.

Model	Parameter	Prior Distribution	Mean/Shape	Variance/Scale
BYM	$\boldsymbol{\beta} = (\beta_1, \beta_2)$	Gaussian	0	1000
	μ	Gaussian	0	1000
	σ_θ^2 and σ_ψ^2	Inverse-Gamma	0.001	0.001
MBYM	$\boldsymbol{\beta} = (\beta_1, \beta_2)$	Gaussian	0	100000
	μ	Flat	-	-
	σ^2	Inverse-Gamma	0.001	0.001
	λ	Uniform [0,1)	0.5	0.5
LLB	$\boldsymbol{\beta} = (\beta_1, \beta_2)$	Gaussian	0	1000
	σ^2	Inverse-Gamma	0.001	0.001
	ρ	Uniform [0,1)	0.5	0.5
LCAR	$\boldsymbol{\beta} = (\beta_1, \beta_2)$	Gaussian	0	1000
	σ^2	Uniform [0,1000)	500	500

5.2. Inference

Posterior inference for all models is based on Markov Chain Monte-Carlo simulation, using a combination of Gibbs sampling and Metropolis-Hastings algorithms. Posterior inference is based on 8000 MCMC samples, which are obtained by running one chain for 100000 samples, by which convergence is assumed to have occurred. We ignore the first 20000 samples as burn-in, and use the remaining 80000 subsequent samples to obtain the posterior distributions of the parameters of interest (a thin of 10 is used to reduce the autocorrelation).

Pilot runs are carried out to establish appropriate burn-in using the Geweke's diagnostic ([9]). Convergence is assessed by visually monitoring the trace and the posterior density plot for each of the parameters.

5.3. Results

Each model is assessed by the resulting Deviance Information Criterion (DIC) ([32]), where a smaller value represents a better fitting model. Table 2 shows the results of the four models.

Table 2: DIC results, which include the effective number of parameters in the model (p.D.).

	BYM	**MBYM**	**LLB**	**LCAR**
DIC	155.3	145.0	159.2	158.0
p.D.	14.3	5.8	18.5	19.5

Table 2 shows that, according to DIC, the MBYM model exhibits the best fit. BYM model is the second best. Following [24], $\lambda = 1$ represents spatial/local smoothing and $\lambda = 0$ represents non-spatial/local smoothing, based on the disease mapping data at hand. In the MBYM the posterior mean value of $\lambda = 0.58$, shows that the data has an higher spatially structured variance than unstructured variance. As already proved by [16], the BYM model shows more robust results in the presence of strong spatial correlation structures, as it seems to be the case here.

Figure 2 shows the posterior median SMR values for the 28 NUTS 3, produced by the MBYM model. Table 3 shows summary measures of the marginal posterior of the parameters of interest obtained by the MBYM model.

Figure 3 displays histograms of the (a) raw SMR and the (b to d) smooth posterior median SMR values for the 28 NUTS 3, produced by the models. The concentration around the interval $[0.5, 1.5]$ on the latter can clearly be seen. Mapping the raw SMRs gives a misleading picture of the risk pattern, whereas any of the four models (plus LLB, which is not presented, but shows the same overall results) give posterior median relative risks less dispersed. This ability of the Bayesian models to "clean" adequately the SMRs from the false patterns created by the Poisson noise had been already referred by [30].

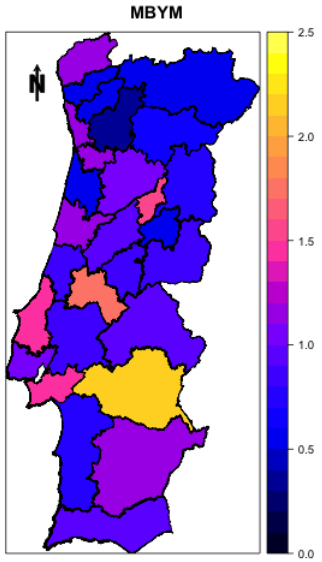

Figure 2: MBYM AAD posterior median SMRs per NUTS 3.

Table 3: MBYM model parameters summary.

Parameter	**Prior distribution**	**Prior mean**	**Prior std**	**MCMC Posterior mean (std)**	**2.5%**	**Median**	**97.5%**
β_0	Flat	0		−0.11 (0.10)	−0.32	−0.11	0.08
β_1	N(0, 100000)	0	100000	−0.23 (0.14)	−0.52	−0.22	0.06
β_2	N(0, 100000)	0	100000	−0.8 (0.13)	−0.34	−0.07	0.18
λ	U[0,1)	0.5	0.5	0.58 (0.25)	0.07	0.61	0.97
σ^2	IG(0.001, 0.001)	1	10	0.61 (0.17)	0.35	0.59	1

The LCAR model is the only one that does not have a single global level of smoothness and therefore any existing discontinuities in the risk pattern can only be concluded from this model. There are 122 neighborhoods (or connections) between the 28 NUTS 3. When applying the LCAR model, the 95% credibility interval of the number of removed connections is [2, 54]. This fact provides evidence that there is information in the data to estimate the number of connections to be removed. Results confirm the known deep cultural roots in the country on the differences between the coast- and the country-side NUTS 3. This is the case of Península de Setúbal and Algarve, two coast-side NUTS 3 sharing physi-

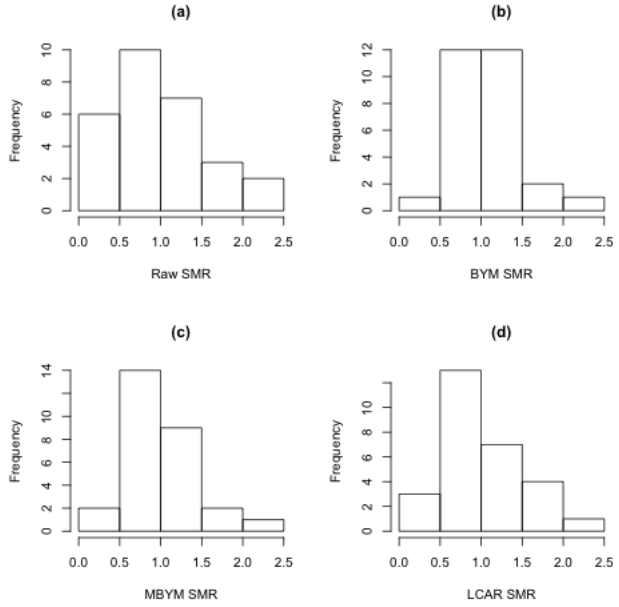

Figure 3: Histograms of the (a) raw SMRs and posterior medians of the (b,c,d) SMRs, for all areas derived by each of the three models, (b) BYM, (c) MBYM and (d) LCAR.

cal borders with the country-side NUTS 3 Alentejo, which are no longer present when data is used to estimate connections.

As mentioned in Subsection 3.2 the goal of DM is not the estimation of associations between covariates and the disease cases, but is to estimate the pattern of disease risk over a geographical region. Nevertheless, due to the fact that the two coefficients (β_1 and β_2) did not show to be significantly different from zero (contrary to expectations mentioned in Section 2), one must remember that this is an ecological study design, and the results must not be interpreted in terms of individual level cause and effect. One possible explanation is ecological bias as the prevalence rate of AAD is higher in younger men. Another possible explanation is that both the random and the covariate effects are confounded, because both are globally smooth in the MBYM model.

6. DISCUSSION

In the past years hierarchical Bayesian models have been developed and refined to achieve statistically precise local estimates of disease risk for each small region. In this study four of those models are assessed and used to estimate the disease risk of AAD at the NUTS 3 level, in Portugal.

In terms of DIC, the MBYM model achieves the best results. The MBYM model derives from the BYM model in an attempt to overcome the known deficiency of the latter, the lack of identifiability. The MBYM is identifiable and facilitates hierarchical modeling of the additive effects of unobserved covariates that might be spatially and randomly varying ([24]). In the present case its superior performance is likely to result from the BYM (and MBYM) model ability of achieving the best results in cases when the spatial correlation structure is strong, as seems to be this case.

The LLB model has consistently shown good results across a variety of cases but in this study, in terms of DIC, it proves to be the most poorly performing. While other authors show that the LLB model is the one achieving the best results ([16, 24]), our study shows otherwise. The performance of each model will depend on the type of data at hand, and none can be defined as the 'gold standard' over others.

The LCAR model is the only model that does not take the neighborhoods as fixed but those emerge from real data, as a random quantity. By doing that, in this example, the known cultural differences (between country- and coast-side) in the country are confirmed.

This study has some particularities when compared with the majority of the published applications:

a) The data use emerged from a survey, which was not plan to have local (at NUTS 3 level) samples with the proper size to allow designed-based estimation, and therefore presents some missing values. To overcome this a frequentist model is used.

b) The complex computations of the first level of the hierarchical Bayesian models do not allow the direct use of the survey estimates. To overcome this the number of lifetime cases of AAD per 100 inhabitants is used.

c) The LCAR model is used as a DM and not as a ESR, and therefore the type of data used for the elicitation of the $\tilde{\mathbf{W}}$ matrix is not previous periods data for the same disease but data from correlated disorders.

The epidemiological study presented in this paper shows substantial evidence of some 'hot spots' in the Center and South of the country allowing the authorities to focus interventions on these 'excess risk' areas.

There are still many opportunities for future work in this area. First the global ICAR's property of tending to negative pair-wise risk dependence as the 'spatial proximity' between two regions is further apart and its potential impact on posterior inference has not been yet sufficiently explored and understood.

Second [7] showed that region effects can be greater (smaller) for specific age groups. We know that AAD is more prevalent in young adult men ([13, 27]). Further research on the region effects on this age-gender group is needed. Third the four models used in this work were GLMM (Generalized linear mixed models), but the linear assumption on the covariate effects might be too restrictive, the usage of a GAMM (generalized additive mixed model) should be explored as it can eventually reveal non-linear relationships.

ACKNOWLEDGMENTS

The Portuguese Mental Health Survey Initiative is carried out in conjunction with the World Health Organization World Mental Health (WMH) Survey Initiative which is supported by the National Institute of Mental Health (NIMH; R01 MH070884), the John D. and Catherine T. MacArthur Foundation, the Pfizer Foundation, the US Public Health Service (R13-MH066849, R01-MH069864, and R01-DA016558), the Fogarty International Center (FIRCA R03-TW006481), the Pan American Health Organization, Eli Lilly and Company, Ortho-McNeil Pharmaceutical, GlaxoSmithKline, and Bristol-Myers Squibb. We thank the staff of the WMH Data Collection and Data Analysis Coordination Centers for assistance with instrumentation, fieldwork, and consultation on data analysis. None of the funder had any role in the design, analysis, interpretation of results, or preparation of this paper. A complete list of all within-country and cross-national WMH publications can be found at http://www.hcp.med.harvard.edu/wmh/.

The Portuguese Mental Health Study was carried out by the Department of Mental Health, Faculty of Medical Sciences, NOVA University of Lisbon, with collaboration of the Portuguese Catholic University, and was funded by Champalimaud Foundation, Gulbenkian Foundation, Foundation for Science and Technology (FCT) and Ministry of Health.

The authors would like to thank Vitor Lobo and Ales Popovic for their useful comments and suggestions.

The authors gratefully acknowledge the valuable comments and suggestions made by the two referees and the editor, all of which have greatly improved the focus and presentation of this work.

REFERENCES

[1] Alonso, J.; Angermeyer, M.C.; Bernert, S.; Bruffaerts, R.; Brugha, T.S.; Bryson, H.; de Girolamo, G.; Graaf, R.; Demyttenaere, K.; Gasquet, I.; Haro, J.M.; Katz, S.J.; Kessler, R.C.; Kovess, V.; Lépine, J.P.; Ormel, J,; Polidori, G.; Russo, L.J.; Vilagut, G.; Almansa, J.; Arbabzadeh-Bouchez, S.; Autonell, J.; Bernal, M.; Buist-Bouwman, M.A.; Codony, M.; Domingo-Salvany, A.; Ferrer, M.; Joo, S.S.; Martínez-Alonso, M.; Matschinger, H.; Mazzi, F.; Morgan, Z.; Morosini, P.; Palacín, C.; Romera, B.; Taub, N. and Vollebergh, W.A. (2004). Prevalence of mental disorders in Europe: results from the European Study of the Epidemiology of Mental Disorders (ESEMeD) project, *Acta Psychiatrica Scandinavica. Supplementum*, **420**, 21–7.

[2] Banerjee, S.; Carlin, B.P. and Gelfand, A.E. (2004). *Hierarchical Modeling and Analysis for Spatial Data*, Chapman&Hall/CRC, New York.

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

[4] Best, N.; Richardson, S. and Thomson, A. (2005). A comparison of Bayesian spatial models for disease mapping, *Statistical Methods in Medical Research*, **14**, 1, 35–59.

[5] Clayton, D. and Kaldor, J. (1987). Empirical Bayes estimates of age-standardized relative risks for use in disease mapping, *Biometrics*, **43**, 3, 671–81.

[6] Coelho, P.S. and Pereira, L.N. (2011). A spatial unit level model for small area estimation, *REVSTAT*, **9**, 2, 155–180.

[7] Dean, C.B.; Ugarte, M.D. and Militino, A.F. (2001). Detecting interaction between random region and fixed age effects in disease mapping, *Biometrics*, **57**, 1, 197–202.

[8] Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition: DSM-IV-TR (1994). American Psychiatric Association.

[9] Geweke, J. (1992). *Evaluating the accuracy of sampling–based approaches to the calculation of posterior moments.* In "Bayesian Statistics 4" (J.M. Bernardo; J.O. Berger; A.P. Dawid and A.F.M. Smith, Eds.), Oxford University Press, Oxford, 169–193.

[10] Haro, J.M.; Arbabzadeh-Bouchez, S.; Brugha, T.S.; de Girolamo, G.; Guyer, M.E.; Jin, R.; Lepine, J.P.; Mazzi, F.; Reneses, B.; Vilagut, G.; Sampson, N.A. and Kessler, R.C. (2006). Concordance of the Composite International with standardized clinical assessments in the WHO World Mental Health Surveys, *Archives of General Psychiatry*, **62**, 6, 593–602.

[11] Held, L. and Rue, H. (2010). *Conditional and intrinsic autoregressions.* In "Handbook of Spatial Statistics" (A.E. Gelfand; P.J. Diggle; M. Fuentes and P. Guttorp, Eds.), CRC Press, Boca Raton, 201–215.

[12] Hodges, J.C. and Reich, B.J. (2010). Adding Spatially-Correlated Errors Can Mess Up the Fixed Effect You Love, *The American Statistician*, **64**, 4, 325–334.

[13] Kalaydjian, A.; Swendsen, J.; Chiu, W.T.; Dierker, L.; Degenhardt, L.; Glantz, M.; Merikangas, K.R.; Sampson, N. and Kessler, R.C. (2009). Sociodemographic predictors of transitions across stages of alcohol use, disorders, and remission in the National Comorbidity Survey Replication, *Comprehensive Psychiatry*, **50**, 4, 299–306.

[14] Kessel, R.C.; Berglund, P.; Demler, O.; Jin, R.; Merikangas, K.R. and Walters, E.E. (2005). Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the National Comorbidity Survey Replication, *International Journal of Methods in Psychiatry Research*, **15**, 4, 167–180.

[15] Lawson, A.B.; Biggeri, A.B.; Boehning, D.; Lesaffre, E.; Viel, J.-F.; Clark, A.; Schlattmann, P. and Divino, F. (2000). Disease mapping models: an empirical evaluation, *Statistics in Medicine*, **19**, 9, 2217–1142.

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

[17] Lee, D. (2013). CARBayes: An R package for Bayesian spatial modeling with conditional autoregressive priors, *Journal os Statistical Software*, **55**, 13, 1–24.

[18] Lee, S.; Guo, W.J.; Tsang, A.; He, Y.L.; Huang, Y.Q.; Zhang, M.Y.; Liu, Z.R.; Shen, Y.C. and Kessler, R.C. (2009). Associations of cohort and socio-demographic correlates with transitions from alcohol use to disorders and remission in metropolitan China, *Addiction (Abingdon, England)*, **104**, 8, 1313–23.

[19] Lee, D. and Mitchell, R. (2012). Boundary detection in disease mapping studies, *Biostatistics (Oxford, England)*, **13**, 3, 415–426.

[20] Lee, D. and Mitchell, R. (2013). Locally adaptive spatial smoothing using conditional auto-regressive models, *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **62**, 4, 593–608.

[21] Lee, D.; Rushworth, A. and Sahu, S.K. (2014). A Bayesian localized conditional autoregressive model for estimating the health effects of air pollution, *Biometrics*, **70**, 2, 419–29.

[22] Leroux, B.G.; Lei, X. and Breslow, N. (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" (M.E. Halloran and D. Berry, Eds.), Springer, New York, 179–191.

[23] Lunn, D.; Spiegelhalter, D.; Thomas, A. and Best, N. (2009) The BUGS project: Evolution, critique, and future directions, *Statistics in Medicine*, **28**, 3049–3067.

[24] MacNab, Y.C. (2011). On Gaussian Markov random fields and Bayesian disease mapping, *Statistical Methods in Medical Research*, **20**, 1, 49–68.

[25] MacNab, Y.C. (2014). On identification in Bayesian disease mapping and ecological-spatial regression models, *Statistical Methods in Medical Research*, **23**, 2, 134–55.

[26] McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models*, Chapman & Hall/CRC, New York.

[27] Neumark, Y.D.; Lopez-Quintero, C.; Grinshpoon, A. and Levinson, D. (2007). Alcohol drinking patterns and prevalence of alcohol-abuse and dependence in the Israel National Health Survey, *The Israel Journal of Psychiatry and Related Sciences*, **44**, 2, 126–35.

[28] Neuroscience of Psychoactive Substance Use and Dependence (2004). World Health Organization, Geneva, **99**, 10, 1361–1362

[29] Rao, J.N.K. (2003). *Small Area Estimation*, Wiley, New York.

[30] Richardson, S.; Thomson, A.; Best, N. and Elliot, P. (2004). Interpreting posterior relative risk estimates in disease mapping studies, *Environmental Health Perspectives*, **112**, 9, 1016–1025.

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

[32] Spiegelhalter, D.J.; Best, N.; Carlin, B.P. and Van der Linde, A. (2002). Bayesian measures of model complexity and fit, *Journal of the Royal Society Statistical Society. Series B (Statistical Methodology)*, **64**, 4, 583–639.

[33] Swendsen, J.; Conway, K.P.; Degenhardt, L.; Dierker, L.; Glantz, M.; Jin, R.; Merikangas, K.R., Sampson, N. and Kessler, R.C. (2009). Socio-demographic risk factors for alcohol and drug dependence: the 10-year follow-up of the national comorbidity survey, *Addiction*, **104**, 8, 1346–1355.

[34] Wakefield, J. (2007). Disease mapping and spatial regression with count data, *Biostatistics (Oxford, England)*, **82**, 2, 158–83.

[35] Wakefield, J. and Lyons, H. (2010). *Spatial aggregation and the ecological fallacy*. In "Handbook of Spatial Statistics" (A.E. Gelfand; P.J. Diggle; M. Fuentes and P. Guttorp, Eds.), CRC Press, Boca Raton, 541–558.

[36] Waller, L.A. and Gotway, C.A. (2004). *Applied Spatial Statistics for Public Health Data*, John Wiley & Sons, Inc.

[37] Waller, L. and Carlin, B. (2010). *Disease mapping*. In "Handbook of Spatial Statistics" (A.E. Gelfand; P.J. Diggle; M. Fuentes and P. Guttorp, Eds.), CRC Press, Boca Raton, 217–244.

[38] WHO, S. A. T. in the D. of M. H. and S. A. of the W. H. O. (2014). Global Status Report on alcohol and health, *Retrieved from http://www.who.int*.

[39] Xavier, M.; Baptista, H.; Mendes, J.M.; Magalhães, J.M. and Caldas-de-Almeida, J. (2013). Implementing the World Mental Health Survey Initiative in Portugal - rationale, design and fieldwork procedures, *International Journal of Mental Health Systems*, **7**, 1, 19.