Online relative risks/rates estimation in spatial and spatio-temporal disease mapping

Aritz Adin[a,b], Tomás Goicoa[a,b], María Dolores Ugarte[a,b,*]

[a] *Department of Statistics, Computer Science and Mathematics, Public University of Navarre, Campus de Arrosadia, 31006 Pamplona, Spain*
[b] *Institute for Advanced Materials (InaMat), Public University of Navarre, Campus de Arrosadia, 31006 Pamplona, Spain*

Abstract

Background and objective: Spatial and spatio-temporal analyses of count data are crucial in epidemiology and other fields to unveil spatial and spatio-temporal patterns of incidence and/or mortality risks. However, fitting spatial and spatio-temporal models is not easy for non-expert users. The objective of this paper is to present an interactive and user-friendly web application (named SSTCDapp) for the analysis of spatial and spatio-temporal mortality or incidence data. Although SSTCDapp is simple to use, the underlying statistical theory is well founded and all key issues such as model identifiability, model selection, and several spatial priors and hyperpriors for sensitivity analyses are properly addressed.

Methods: The web application is designed to fit an extensive range of fairly complex spatio-temporal models to smooth the very often extremely variable standardized incidence/mortality risks or crude rates. The application is built with the R package shiny and relies on the well founded integrated nested Laplace approximation technique for model fitting and inference.

Results: The use of the web application is shown through the analysis of Spanish spatio-temporal breast cancer data. Different possibilities for the analysis regarding the type of model, model selection criteria, and a range of graphical

*Corresponding author
Email addresses: aritz.adin@unavarra.es (Aritz Adin), tomas.goicoa@unavarra.es (Tomás Goicoa), lola@unavarra.es (María Dolores Ugarte)

as well as numerical outputs are provided.

Conclusions: Unlike other software used in disease mapping, SSTCDapp facilitates the fit of complex statistical models to non-experts users without the need of installing any software in their own computers, since all the analyses and computations are made in a powerful remote server. In addition, a desktop version is also available to run the application locally in those cases in which data confidentiality is a serious issue.

Keywords: Areal data, disease mapping, R-INLA, shiny, small areas, spatio-temporal models

---

1. Introduction

Disease mapping deals with areal data from non-overlapping units focussing on the estimation of the spatial and spatio-temporal evolution of incidence or mortality patterns. The great variability inherent to classical estimation measures, such as standardized mortality/incidence ratios or crude rates, makes it necessary the use of models to smooth risks borrowing information from spatial and temporal neighbors. Research on spatial and spatio-temporal disease mapping has been focused on generalized linear mixed models within a general Bayesian framework. Two main approaches have been followed for model fitting and inference, the empirical Bayes and the fully Bayes approach. In the empirical Bayes approach, inference commonly relies on the well known penalized quasi-likelihood (PQL) technique, popularized by Breslow and Clayton [1]. The fully Bayes approach provides posterior marginal distributions of the target parameters instead of a single point estimate. However, the posterior distributions are not usually available in closed form and Markov chain Monte Carlo (MCMC) algorithms have to be used (see for example Gilks et al. [2]). WinBUGS [3] has traditionally been the most used software for Bayesian inference in disease mapping using MCMC. However, the new program Stan [4] is becoming popular for full Bayesian inference (see for example Carpenter et al. [5]). The

key difference with WinBUGS is that Stan's MCMC techniques are based on Hamiltonian Monte Carlo, which is more efficient and robust than traditional Gibbs sampling or Metropolis-Hastings algorithms. As MCMC methods are computationally intensive when dealing with complex models, a new approximate method for Bayesian inference in latent Gaussian models, a subclass of structured additive regression models which are suitable for many practical applications, has been developed by Rue et al. [6] as an alternative to MCMC. The method uses integrated nested Laplace approximations (INLA) and numerical integration to estimate the posterior marginal distributions of the quantities of interest. Since the latent fields of the model are assumed to be described by a Gaussian Markov Random Field [7], INLA uses numerical algorithms for sparse matrices to speed up computations in comparison with MCMC methods. See Rue et al. [6] or Blangiardo and Cameletti [8] for details about the approximate Bayesian inference strategy of INLA.

Different tools for Bayesian spatial and spatio-temporal modelling of areal count data are available for users. BayesX [9] is a free software written in C++ that uses numerically efficient sparse matrix architectures for estimating regression models with structured additive predictors. In addition, a fully interactive `R` interface to BayesX is available through the "R2BayesX" package [10]. Some other `R` packages have been also developed for areal (lattice) data such as the "surveillance" package [11] for modeling epidemic data, the "plm" [12] and "splm" [13] packages for modeling econometric spatial panel data, and both the "CARBayes" [14] and "CARBayesST" [15] packages to fit a range of spatial and spatio-temporal conditional autoregressive models respectively. Some examples on how to fit spatial disease mapping models using BayesX, CARBayes and INLA are given in Bivand et al. [16, Chapter 10.4].

The INLA approach for approximate Bayesian inference in latent Gaussian models has become a widespread tool in several fields of scientific research, mainly due to its great versatility, accuracy, and reduction of computational costs. An interface with the free statistical software `R`, called R-INLA, is available allowing model specification and fitting within `R`. Documentation for the

package, many worked examples, and a discussion forum are also available in the R-INLA website `http://www.r-inla.org/`. However, the generality of R-INLA makes its use complex for non-expert users, so some purpose-built packages defined on top of INLA have been developed for specific set of models. See Rue et al. [17] for a review of recent examples of applications using the R-INLA package.

Very recently, the `R` package "SpatialEpiApp" has been developed for the analysis of spatial and spatio-temporal disease data [18] which generates a web application using "shiny" [19]. It is designed for visualizing disease data, estimating risks using INLA, and detecting clusters using the scan statistics implemented in the SaTScan™ software [20]. Although the application provides a wide range of interactive data visualization tools, it has two main limitations: (i) the user must have installed on its own computer all the `R` packages and software dependencies, and (ii) only the spatial convolution model proposed by Besag, York and Mollié [21] and the spatio-temporal model with linear trend described in Bernardinelli et al. [22] are implemented, which may be very limited in many real data analyses.

Although there exist software that allows to fit a wide range of spatial and spatio-temporal models for areal data, in most cases certain programming skills are required to use them properly. In addition, using appropriate prior distributions and solving identifiability problems in the models are not always straightforward. This is what has motivated us to develop the interactive web application SSTCDapp, facilitating the use of fairly complex spatial and spatio-temporal disease mapping models using R-INLA for users in many areas, including epidemiologists and public health researchers, as well as providing additional tools that are useful for a detailed analysis of the model results. SSTCDapp is designed to perform descriptive analyses in space and time of mortality/incidence risks or rates, and to fit a wide variety of spatial and spatio-temporal hierarchical models commonly used in disease mapping. It has been developed with "shiny", a package to build interactive web applications in the `R` software environment. The use of this package is becoming very popular

nowadays and many different environmental modelling applications are being developed [23, 24, 25].

The main objective of the SSTCDapp application is to make available to any potential user a simple and intuitive web tool that allows to fit a wide variety of spatial and spatio-temporal hierarchical models commonly used in disease mapping without installing any software in its computer, since all the analyses and computations are made in a powerful remote server. In addition, a desktop version is also available to run the application locally in those cases in which data confidentiality could be a serious issue. The application may also be used for the analysis of similar problems in other fields such as ecology, criminology, gender-based violence, road-traffic accidents or veterinary.

The rest of this paper is laid out as follows. Section 2 outlines the spatial and spatio-temporal models that are currently available in SSTCDapp, as well as other implemented features. In Section 3 the main functionalities of the application are described and an example illustrating how to analyze breast cancer mortality data in Spanish provinces is given. Finally, a concluding discussion and a summary of future improvements of the application are given in Section 4.

2. Methodology

This section outlines the spatial and spatio-temporal Bayesian hierarchical models available in SSTCDapp. Details about model identifiability, model selection criteria, and hyperprior distributions are also given. In what follows we suppose that we have a region divided into n small areas labelled as $i = 1, \ldots, n$; for a given area i, O_i will denote the observed number of cases and N_i the population at risk.

2.1. Classical risk estimation measures

The simplest mortality/incidence indicator is the *crude rate* (CR), which is defined as the number of cases per 100,000 inhabitants. That is,

$$CR_i = \frac{O_i}{N_i} \times 100,000 \quad \text{for} \quad i = 1, \ldots, n.$$

These rates are often age standardized. The underlying reason is that two populations that have the same age-specific mortality/incidence rates for a specific disease will have different crude rates if the age distributions of the two populations are different. To standardize rates, both direct and indirect age standardization methods can be performed [26].

2.1.1. Standardized rates

The direct standardization method provides age-standardized mortality/incidence rates that would have been observed in a population with the same age structure of a certain reference population, called the standard population. So, the *standardized rate* (SR) for each area is computed as

$$SR_i = \frac{\sum_{j=1}^{J} P_j * \frac{O_{ij}}{N_{ij}}}{\sum_{j=1}^{J} P_j} \times 100,000 \quad \text{for} \quad i = 1, \ldots, n$$

where J is the number of age-groups, P_j is the standard population in the j^{th} age-group, while O_{ij} and N_{ij} are the number of counts and population at risk in area i and age-group j, respectively. By default, the age distribution of the 2013 European Standard Population is used to compute these rates in SSTCDapp.

2.1.2. Standardized mortality/incidence ratio

The indirect standardization method uses the same age-specific rates, generally those computed using the information from all the areas together, applied to the age structure of the population at risk in each geographical unit. In SSTCDapp, the expected number of cases for each area is computed as

$$e_i = \sum_{j=1}^{J} N_{ij} \frac{O_j}{N_j} \quad \text{for} \quad i = 1, \ldots, n$$

where $O_j = \sum_{i=1}^{n} O_{ij}$ and $N_j = \sum_{i=1}^{n} N_{ij}$ are the number of cases and the population at risk in the j^{th} age-group, respectively. Finally, the standardized mortality/incidence ratio (SMR or SIR) is defined as the number of observed cases divided by the number of expected cases. These measures are extremely variable when analyzing rare diseases (with few number of cases) or low-populated

areas. If this is the case, statistical models are needed to provide reliable risk or rate estimates.

2.2. Spatial models for disease mapping

Let us consider that the interest lies in estimating the relative risk r_i of mortality/incidence of a disease in area i. Conditional on these risks, the number of counts O_i is assumed to be Poisson distributed with mean $\mu_i = e_i r_i$. That is,

$$\begin{aligned} O_i|r_i &\sim Poisson(\mu_i = e_i r_i) \quad \text{for} \quad i = 1, \ldots, n, \\ \log \mu_i &= \log e_i + \log r_i. \end{aligned} \tag{1}$$

Here, $\log e_i$ is an offset and depending on the specification of $\log r_i$ different models are defined.

Similarly, if the interest lies in estimating the region specific rate of mortality/incidence of a disease in area i, the model of Equation (1) can be reformulated as

$$\begin{aligned} O_i|r_i &\sim Poisson(\mu_i = N_i r_i) \quad \text{for} \quad i = 1, \ldots, n, \\ \log \mu_i &= \log N_i + \log r_i, \end{aligned}$$

where now the population at risk of each area N_i is the offset of the Poisson log-linear model. In what follows, we will refer to r_i as the relative risk of area i. Most of the spatial disease mapping models in the literature are based on conditional autoregressive (CAR) prior distributions. In the simplest model the log-risk is modelled as

$$\log r_i = \eta + \xi_i, \tag{2}$$

where η is an intercept representing an overall level of risk and $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_n)'$ is a spatially structured random effect. In what follows we describe the different CAR prior distributions that are currently implemented in SSTCDapp.

2.2.1. Intrinsic CAR model

The intrinsic conditional autoregressive (iCAR) prior distribution [27] commonly used in disease mapping is defined as

$$\boldsymbol{\xi} \sim N(\mathbf{0}, [\tau_\xi \mathbf{R}_s]^-), \tag{3}$$

where τ_ξ is a precision parameter and $\mathbf{R}_s$ is the $n \times n$ spatial neighborhood matrix with diagonal elements equal to the number of neighbors of each area and non-diagonal elements $(\mathbf{R}_s)_{ij} = -1$ if areas i and j are neighbors and $(\mathbf{R}_s)_{ij} = 0$ otherwise. Here, two areas are considered as neighbors if they share a common border. The symbol $^-$ denotes the Moore-Penrose generalized inverse of a matrix. The intrinsic CAR prior is improper, i.e., its precision matrix is not of full rank. Clearly $\sum_j (\mathbf{R}_s)_{ij} = 0, \forall i$, that is $\mathbf{R}_s \mathbf{1}_n = 0$, where $\mathbf{1}_n$ is a vector of ones of length n. Hence, an identifiability problem arises for the intercept in Equation (2). The problem can be solved by imposing the sum-to-zero constraint $\sum_{i=1}^{n} \xi_i = 0$ or by deleting the intercept (see for example, Eberly and Carlin [28]).

2.2.2. BYM model

The iCAR prior distribution only accounts for spatial correlation structures, and hence, it is not appropriate if the variability is not spatially structured. Besag, York and Mollié [21] proposed a model (hereafter BYM model) which includes two spatial random effects: one assuming an iCAR prior for the spatially structured variability, and another one assuming a Gaussian exchangeable prior to model unstructured heterogeneity. That is,

$$\boldsymbol{\xi} = \mathbf{u} + \mathbf{v}; \quad \text{with} \quad \begin{array}{l} \mathbf{u} \sim N(\mathbf{0}, [\tau_u \mathbf{R}_s]^-), \\ \mathbf{v} \sim N(\mathbf{0}, \tau_v^{-1} \mathbf{I}_n), \end{array} \tag{4}$$

where τ_u and τ_v are the precision parameters of the structured and unstructured spatial effects respectively, and $\mathbf{I}_n$ is an identity matrix of dimension $n \times n$. However, the variance components in the BYM model are not identifiable from the data [29], and hence only the sum $u_i + v_i$ is identifiable. To solve identifiability problems with the intercept term, the sum-to-zero constraint $\sum_{i=1}^{n} (u_i + v_i) = 0$ must be imposed.

2.2.3. Leroux model

Leroux et al. [30] proposed an alternative formulation to model both spatially unstructured and structured variation in a single set of random effects

(hereafter LCAR model), which is given by

$$\boldsymbol{\xi} \sim N(\mathbf{0}, [\tau_\xi(\lambda_\xi \mathbf{R}_s + (1-\lambda_\xi)\mathbf{I}_n)]^{-1}), \tag{5}$$

where τ_ξ is a precision parameter and λ_ξ is a spatial smoothing parameter taking values between 0 and 1. Note that $\lambda_\xi = 0$ corresponds to the unstructured prior $\boldsymbol{\xi} \sim N(\mathbf{0}, \tau_\xi^{-1}\mathbf{I}_n)$, while $\lambda_\xi = 1$ corresponds to the iCAR prior $\boldsymbol{\xi} \sim N(\mathbf{0}, [\tau_\xi \mathbf{R}_s]^-)$. The covariance matrix of the LCAR model is of full rank whenever $0 \leq \lambda_\xi < 1$, but a confounding problem still remains. In practice, retaining the intercept implicitly included in the LCAR model has no advantage and provokes a variance inflation of the fixed-effects intercept [31]. So, a sum-to-zero constraint $\sum_{i=1}^n \xi_i = 0$ has to be considered.

2.2.4. A modified BYM model (BYM2)

Riebler et al. [32] consider a modification of the model proposed by Dean et al. [33] which addresses both the identifiability and scaling issue of the BYM model, hereafter BYM2 model. The spatial random effect is reparameterized as

$$\boldsymbol{\xi} = \frac{1}{\sqrt{\tau_\xi}}(\sqrt{\lambda_\xi}\mathbf{u}_\star + \sqrt{\mathbf{1}-\lambda_\xi}\mathbf{v}), \tag{6}$$

where $\mathbf{u}_\star$ is the scaled intrinsic CAR model with generalized variance equal to one and $\mathbf{v}$ is the unstructured random effect. The variance of the random effect is expressed as a weighted average of the covariance matrices of the structured and unstructured spatial components (unlike the LCAR model which considers a weighted combination of the precision matrices), i.e.,

$$\mathrm{Var}(\boldsymbol{\xi}|\tau_\xi) = \frac{1}{\tau_\xi}(\lambda_\xi \mathbf{R}_\star^- + (1-\lambda_\xi)\mathbf{I}_n),$$

where $\mathbf{R}_\star^-$ indicates the generalised inverse of the scaled spatial precision matrix [34]. As in the previous models, a sum-to-zero constraint $\sum_{i=1}^n \xi_i = 0$ must be imposed to avoid identifiability problems.

2.3. Spatio-temporal models for disease mapping

Suppose now that for each area i, data are available for different time periods labeled as $t = 1, \ldots, T$. The classical risk estimation measures described for

spatial count data are similarly defined in the spatio-temporal context. For example, using the indirect standardization method, the number of expected cases for area i and time t is now defined as

$$e_{it} = \sum_{j=1}^{J} N_{itj} \frac{O_j}{N_j} \quad \text{for} \quad i = 1, \ldots, n; \quad t = 1, \ldots, T,$$

where $O_j = \sum_{i=1}^{n} \sum_{t=1}^{T} O_{itj}$ and $N_j = \sum_{i=1}^{n} \sum_{t=1}^{T} N_{itj}$ are the number of cases and the population at risk in the j^{th} age-group, respectively.

If the interest lies in estimating the relative mortality/incidence risk of a disease in area i and time t, the model of Equation (1) is extended as follows

$$\begin{aligned} O_{it}|r_{it} &\sim Poisson(\mu_{it} = e_{it} r_{it}) \quad \text{for} \quad i = 1, \ldots, n; \quad t = 1, \ldots, T, \\ \log \mu_{it} &= \log e_{it} + \log r_{it}. \end{aligned} \tag{7}$$

The non-parametric models based on CAR priors for spatial random effects, random walk priors for temporal random effects, and different types of spatio-temporal interactions described in Knorr-Held [35] are currently implemented in SSTCDapp. The log-risks are modelled as

$$\log r_{it} = \eta + \xi_i + \phi_t + \gamma_t + \delta_{it}, \tag{8}$$

where η is an intercept representing an overall level of risk, ξ_i is the spatial component, ϕ_t and γ_t represent the unstructured and structured temporal effects respectively, and δ_{it} is the space-time interaction effect. If the interaction term is dropped, an additive model is obtained. The BYM model (4) was originally proposed by Knorr-Held [35] as the prior distribution for the spatial random effect $\boldsymbol{\xi}$, while the LCAR model (5) is considered instead by Ugarte et al. [36]. In addition to these models, the iCAR model (3) and the BYM2 model (6) have been also implemented as spatial prior distributions when fitting spatio-temporal models. An exchangeable prior distribution is given to the temporal random effect $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_T)'$, that is

$$\boldsymbol{\phi} \sim N(\mathbf{0}, \tau_{\phi}^{-1} \mathbf{I}_T),$$

Interaction	$\mathbf{R}_\delta$	Spatial correlation	Temporal correlation
Type I	$\mathbf{I}_n \otimes \mathbf{I}_T$	–	–
Type II	$\mathbf{I}_n \otimes \mathbf{R}_t$	–	✓
Type III	$\mathbf{R}_s \otimes \mathbf{I}_T$	✓	–
Type IV	$\mathbf{R}_s \otimes \mathbf{R}_t$	✓	✓

Table 1: Specification for the different types of space-time interactions.

where τ_ϕ is a precision parameter and $\mathbf{I}_T$ is an identity matrix of dimension $T \times T$. For the temporally structured random effect $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_T)'$, random walks of first (RW1) or second order (RW2) can be assumed, i.e.,

$$\boldsymbol{\gamma} \sim N(\mathbf{0}, [\tau_\gamma \mathbf{R}_t]^-),$$

where τ_γ is a precision parameter and $\mathbf{R}_t$ is the $T \times T$ structure matrix of a RW1/RW2 (see for example, Rue and Held [7], pp. 95 and 110). In practice the temporal effect of the data is usually structured, so the uncorrelated temporal component ϕ_t can be removed. Finally, the following prior distribution is assumed for the space-time interaction random effect $\boldsymbol{\delta} = (\delta_{11}, \ldots, \delta_{1T}, \ldots, \delta_{n1}, \ldots, \delta_{nT})'$

$$\boldsymbol{\delta} \sim N(\mathbf{0}, [\tau_\delta \mathbf{R}_\delta]^-).$$

Here, τ_δ is a precision parameter and $\mathbf{R}_\delta$ is the $nT \times nT$ matrix obtained as the Kronecker product of the corresponding spatial and temporal structure matrices. As shown in Table 1, four types of interactions can be considered.

2.3.1. Identifiability constraints

Similar to the spatial case, identifiability problems arise in model (8) because the overall level can be absorbed by both the spatial and temporal effects. In addition, the interaction term is confounded with the main effects. To ensure model identifiability, sum-to-zero constraints are usually imposed over the random effects of the model (see Table 2). Details on the derivation of these constraints can be found in Goicoa et al. [31].

Interaction	RW1 prior for $\boldsymbol{\gamma}$	RW2 prior for $\boldsymbol{\gamma}$
Type I	$\sum_{i=1}^{n}\xi_i = 0, \quad \sum_{t=1}^{T}\gamma_t = 0,$ $\sum_{i=1}^{n}\sum_{t=1}^{T}\delta_{it} = 0$	$\sum_{i=1}^{n}\xi_i = 0, \quad \sum_{t=1}^{T}\gamma_t = 0,$ $\sum_{i=1}^{n}\sum_{t=1}^{T}\delta_{it} = \sum_{i=1}^{n}\sum_{t=1}^{T}t\delta_{it} = 0$
Type II	$\sum_{i=1}^{n}\xi_i = 0, \quad \sum_{t=1}^{T}\gamma_t = 0,$ $\sum_{t=1}^{T}\delta_{it} = 0$, for $i = 1, \ldots, n$	$\sum_{i=1}^{n}\xi_i = 0, \quad \sum_{t=1}^{T}\gamma_t = \sum_{t=1}^{T}t\gamma_t = 0,$ $\sum_{t=1}^{T}\delta_{it} = 0$, for $i = 1, \ldots, n$
Type III	$\sum_{i=1}^{n}\xi_i = 0, \quad \sum_{t=1}^{T}\gamma_t = 0,$ $\sum_{i=1}^{n}\delta_{it} = 0$, for $t = 1, \ldots, T$	$\sum_{i=1}^{n}\xi_i = 0, \quad \sum_{t=1}^{T}\gamma_t = 0,$ $\sum_{i=1}^{n}\delta_{it} = 0$, for $t = 1, \ldots, T$
Type IV	$\sum_{i=1}^{n}\xi_i = 0, \quad \sum_{t=1}^{T}\gamma_t = 0,$ $\sum_{t=1}^{T}\delta_{it} = 0$, for $i = 1, \ldots, n$ $\sum_{i=1}^{n}\delta_{it} = 0$, for $t = 1, \ldots, T$	$\sum_{i=1}^{n}\xi_i = 0, \quad \sum_{t=1}^{T}\gamma_t = 0,$ $\sum_{t=1}^{T}\delta_{it} = 0$, for $i = 1, \ldots, n$ $\sum_{i=1}^{n}\delta_{it} = 0$, for $t = 1, \ldots, T$

Table 2: Identifiability constraints for the spatio-temporal CAR models in Equation (8).

2.3.2. Posterior patterns and risk variability decomposition

A decomposition of the estimated log-risks in Equation (7) is given by Adin et al. [37] to make the smoothing effects comparable when fitting different spatio-temporal disease mapping models. For this purpose, the following posterior intercept (η^*), spatial (ξ_i^*), temporal (γ_t^*), and spatio-temporal (δ_{it}^*) patterns are defined

$$\begin{aligned} \eta^* &= \tfrac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\log r_{it}, \\ \xi_i^* &= \tfrac{1}{T}\sum_{t=1}^{T}\log r_{it} - \eta^*, \\ \gamma_t^* &= \tfrac{1}{n}\sum_{i=1}^{n}\log r_{it} - \eta^*, \\ \delta_{it}^* &= \log r_{it} - \xi_i^* - \gamma_t^* - \eta^*. \end{aligned} \tag{9}$$

It can be checked that the estimated log-risks are expressed as the sum of these patterns, i.e., $\log r_{it} = \eta^* + \xi_i^* + \gamma_t^* + \delta_{it}^*$. In addition, since these patterns are centered at zero, the total amount of variability of the overall log-risks can be decomposed as the sum of the spatial, temporal, and spatio-temporal variabilities as follows

$$\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\left(\log r_{it} - \overline{\log r_{it}}\right)^2 = \frac{1}{n}\sum_{i=1}^{n}(\xi_i^*)^2 + \frac{1}{T}\sum_{t=1}^{T}(\gamma_t^*)^2 + \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}(\delta_{it}^*)^2.$$

This decomposition allows as to compute the percentage of variability explained by the spatial, temporal, and spatio-temporal terms.

2.4. Model selection criteria

Some criteria based on the deviance are computed with the SSTCDapp application to compare different models in terms of model fitting and complexity. The deviance information criterion (DIC) [38] is the most commonly used measure of model fit based on deviance for Bayesian models, which is computed as the sum of the posterior mean of the deviance (a measure of goodness of fit) and the number of effective parameters (a measure of complexity). A corrected version of the DIC proposed by Plummer [39] has been also considered in SSTCDapp, because it has been shown that DIC values may under-penalize complex models in disease mapping. The recently derived Watanabe-Akaike information criterion (WAIC) [40] which is recommended by Gelman et al. [41] over the DIC criterion, is also computed by the SSTCDapp application. WAIC is a method for estimating pointwise out-of-sample prediction accuracy from a fitted Bayesian model, and unlike DIC, is invariant to parametrization and also works for singular models. The logarithmic score [42], a scoring rule to compare models in terms of their predictive performance, is also computed when fitting a battery of models with different types of spatio-temporal interaction random effects.

2.5. Prior distribution for the hyperparameters

Prior distributions for the precision parameters have to be specified in all the models for latent Gaussian fields described in Section 2.2. By default,

log-Gamma distributions are given to the log-precision parameters in R-INLA. However, these priors may lead to wrong results and have been criticized in the literature [43, 44]. In the SSTCDapp application, improper uniform prior distribution on the positive real line are considered for the standard deviations, i.e., $\sigma = 1/\sqrt{\tau} \sim U(0, \infty)$. In addition, a standard uniform distribution is given to the spatial smoothing parameter λ_ξ when fitting the LCAR or BYM2 model for the spatial random effect. See Ugarte et al. [45] for details about how to implement these prior distributions in R-INLA.

The penalised complexity (PC) priors [44] are also available in SSTCDapp when scaling intrinsic GMRFs [34]. If PC priors are used for the precision of a Gaussian random effect, the parameters U and α must be specified so that $P(\sigma > U) = \alpha$. The default values in R-INLA $(U, \alpha) = (1, 0.01)$ are considered when fitting iCAR and RW1/RW2 prior distributions, as well as for the corresponding space-time interaction effect. If the BYM2 model is selected for the spatial random effect $\boldsymbol{\xi}$, the values $(U, \alpha) = (0.5, 0.5)$ are given to the probability statement $P(\lambda_\xi > U) = \alpha$.

3. Interactive web application

SSTCDapp is an interactive web application developed with "shiny" for the analysis of spatial and spatio-temporal areal count data, and it is addressed at `https://emi-sstcdapp.unavarra.es/`.

3.1. Description and functionalities

Every new user must register for the first time and a password will be sent to the user's e-mail address. The password is required to login in the SSTCDapp application and to create a personal account. In this way, the user will be able to submit a model on a remote server and collect the results when the computations are finished. The application uses SSH for data transfer to the remote server when fitting the INLA model. No user has access to the data and results of any other user, and the data files uploaded by the users are automatically deleted

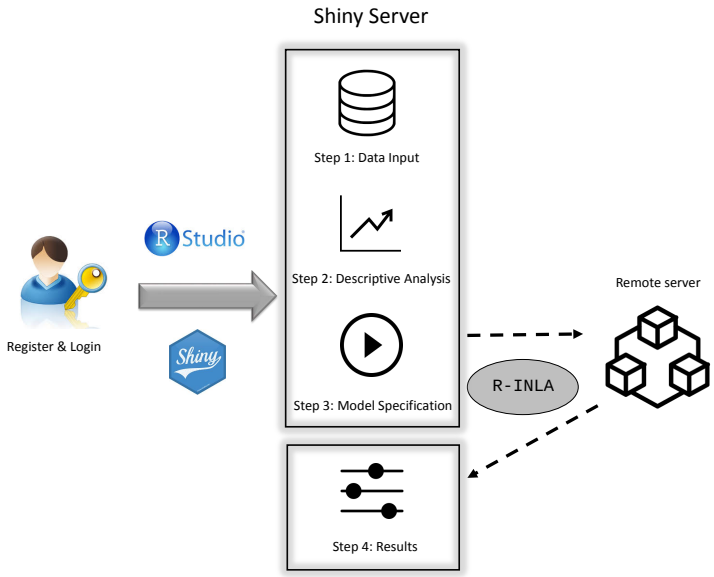


Figure 1: Workflow of the SSTCDapp application.

from the application once they have logged out. A desktop version is also available with the source code of the application to be run locally if needed, fully guaranteeing data confidentiality. As described in Figure 1, the application is structured into four main parts organized in tabs:

1. **Data Input**: The data and the associated cartography are uploaded by the user, and automatically previewed on the screen. Several formats for both data and cartography are supported.

2. **Descriptive Analyses**: The target variables are selected and standardized mortality ratios (SMR) or standardized rates (SR) are computed. Descriptive graphics of the spatial, temporal, and spatio-temporal distribution for the variables of interest (crude rates, SMR or SR) are generated.

3. **Model Specification**: The spatial or spatio-temporal models described above can be fitted. The model is submitted to a remote server and once the calculations are finished, the user will receive a notification by email.

4. **Results**: Summary measures are provided for the posterior distribution

of model hyperparameters, relative mortality/incidence risks (or rates), and spatial, temporal, and spatio-temporal patterns. Maps with the geographical distribution of the disease risks and area-specific temporal evolutions can be generated.

In addition, it includes a *Desktop version* tab to download a zip file with the local version of the application, and a *Help* tab with a detailed user guide and a set of tutorials that show how to fit spatial and spatio-temporal disease mapping models using SSTCDapp. A complete description of the main functionalities can be found in the Appendix.

3.2. Example: Estimating breast cancer mortality risks in Spanish provinces

In this example female breast cancer mortality data in continental Spain during the period 1990-2010 is used to illustrate how to fit spatio-temporal disease mapping models using SSTCDapp. These data were analyzed in Ugarte et al. [46] using different possibilities of modelling the space-time interaction using B-splines in Bayesian disease mapping. All the results shown in this section can be reproduced by downloading the attached files of the tutorial from the *Help* tab of the application. The data files needed to run the example are the following.

- `Breast_Cancer.txt`: A text file with female breast cancer mortality cases registered in the 47 provinces of continental Spain from 1990 to 2010. Data are disaggregated by area, year, and the following age-groups: 1=[0,5), 2=[5,10), ..., 17=[80,85), 18=[+85). The file contains an identification variable for the provinces (`Region`), the year of death (`Year`), the age-group (`Age_group`), the number of observed cases (`Cases`), and the population at risk (`Pop`).

- `Carto_SpainPROV.Rdata`: An `Rdata` file containing an spatial data object of class `SpatialPolygonDataFrame` with the cartography of the Spanish provinces.

- `SpainPROV_nbMatrix.txt` and `SpainPROV_adjacencyMatrix.txt`: Text files with the spatial neighborhood and adjacency matrices.

Firstly, the input data must be uploaded. The user needs to select the data format in the "Data file options" box (in this case `.txt`). Then, the `Breast_Cancer.txt` file is uploaded using the browser of the "Upload data file" box. Secondly, the cartography of the Spanish provinces should be uploaded. The `.Rdata` format in the "Map file options" box should be chosen, uploading the `Carto_SpainPROV.Rdata` file using the browser of the "Upload map file" box. The user then selects `ID.area` as the area variable in the map (see Figure A1).

The target variables must be chosen in the *Variable Selection* tab to compute the standardized mortality ratios (SMRs). Then, the geographical distribution of these ratios can be plotted using the uploaded cartography. The user selects `Region` as the area variable in the "Area" box, `Year` as the time variable in the "Time" box, `Pop` as the population variable in the "Population" box, and `Cases` as the observed cases variable in the "Counts" box. To compute the number of expected cases using the indirect standardization method, `Age_group` has to be chosen as auxiliary variable. Then, the `"Compute aggregated data"` button should be pressed before moving to the next tab (see Figure A2).

Finally, descriptive graphs of the spatial distribution (for the whole period) and the temporal evolution (for the whole Spain) of the SMRs for breast cancer data are generated in the *Graphical Outputs* tab by selecting the `Standardized mortality/incidence ratios` variable (see Figure A3).

The spatial and temporal prior distributions, as well as the corresponding space-time interaction must be selected in the *Model Specification* tab. In addition, the spatial neighborhood structure of the regions must be defined using one of the following options: (i) use the previously uploaded cartography to automatically compute the spatial neighborhood matrix, (ii) upload the spatial neighborhood matrix file `SpainPROV_nbMatrix.txt`, or (iii) upload the spatial adjacency matrix file `SpainPROV_adjacencyMatrix.txt`. Once the spatial neighborhood has been defined, the user presses the `"Show neighborhood graph"` button

to check whether the uploaded neighborhood structure and the spatial regions (Spanish provinces) are consistent. A plot and a summary of the spatial neighborhood structure are shown on the screen, where no disconnected areas are observed (see Figure A4). Finally, the default option `simplified.laplace` approximation strategy has been considered to compute the posterior marginal distribution of the random effects.

Some advanced options are also available in SSTCDapp: (i) by default, posterior distributions of the spatial, temporal, and spatio-temporal patterns are computed; (ii) a battery of models with all types of interactions can be fitted simultaneously, in order to select the model with the most appropriate interaction effect according to some model selection criteria; (iii) the model scaling option is available for intrinsic GMRFs, using PC-priors for the precision parameters (`iCAR` and `RW1/RW2` models) and the spatial smoothing parameter (`BYM2` model); and (iv) the `R` code to fit the models in INLA can be downloaded to be run locally in the user's computer if needed (e.g., to change the hyperprior distributions or other arguments of the `inla()` function).

If the user does not have good computational facilities, he could press the `"Run INLA"` button to submit the model into the remote server. Once the calculations are finished the user will receive a notification by email. The fitted model can be imported into the application using the `"Retrieve selected model(s)"` tab. A summary of the fitted model given by the SSTCDapp is shown in Figure A5. Since posterior patterns have been computed, it is possible to decompose the percentage of variability of the overall risk explained by the estimated spatial (56%), temporal (37%), and spatio-temporal (7%) patterns. If multiple models are selected to be retrieved, a table with different model selection criteria and computational time (in seconds) is printed. In Table A1 the results obtained when fitting all types of space-time interactions are shown, where all the model selection criteria suggest a Type IV (completely structured) space-time interaction as the best model. Consequently, we show results of the selected model next.

Maps of the posterior mean of province-specific relative risks $\zeta_i = \exp(\xi_i^*)$

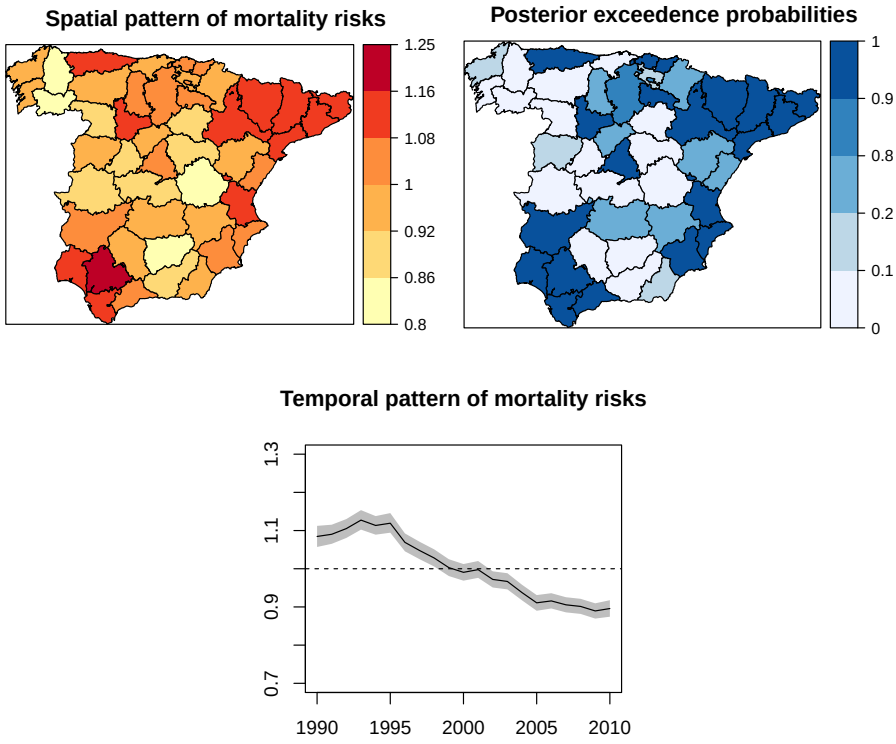


Figure 2: Posterior mean of province-specific relative risks $\zeta_i = \exp(\xi_i^*)$ compared with the whole Spain, and posterior probabilities, $P(\zeta_i > 1|\mathbf{O})$ (top) and global temporal patterns (bottom) of female breast cancer mortality.

compared with the whole Spain, and posterior probabilities, $P(\zeta_i > 1|\mathbf{O})$, are represented at the top of Figure 2. In addition, the global temporal risk pattern $\exp(\gamma_t^*)$ and 95% two-sided credible interval is visualized at the bottom of Figure 2. In Figure 3, maps with the estimated spatio-temporal evolution of breast cancer mortality relative risks and posterior exceedence probabilities $P(r_{it} > 1|\mathbf{O})$ are shown. Finally, plots of area-specific relative risk evolutions for six selected provinces and the corresponding 95% two-sided credible intervals are represented in Figure 4. The colors used in the bands are associated to the posterior exceedence probabilities of relative risks being greater than one. For example the dark blue color means that the risk of the province in those years

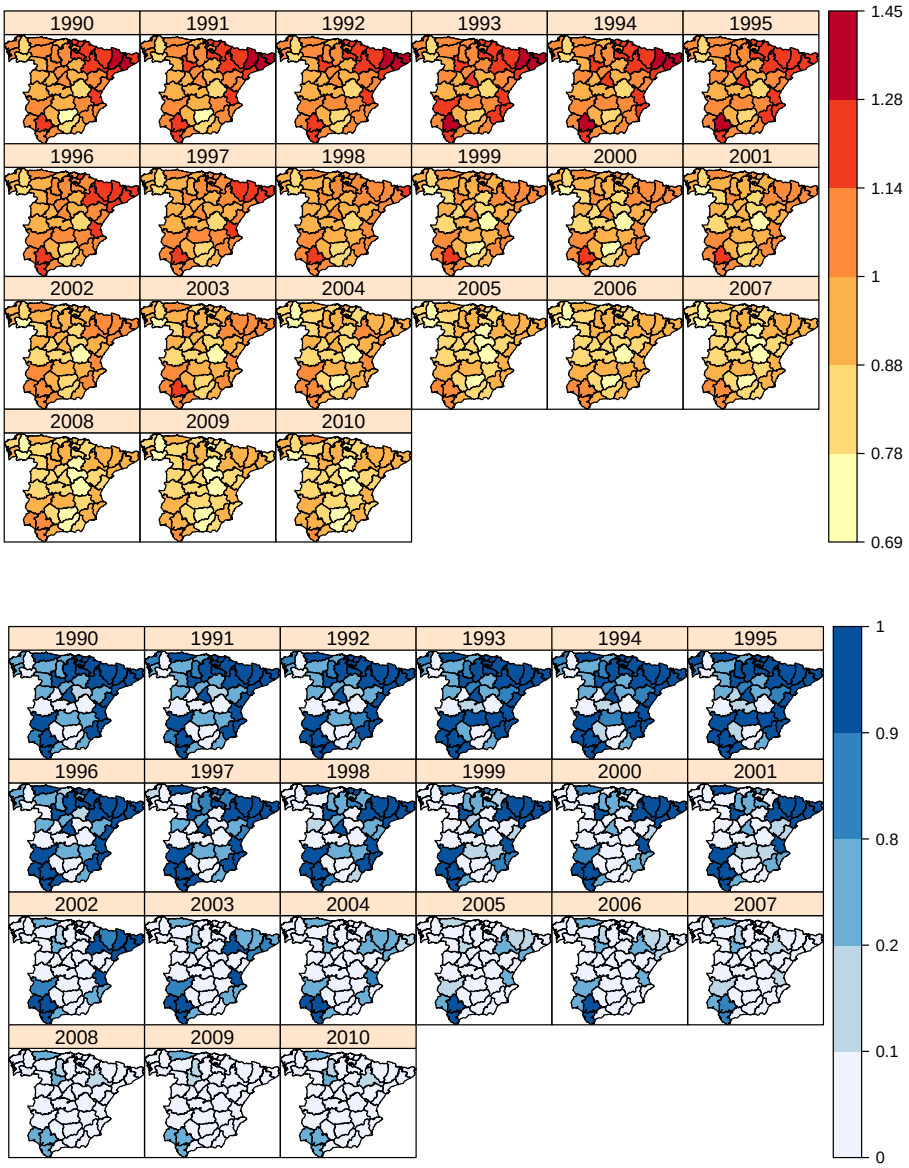


Figure 3: Posterior means of relative risks r_{it} (top) and posterior exceedence probabilities, $P(r_{it} > 1|\mathbf{O})$, (bottom) for female breast cancer mortality in Spanish provinces.

is greater than the risk of the whole of Spain with a high probability.

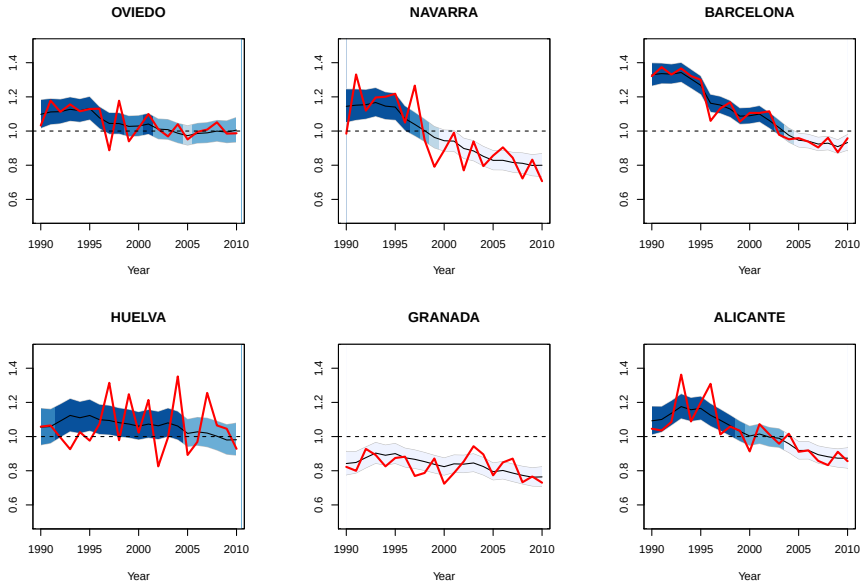


Figure 4: Temporal evolution of female breast cancer mortality relative risks r_{it} for six selected Spanish provinces and 95% two-sided credible intervals. The colors used in the bands are associated to the posterior exceedence probabilities of relative risks being greater than one.

4. Discussion and future work

In this paper we describe the interactive web application SSTCDapp for the analysis of relative risk/rates in spatial and spatio-temporal disease mapping. It provides several graphical tools for the descriptive analysis of the data, as well as the possibility to fit conditional autoregressive (CAR) models with different prior distributions for the spatial, temporal, and space-time interaction random effects. The online version of the application is addressed at `https://emi-sstcdapp.unavarra.es/` and it only requires registration to be used. In addition, a desktop version is also available to run SSTCDapp locally if needed, which avoids uploading the data to the online application fully guaranteeing data confidentiality. It can be downloaded from the left bar menu of the application, and contains the source code and software requirement for its local installation. Unlike the web application, all the computations are made in the

user's computer, including model fitting using the `R-INLA` package, which could require a high RAM/CPU memory usage. In addition, the user can not close the application until the computations are finished. In consequence, the option to fit a battery of models with different types of spatio-temporal interactions is disabled.

A detailed user guide and a set of tutorials are also provided with the application. These tutorials show some examples to estimate spatial and spatio-temporal relative risks/rates with a small number of regions in short computational times. The larger the number of spatial and/or temporal units of the data, the higher the computational cost to fit conditional autoregressive (CAR) models with INLA. This is mainly due to two reasons: (1) the dimension of the precision matrices of the spatial, temporal, and space-time interaction random effects (and therefore the estimation of their posterior distributions), and (2) the number of sum-to-zero constraints imposed to ensure model identifiability, which depends on the number of areas and time points. In these cases, large computational time and high memory usage will be consumed by INLA. Currently, the models are fitted in a remote server with 32 cores and 96GB of RAM, which is powerful enough to fit fairly large data sets in a reasonable time. Note also that INLA models submitted by different users of the application are managed through a queue system in the remote server to avoid memory overflow problems.

Future development of the application will be the integration of `sf` (simple feature) objects [47] as cartography files to generate maps, to include interactive data visualization maps using the `R` package "tmap" [48], and to implement other spatio-temporal proposals such as B-spline models accounting for both spatial and temporal correlation [46], models including age-specific patterns [49], or models to estimate disease risks in the presence of local discontinuities and clusters [50].

5. Conclusions

The SSTCDapp was mainly developed to estimate relative risks or rates using spatial and spatio-temporal disease mapping models. It also provides separate spatial, temporal, and spatio-temporal patterns together with the corresponding exceedence probabilities and/or credibility intervals. Apart from disease mapping problems, it can be used to analyze a wide range of areal count data like gender-based violence, criminology or veterinary data. The key advantage of this application in comparison with other software commonly used in disease mapping is that it provides an user-friendly interface that facilitates the fit of fairly complex models without the need of installing any software in the user's computer. Despite the appeal and easy use of the application, it relies on complex and well founded statistical methodology, and key issues such as model identifiability, model selection, and priors and hyperpriors choices are properly addressed.

Conflict of interest

The authors do not have financial and personal relationships with other people or organizations that could inappropriately influence (bias) their work.

Acknowledgments

We would like to thank Håvard Rue, Andrea Riebler, and Haakon Bakka for their comments and contributions to improve this application. This work has been supported by grants from the Spanish Ministry of Economy and Competitiveness (Project MTM2014-51992-R), the Health Department of the Navarre Government (Project 113, Res.2186/2014) and by Project MTM2017-82553-R (AEI/FEDER, UE).

References

[1] Breslow NE, Clayton DG. Approximate inference in generalized linear mixed models. Journal of the American Statistical Association 1993;88(421):9–25. doi:10.1080/01621459.1993.10594284.

[2] Gilks W, Richardson S, Spiegelhalter D. Markov Chain Monte Carlo in Practice. Chapman & Hall, London; 1996.

[3] Spiegelhalter D, Thomas A, Best N, Lunn D. WinBUGS: version 1.4 User Manual; 2003. URL: https://www.mrc-bsu.cam.ac.uk/software/bugs/the-bugs-project-winbugs/.

[4] Stan Development Team . Stan Modeling Language User's Guide and Reference Manual; 2017. URL: http://mc-stan.org/documentation/; Stan Version 2.17.0.

[5] Carpenter B, Gelman A, Hoffman M, Lee D, Goodrich B, Betancourt M, et al. Stan: A probabilistic programming language. Journal of Statistical Software 2017;76(1):1–32. doi:10.18637/jss.v076.i01.

[6] Rue H, Martino S, Chopin N. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 2009;71(2):319–92. doi:10.1111/j.1467-9868.2008.00700.x.

[7] Rue H, Held L. Gaussian Markov Random Fields: Theory and Applications; vol. 104. Chapman & Hall/CRC; 2005.

[8] Blangiardo M, Cameletti M. Spatial and Spatio-temporal Bayesian Models with R-INLA. John Wiley & Sons; 2015. URL: https://sites.google.com/a/r-inla.org/stbook/.

[9] Brezger A, Kneib T, Lang S. BayesX: Analyzing Bayesian Structural Additive Regression Models. Journal of Statistical Software 2005;14(11):1–22. doi:10.18637/jss.v014.i11.

[10] Umlauf N, Adler D, Kneib T, Lang S, Zeileis A. Structured Additive Regression Models: An R Interface to BayesX. Journal of Statistical Software 2015;63(21):1–46. doi:10.18637/jss.v063.i21.

[11] Meyer S, Held L, Höhle M. Spatio-Temporal Analysis of Epidemic Phenomena Using the R Package surveillance. Journal of Statistical Software 2017;77(11):1–55. doi:10.18637/jss.v077.i11.

[12] Croissant Y, Millo G. Panel Data Econometrics in R: The plm Package. Journal of Statistical Software 2008;27(2):1–43. doi:10.18637/jss.v027.i02.

[13] Millo G, Piras G. splm: Spatial Panel Data Models in R. Journal of Statistical Software 2012;47(1):1–38. doi:10.18637/jss.v047.i01.

[14] Lee D. CARBayes: An R package for Bayesian spatial modeling with conditional autoregressive priors. Journal of Statistical Software 2013;55(13):1–24. doi:10.18637/jss.v055.i13.

[15] Lee D, Rushworth A, Napier G. Spatio-Temporal Areal Unit Modelling in R with Conditional Autoregressive Priors Using the CARBayesST Package. Journal of Statistical Software 2018;84(9):1–39. doi:10.18637/jss.v084.i09.

[16] Bivand RS, Pebesma EJ, Gómez-Rubio V. Applied spatial data analysis with R. 2nd edition. Springer-Verlag, New York; 2013. doi:10.18637/jss.v063.i20.

[17] Rue H, Riebler A, Sørbye SH, Illian JB, Simpson DP, Lindgren FK. Bayesian computing with INLA: a review. Annual Review of Statistics and Its Application 2017;4:395–421. doi:10.1146/annurev-statistics-060116-054045.

[18] Moraga P. SpatialEpiApp: A Shiny web application for the analysis of spatial and spatio-temporal disease data. Spatial and Spatio-Temporal Epidemiology 2017;23:47–57. doi:10.1016/j.sste.2017.08.001.

[19] Chang W, Cheng J, Allaire J, Xie Y, McPherson J. shiny: Web Application Framework for R; 2018. URL: https://CRAN.R-project.org/package=shiny; R package version 1.1.0.

[20] Kulldorff M. SaTScan™ v9.6: Software for the spatial and space-time scan statistics; 2018. URL: http://www.satscan.org/.

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

[22] Bernardinelli L, Clayton D, Pascutto C, Montomoli C, Ghislandi M, Songini M. Bayesian analysis of space-time variation in disease risk. Statistics in Medicine 1995;14(21-22):2433–43. doi:10.1002/sim.4780142112.

[23] Xu W, Collingsworth P, Bailey B, Mazur MC, Schaeffer J, Minsker B. Detecting spatial patterns of rivermouth processes using a geostatistical framework for near-real-time analysis. Environmental Modelling & Software 2017;97:72–85. doi:10.1016/j.envsoft.2017.06.049.

[24] Hossard L, Bregaglio S, Philibert A, Ruget F, Resmond R, Cappelli G, et al. A web application to facilitate crop model comparison in ensemble studies. Environmental Modelling & Software 2017;97:259–70. doi:10.1016/j.envsoft.2017.08.008.

[25] Morley DW, Gulliver J. A land use regression variable generation, modelling and prediction tool for air pollution exposure assessment. Environmental Modelling & Software 2018;105:17–23. doi:10.1016/j.envsoft.2018.03.030.

[26] Ugarte MD. Mortality. In: Kattan MW, editor. Encyclopedia of Medical Decision Making. Thousand Oaks, CA: SAGE Publications; 2009, p. 788–92.

[27] Besag J. Spatial interaction and the statistical analysis of lattice systems. Journal of the Royal Statistical Society Series B (Methodological) 1974;:192–236.

[28] Eberly LE, Carlin BP. Identifiability and convergence issues for Markov chain Monte Carlo fitting of spatial models. Statistics in Medicine 2000;19(17-18):2279–94. doi:10.1002/1097-0258(20000915/30)19:17/18\%3C2279::AID-SIM569\%3E3.0.CO;2-R.

[29] MacNab YC. On Gaussian Markov random fields and Bayesian disease mapping. Statistical Methods in Medical Research 2011;20(1):49–68. doi:10.1177/0962280210371561.

[30] Leroux BG, Lei X, Breslow N. Estimation of disease rates in small areas: A new mixed model for spatial dependence. In: Halloran M, Berry D, editors. Statistical Models in Epidemiology, the Environment, and Clinical Trials. Springer-Verlag: New York; 1999, p. 179–91.

[31] Goicoa T, Adin A, Ugarte MD, Hodges JS. In spatio-temporal disease mapping models, identifiability constraints affect PQL and INLA results. Stochastic Environmental Research and Risk Assessment 2018;32(3):749–70. doi:10.1007/s00477-017-1405-0.

[32] Riebler A, Sørbye SH, Simpson D, Rue H. An intuitive Bayesian spatial model for disease mapping that accounts for scaling. Statistical Methods in Medical Research 2016;25(4):1145–65. doi:10.1177/0962280216660421.

[33] Dean CB, Ugarte MD, Militino AF. Detecting interaction between random region and fixed age effects in disease mapping. Biometrics 2001;57(1):197–202. doi:10.1111/j.0006-341X.2001.00197.x.

[34] Sørbye SH, Rue H. Scaling intrinsic Gaussian Markov random field priors in spatial modelling. Spatial Statistics 2014;8:39–51. doi:10.1016/j.spasta.2013.06.004.

[35] Knorr-Held L. Bayesian modelling of inseparable space-time variation in disease risk. Statistics in Medicine 2000;19(17-18):2555–67. doi:10.1002/1097-0258(20000915/30)19:17/18\%3C2555::AID-SIM587\%3E3.0.CO;2-\%23.

[36] Ugarte MD, Adin A, Goicoa T, Militino AF. On fitting spatio-temporal disease mapping models using approximate Bayesian inference. Statistical Methods in Medical Research 2014;23(6):507–30. doi:10.1177/0962280214527528.

[37] Adin A, Martínez-Beneito M, Botella-Rocamora P, Goicoa T, Ugarte MD. Smoothing and high risk areas detection in space-time disease mapping: a comparison of P-splines, autoregressive, and moving average models. Stochastic Environmental Research and Risk Assessment 2017;31(2):403–15. doi:10.1007/s00477-016-1269-8.

[38] Spiegelhalter DJ, Best NG, Carlin BP, Van Der Linde A. Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 2002;64(4):583–639. doi:10.1111/1467-9868.00353.

[39] Plummer M. Penalized loss functions for Bayesian model comparison. Biostatistics 2008;9(3):523–39. doi:10.1093/biostatistics/kxm049.

[40] Watanabe S. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. Journal of Machine Learning Research 2010;11(Dec):3571–94.

[41] Gelman A, Hwang J, Vehtari A. Understanding predictive information criteria for Bayesian models. Statistics and Computing 2014;24(6):997–1016. doi:10.1007/s11222-013-9416-2.

[42] Gneiting T, Raftery AE. Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association 2007;102(477):359–78. doi:10.1198/016214506000001437.

[43] Carroll R, Lawson A, Faes C, Kirby R, Aregay M, Watjou K. Comparing inla and openbugs for hierarchical poisson modeling in disease mapping. Spatial and Spatio-Temporal Epidemiology 2015;14:45–54. doi:10.1016/j.sste.2015.08.001.

[44] Simpson DP, Rue H, Martins TG, Riebler A, Sørbye SH. Penalising model component complexity: A principled, practical approach to constructing priors. Statistical Science 2017;32(1):1–28. doi:10.1214/16-STS576.

[45] Ugarte MD, Adin A, Goicoa T. Two-level spatially structured models in spatio-temporal disease mapping. Statistical Methods in Medical Research 2016;25(4):1080–100. doi:doi.org/10.1177/0962280216660423.

[46] Ugarte MD, Adin A, Goicoa T. One-dimensional, two-dimensional, and three dimensional B-splines to specify space-time interactions in Bayesian disease mapping: model fitting and model identifiability. Spatial Statistics 2017;22(2):451–68. doi:10.1016/j.spasta.2017.04.002.

[47] Pebesma E. Simple Features for R: Standardized Support for Spatial Vector Data. The R Journal 2018;URL: https://journal.r-project.org/archive/2018/RJ-2018-009/index.html.

[48] Tennekes M. tmap: Thematic Maps in R. Journal of Statistical Software 2018;84(6):1–39. doi:10.18637/jss.v084.i06.

[49] Goicoa T, Ugarte M, Etxeberria J, Militino A. Age–space–time CAR models in Bayesian disease mapping. Statistics in Medicine 2016;35(14):2391–405. doi:10.1002/sim.6873.

[50] Adin A, Lee D, Goicoa T, Ugarte MD. A two-stage approach to estimate spatial and spatio-temporal disease risks in the presence of local discontinuities and clusters. Statistical Methods in Medical Research *(in press)* 2018;:1–19doi:10.1177/0962280218767975.

Appendix

The appendix describes the main functionalities of SSTCDapp.

Data Input

The input data should be provided in a single file. Data files in both `txt` and `csv` plain text formats are supported, with cases corresponding to rows and variables to columns in the file. An `Rdata` file containing a single dataframe can be also uploaded. The file must contain at least the names or IDs of each area and time period (the latter only if spatio-temporal data are analyzed), the observed number of cases, and either the population at risk or the number of expected cases.

The cartography of the region under study associated to the input data needs to be included to generate maps with the spatial or spatio-temporal distribution of the disease. Several cartography file formats are supported, such as `Rdata` or `rds` extensions containing an spatial data object of both `SpatialPolygon` or `SpatialPolygonDataFrame` classes provided by the "sp" package [16]. Commonly used shapefile formats are also supported. In this case, all the related files (at least those with `shp`, `shx` and `dbf` filename extensions) must be uploaded. The areas (polygons) in the cartography file must match those of the input data file.

In Figure A1, a screenshot of the *Data file* and *Map file* input tabs are included when analyzing female breast cancer mortality risks in Spanish provinces.

Descriptive Analyses

First, the target variables are selected and the aggregated data is computed, with rows corresponding to areas in the case of spatial count data, or unique combinations of areas and time points if spatio-temporal data are analyzed. If needed, standardized mortality/incidence ratios or standardized rates are also calculated. Finally, several graphs are generated for the descriptive analyses of the risks/rates.

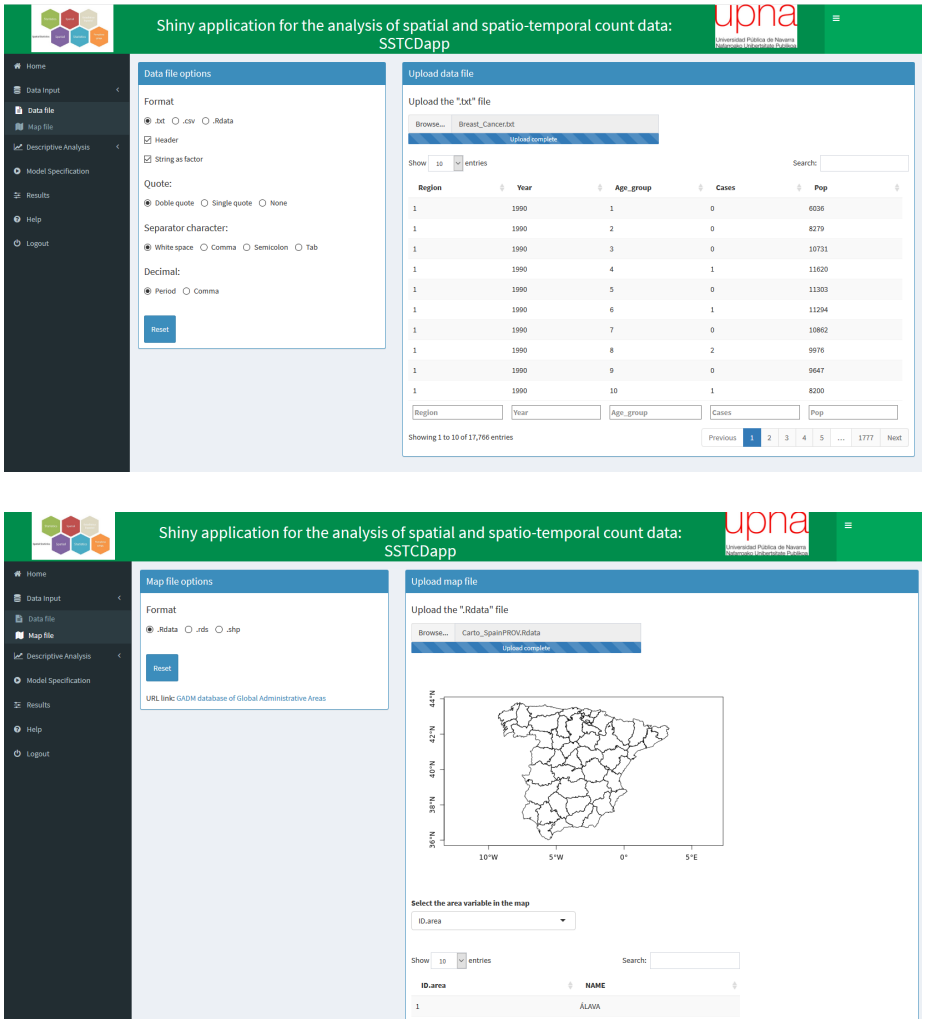


Figure A1: *Data file* (top) and *Map file* (bottom) input tabs in SSTCDapp when analyzing female breast cancer mortality data in Spanish provinces.

The following variables must be selected from the input data file: the variables with the names or IDs of the areas and time points (leave the latter blank for purely spatial analysis), the observed number of cases, and the population at risk. Additionally, the offset variable of the Poisson log-linear model must be specified. If the interest lies in analyzing relative risks, the offset corresponds to

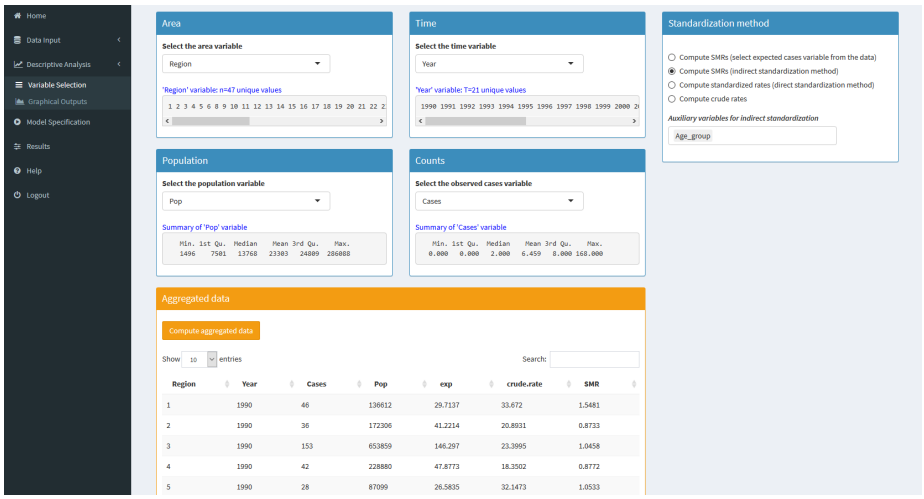
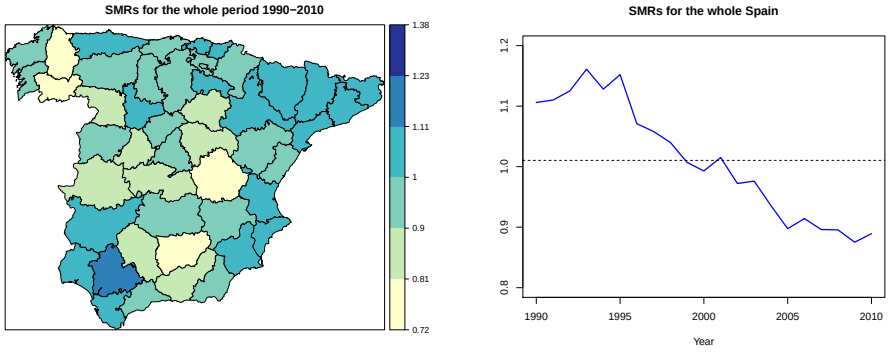


Figure A2: *Variable Selection* tab in SSTCDapp when analyzing female breast cancer mortality data in Spanish provinces.

the number of expected cases. This variable can be selected from the input data file, or can be computed using the indirect standardization method by selecting the auxiliary variables (which usually correspond to age-groups). If the interest lies in analyzing rates, the offset of the model corresponds to the population at risk. The application includes the option to compute standardized rates using the direct standardization method. By default, the age distribution of the 2013 European Standard Population is used to compute these rates. However, an external standard population file can be also uploaded. When all these variables are selected, the `"Compute aggregated data"` button should be pressed before moving to the next tab. In Figure A2, a screenshot of the *Variable Selection* tab is included when analyzing female breast cancer mortality risks in Spanish provinces.

Then, the application allows to generate descriptive graphics of the spatial distribution (a map with the geographical distribution for the whole study period), temporal evolution (a line chart with the temporal evolution for the whole area), and spatio-temporal distribution (maps with the geographical distribution for each time point). The following variables of interest can be selected:



(a) Spatial distribution

(b) Temporal evolution

Figure A3: Spatial distribution (for the whole period) and temporal evolution (for the whole Spain) of breast cancer standardized mortality ratio (SMR). These plots are generated from the *Graphical Outputs* tab in SSTCDapp.

crude rates, standardized rates, or standardized mortality/incidence ratios. In Figure A3, the spatial distribution (for the whole period) and temporal evolution (for the whole Spain) of female breast cancer standardized mortality ratio are represented.

Model Specification

Different spatial and spatio-temporal models commonly used in disease mapping are available.

- Spatial prior distribution: `iCAR`, `BYM`, `Leroux` or `BYM2` prior distributions can be selected for the spatial random effect (see Section 2.2). If a cartography file has been uploaded, the spatial neighborhood matrix can be automatically computed from the map. Alternatively, a file containing the neighborhood structure can be uploaded. An option to check whether the neighborhood structure and the spatial regions are consistent has been included, which automatically shows a warning message if the spatial neighborhood structure is not a connected graph.

- Temporal prior distribution: `RW1` or `RW2` prior distributions can be selected for

the temporally structured random effect (see Section 2.3). An unstructured temporal random effect can be also included in the model.

- Spatio-temporal prior distribution: `None` (additive model), `TypeI`, `TypeII`, `TypeIII` or `TypeIV` prior distributions can be selected for the space-time interaction random effect. See Table 1 for details about the correlation structure of the different interaction types.

Three approximation strategies are available in R-INLA to compute the marginal distributions for the latent effects of the model: `gaussian`, `simplified.laplace` or `laplace`. The Gaussian approximation is the fastest option and often gives reasonable results, but it may be inaccurate. The "full" Laplace approximation is very accurate, but it can be computationally expensive. The simplified Laplace approximation (default option in SSTCDapp) offers a tradeoff between accuracy and computing time. In addition, different numerical integration methods are also available in R-INLA to approximate the posterior distribution of model hyperparameters: the central composite design (`ccd`), the grid exploration (`grid`), the empirical Bayes (`eb`), or the `auto` strategies. The latter is the default option in INLA, and automatically chooses the integration strategy depending on the number of hyperparameters. See Rue et al. [6] for details about the approximation strategies and integration strategies in R-INLA.

The `"Run INLA"` button should be pressed to submit the selected model into a remote server, and the user can monitor the status of the model fitting at any time through the application. Each model has a unique identifier, necessary to manage the model retrieve and/or delete options. A name for the model can be also specified by the user. The users will receive an email once the model calculations are done, so they do not have to wait with the application open until the model has finished. In Figure A4, a screenshot of the *Model Specification* tab is included when analyzing female breast cancer mortality risks in Spanish provinces.

The models fitted on the remote server can be retrieved into the application through the `"Import model(s)"` button. If a single model is selected, a

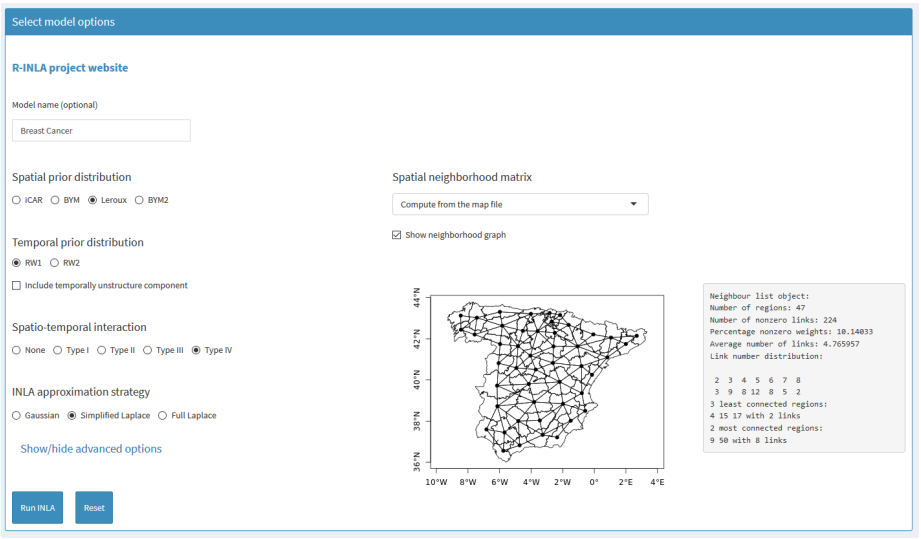


Figure A4: *Model Specification* tab in SSTCDapp when analyzing female breast cancer mortality data in Spanish provinces.

summary is printed on the screen with information related to the INLA version, model name, computation time, posterior distributions of fixed effects and model hyperparameters, and deviance information criteria among others. In Figure A5, the summary output given by SSTCDapp when fitting a model with LCAR prior distribution for space, a RW1 prior distribution for time and TypeIV space-time interaction is shown.

If multiple models are selected, a table is printed with the prior distributions selected for the spatial, temporal, and interaction random effects, the INLA approximation used and the integration strategy, as well as different model comparison measures. The models can be deleted from the remote server manually. Otherwise, they will be automatically removed 7 days after their execution has finished. We recommend to save the imported models as `Rdata` files. In Table A1, the results obtained when fitting all types of space-time interactions are shown.

```
INLA version ............: 17.06.20
INLA date ...............: Tue 20 Jun 12:36:50 JST 2017
INLA hgid ...............: Version_17.06.20
INLA-program hgid .......: Version_17.06.20
Main web-page ...........: www.r-inla.org
Download-page ...........: inla.r-inla-download.org
Email support ...........: help@r-inla.org
                         : r-inla-discussion-group@googlegroups.com


Model name:  Breast_Cancer_TypeIV

CPU used
        Total : 43.701 seconds

Fixed effects:
              mean     sd 0.025quant 0.5quant 0.975quant    mode kld
(Intercept) -0.0384 0.0042    -0.0467  -0.0384    -0.0303 -0.0384   0

Random effects:
Name      Model
 ID.area   Generic1 model
ID.year   RW1 model
ID.area.year   Generic0 model

Model hyperparameters:
                              mean       sd 0.025quant  0.5quant 0.975quant      mode
Precision for ID.area        45.569  16.8925    20.7722   42.8732    86.2287   37.8597
Beta for ID.area              0.421   0.1981     0.0956    0.4047     0.8208    0.3295
Precision for ID.year      1749.868 776.7236   677.6205 1605.5543  3667.8590 1344.8481
Precision for ID.area.year 1025.171 289.3030   582.8889  982.2402  1711.4295  901.9874

Expected number of effective parameters(std dev): 122.61(10.98)
Number of equivalent replicates : 8.05

Deviance Information Criterion (DIC) ...: 7206.08
Effective number of parameters .........: 123.78

Watanabe-Akaike information criterion (WAIC) ...: 7207.85
Effective number of parameters .................: 111.38

Marginal log-Likelihood:  -4366.96
CPO and PIT are computed

Posterior marginals for linear predictor and fitted values computed


Decomposition of the total amount of variability of log-risks:
Spatial = 56.33% | Temporal = 37.07% | Space-time = 6.61%
```

Figure A5: Summary output given by SSTCDapp when fitting a model with LCAR prior distribution for space, a RW1 prior distribution for time and TypeIV space-time interaction (when analyzing female breast cancer mortality data in Spanish provinces).

Results

Finally, summary measures and several graphs/tables are provided for the posterior distributions of model hyperparameters, relative mortality/incidence

Interaction	deviance	p.eff	DIC	DICc	WAIC	LS	Time
Additive	7290.68	58.10	7348.78	7352.95	7368.69	3684.87	2 sec.
Type I	7298.82	195.93	7294.75	7388.09	7296.37	3665.61	6 sec.
Type II	7275.84	144.66	7220.50	7255.24	7222.68	3616.50	44 sec.
Type III	7295.16	167.31	7262.47	7337.15	7262.54	3644.39	15 sec.
Type IV	7282.30	123.78	7206.08	7233.22	7207.85	3607.88	62 sec.

Table A1: Model comparison results given by SSTCDapp when fitting a model with LCAR prior distribution for space, a RW1 prior distribution for time and all types of space-time interaction using the `simplified.laplace` approximation (when analyzing female breast cancer mortality data in Spanish provinces).

risks (or rates), and spatial, temporal, and spatio-temporal patterns (see Section 2.3). These results are computed for the currently imported model or any previously saved model.

- Model description: The main characteristics of the imported/loaded model are shown on the screen, such as the model ID and name, the prior distributions for the space, time and interaction random effects, and the INLA approximation and integration strategies.

- Hyperparameter distribution: Posterior marginal distributions of model hyperparameters are displayed in both precision or variance scale. Specifically, a table of summary statistics and plots of the posterior marginal distributions are generated.

- Posterior relative risks (or rates): Maps of posterior means and posterior exceedence probabilities of being greater than a specified threshold value are plotted. A table with summary statistics of posterior estimates can be also exported. If spatio-temporal relative risks have been computed, graphs with area-specific temporal evolutions are also generated (posterior mean estimates and 95% two-sided credible intervals).