Scalable Bayesian models for spatio-temporal count data

Aritz Adin
aritz.adin@unavarra.es
(joint work with E. Orozco-Acosta and M.D. Ugarte)
Departament of Statistics, Computer Science and Mathematics and INAMAT2
Public University of Navarre (UPNA)

New Bridges between Mathematics and Data Science

November 8-11, 2021

Spatial Statistics
Public University of Navarre

upna
Universidad Pública de Navarra
Nafarroako Unibertsitate Publikoa

1 Introduction

2 Statistical models in space-time disease mapping
2.1 Classical risk estimation measures
2.2 Spatio-temporal CAR models

3 Scalable Bayesian model proposal
3.1 R package `bigDM`
3.2 Methodology

4 Results
4.1 Data analysis: lung cancer mortality risks
4.2 Simulation study

5 Conclusions and further work

Introduction

- Disease mapping deals with areal count data from non-overlapping units focussing on the estimation of the geographical distribution of a disease and its evolution in time.

- The development of statistical techniques for disease mapping has been tremendous in the last few years, mainly due to the availability of information from modern registers with high quality data recorded throughout many years and regions.

- The information acquired from these analyses is of great interest for health researchers, epidemiologists and policy makers as it helps to
  - formulate hypotheses about the disease's etiology
  - look for main risk factors
  - allocate economic resources efficiently in prevention or intervention programs

Introduction

- Three main inferential goals in disease mapping:
  1. To provide accurate estimates of mortality/incidence risks or rates in space and time
  2. To unveil underlying spatial and spatio-temporal patterns
  3. To detect high-risk areas or hotspots

- The great variability inherent to classical estimation measures, makes it necessary to use statistical models to smooth risks borrowing information from spatial and temporal neighbors.

- Despite the enormous expansion of modern computers and the development of new software and estimation techniques to make fully Bayesian inference, dealing with massive data is still computationally challenging.

Introduction

- Mixed Poisson models including conditional autoregressive (CAR) priors for space and random walk priors for time including space-time interactions (Knorr-Held, 2000) are typical models in space-time disease mapping.
- Other approaches based on reduced rank multidimensional P-splines have been also proposed in this field (see for example Ugarte et al., 2017).
- However, are these smoothing methods feasible when analyzing 'very' large spatio-temporal datasets?
- **Objective**: To propose a scalable Bayesian modeling approach to smooth mortality or incidence risks in a high-dimensional spatio-temporal disease mapping context.

Classical risk estimation measures

- Classical risk estimation measures such as the standardized mortality ratio (SMR), are extremely variable when analyzing rare diseases (with few cases) or low-populated areas.
- Therefore, the use of **statistical models** to smooth risks borrowing information from spatial and temporal neighbors is necessary.

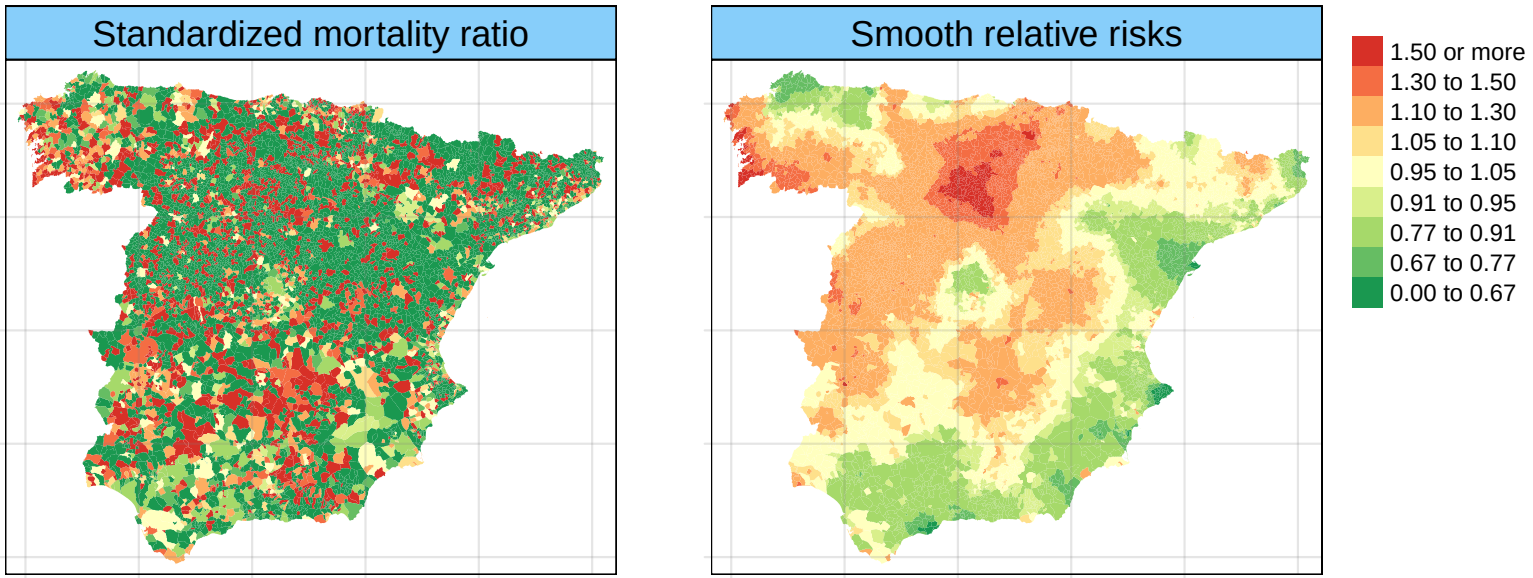


Figure 1: Maps with SMRs and smooth relative risks in the municipalities of Spain.

Statistical models in space-time disease mapping

Let us assume that the region under study is divided into contiguous small areas labeled as $i = 1, \ldots, S$, and data are available for consecutive time periods labeled as $t = 1, \ldots, T$.

- O_{it} denotes the number of observed cases for area i and time t.
- E_{it} denotes the number of expected cases for area i and time t.
- r_{it} denotes the relative risk of mortality (incidence).

Then,

$$\begin{aligned} O_{it}|r_{it} &\sim Poisson(\mu_{it} = E_{it} r_{it}) \\ \log \mu_{it} &= \log E_{it} + \log r_{it} \end{aligned}$$

Depending on the specification of $\log r_{it}$, different models are defined.

Spatio-temporal CAR models

Slight modifications of the **spatio-temporal CAR models** described by Knorr-Held (2000) were considered by Ugarte et al. (2014)

$$\log r_{it} = \alpha + \xi_i + \gamma_t + \delta_{it}$$

- α is a global intercept.
- ξ is a spatially structured random effect with a CAR prior distribution.
- γ_t is a temporally structured random effect that follows a random walk prior distribution.
- δ_{it} is a spatio-temporal random effect (four types of interactions).

- These models are flexible enough to describe real situations, and their interpretation is simple and attractive.
- However, the models are typically not identifiable and appropriate sum-to-zero constraints must be imposed over the random effects (Goicoa et al., 2018).
- We will refer to this model as the **Global model**.

Scalable Bayesian model proposal

In this work, we extend the scalable Bayesian spatial model proposed by Orozco-Acosta et al. (2021) based on the idea of "divide-and-conquer" so that local spatio-temporal models can be simultaneously fitted.

- Several scalable spatial models for high-dimensional areal count data are already implemented in the `R` package **bigDM**, available at

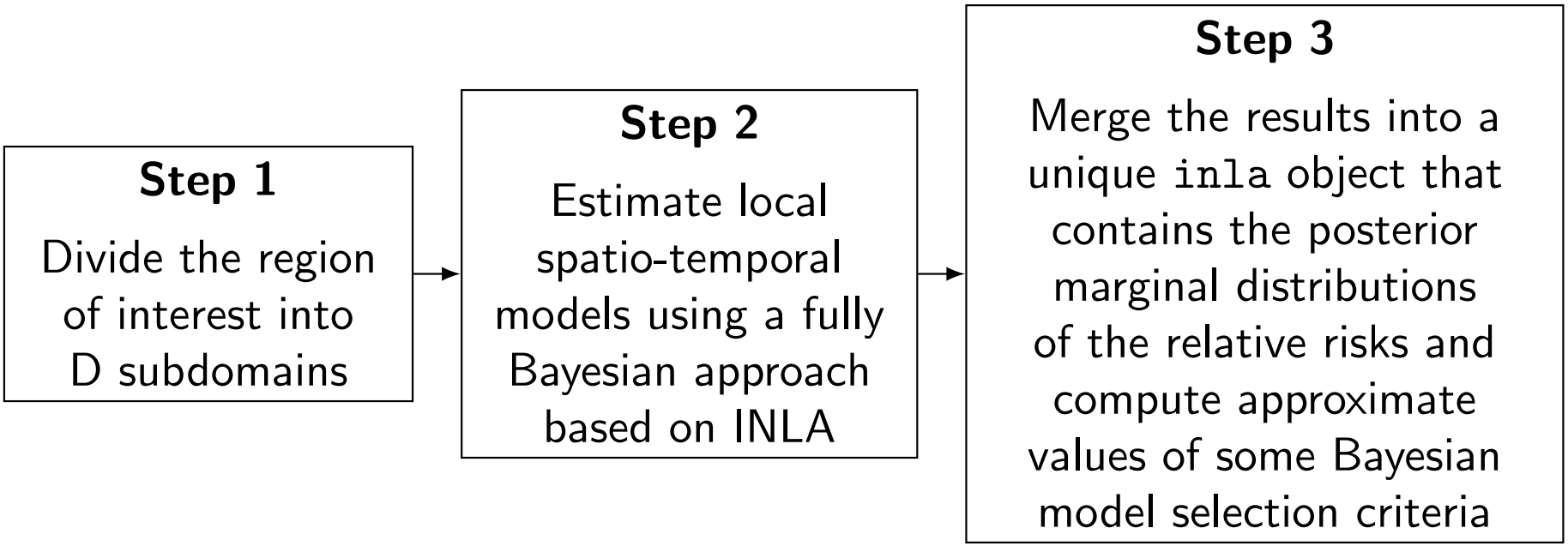
  `https://github.com/spatialstatisticsupna/bigDM`

- Inference is fully Bayesian using the well-known integrated nested Laplace approximation (INLA; Rue et al., 2009) technique through the `R-INLA` package.

- Parallel or distributed computation strategies can be performed to speed up computations by using the `future` package (Bengtsson, 2020).

Scalable Bayesian model proposal

Our modeling approach consists of three main steps:



Step 1: divide the data

- Instead of considering global random effects whose correlation structures are based on the whole spatial/temporal neighbourhood graphs of the areal-time units, we propose to divide the data into $D = D_s \times D_t$ subdomains, where D_s and D_t denote the number of spatial and temporal partitions, respectively.

- Extending the methodology described in Orozco-Acosta et al. (2021), we define **Disjoint** and **k-order neighbourhood models** for estimating spatio-temporal disease risks.

Step 2: local spatio-temporal models

- **Disjoint model**
  - A partition of the spatio-temporal domain $\mathcal{D} = \mathcal{D}^s \times \mathcal{D}^t$ into D sub-domains is defined, so that $\mathcal{D} = \bigcup_{d=1}^{D} \mathcal{D}_d$ where $\mathcal{D}_j \cap \mathcal{D}_k = \emptyset$ for all $j \neq k$.
  - If we denote as A_{it} to the small area i in time period t, note that each area-time unit A_{it} belongs to a single sub-domain.

- **k-order neighbourhood model**
  - Assuming independence between areas belonging to different sub-domains could be very restrictive and it may lead to border effects
  - We avoid this undesirable issue by adding neighbouring area-time units to each partition of the spatial and/or temporal sub-domain $\mathcal{D}^s$ and $\mathcal{D}^t$, respectively.

Toy example: purely spatial partition ($D_t = 1$)

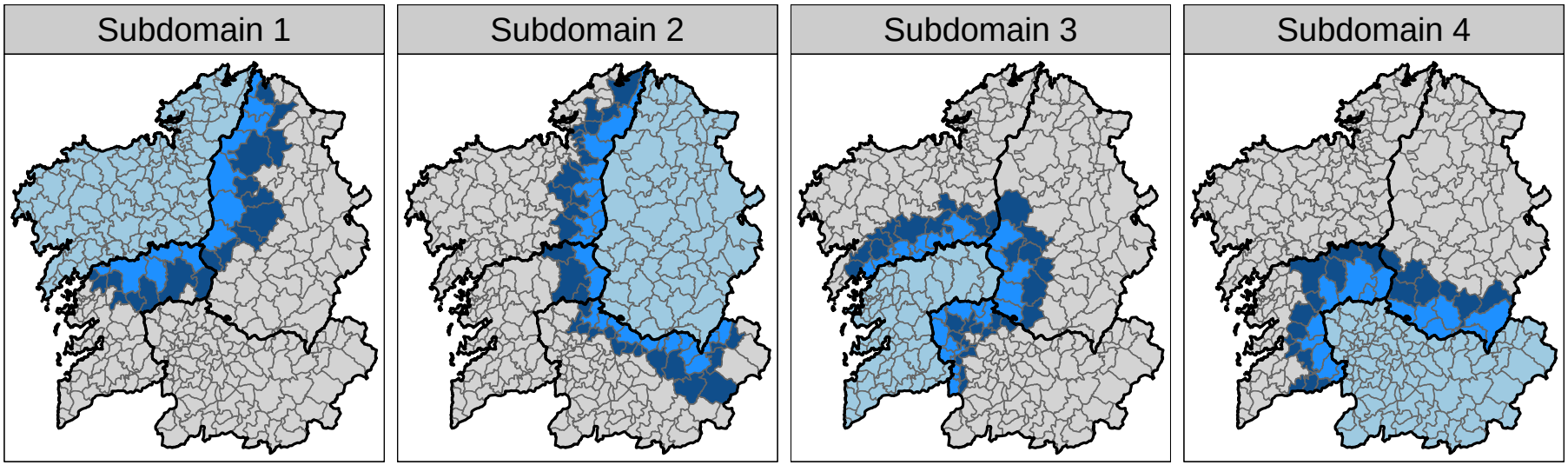


Figure 2: Toy example of a purely spatial partition using the disjoint and 1st/2nd-order neighbourhood models.

Step 3: merge the results

- **Disjoint model**
  - The log-risk surface $\log \boldsymbol{r} = (\log \boldsymbol{r}_1, \ldots, \log \boldsymbol{r}_D)'$ is just the union of the posterior marginal estimates of each spatio-temporal sub-model.
- ***k*-order neighbourhood model**
  - Since multiple relative risk estimates are obtained for some A_{it} units, we compute mixture distributions of the posterior probability density functions estimated from the different local spatio-temporal models to obtain a single posterior distribution for each r_{it}.
  - We use the *conditional predictive ordinate* (CPO), a diagnostic measure to detect discrepant observations from a given model (Pettit, 1990), to compute the weights of the mixture distribution.

$$CPO_{it} = Pr(O_{it} = o_{it} | \mathbf{o}_{-it})$$

- Approximations to model selection criteria such as DIC (Spiegelhalter et al., 2002) and WAIC (Watanabe, 2010) are also derived.

Toy example: mixture distribution

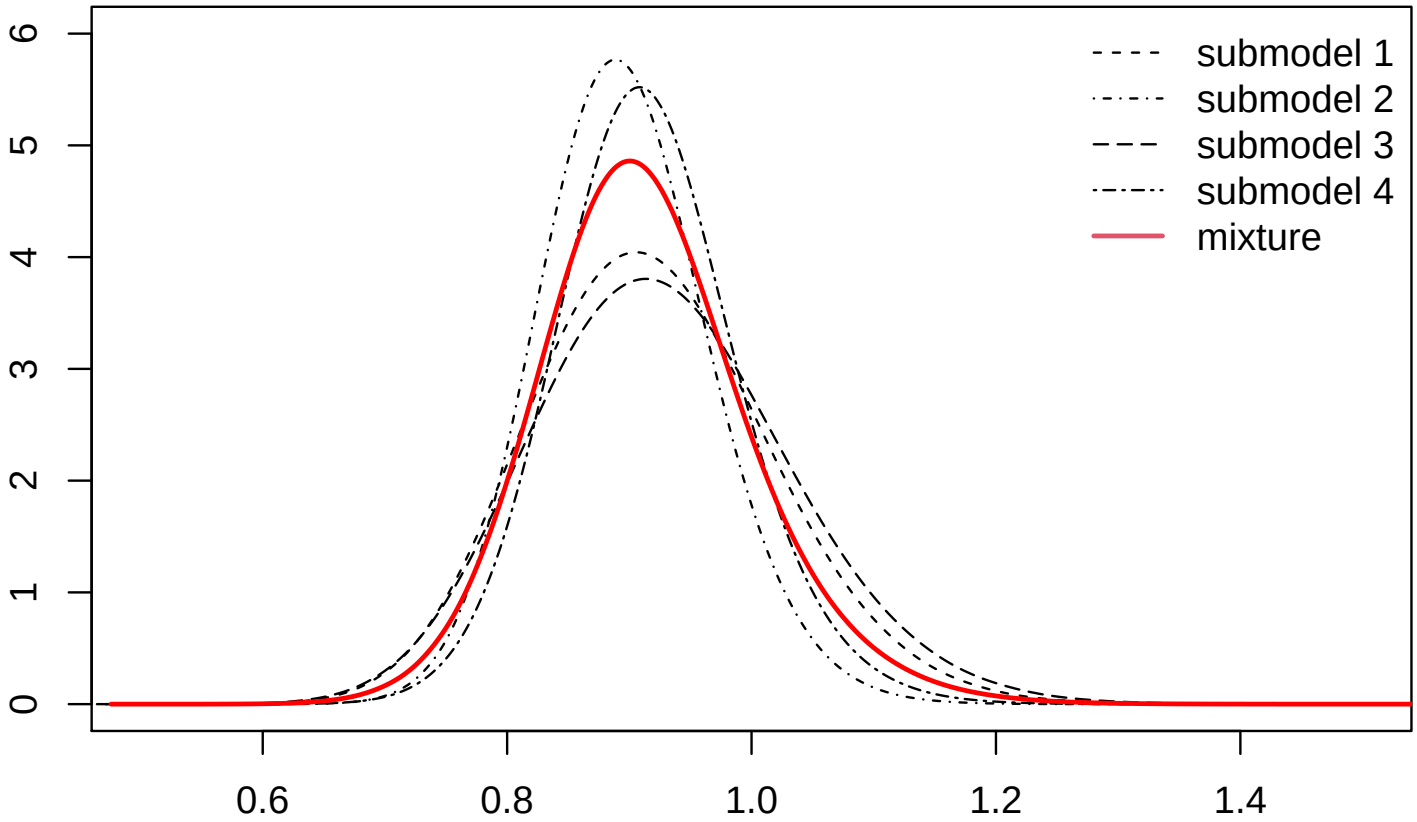


Figure 3: Toy example of a mixture distribution of posterior marginal estimates of relative risks.

Data analysis: lung cancer mortality risks

We illustrate the models's behaviour by estimating lung cancer mortality risks in the $S = 7907$ municipalities of continental Spain during the period 1991-2015 (T=25).

- **Main problem**: Computationally unfeasible to fit Type II and Type IV interaction *Global* models
  - Huge dimension of the spatio-temporal structure matrix

$$197\,675 \times 197\,675\,(\approx 4x10^{10}\text{elements})$$

  - High number of identifiability constraints over the spatio-temporal interaction ($\approx 8\,000$ constraints)

- In contrast, we are able to fit our scalable model proposals reducing the RAM/CPU memory usage and computational time substantially.

Data analysis: lung cancer mortality risks

Table 1: Model selection criteria and computational time (in minutes) using the `simplified.laplace` approximation strategy (R-INLA stable version 21.02.23).

Model	Interaction	$\bar{D}$	p_D	DIC	WAIC	T.run	T.merge	T.total
Global	Type I	333787	2984	336771	336802	663	–	663
	Type II	–	–	–	–	–	–	–
	Type III	333573	2968	336541	336564	3845	–	3845
	Type IV	–	–	–	–	–	–	–
Disjoint	Type I	332260	3999	336259	336267	10	6	16
	Type II	332281	3801	336082	336151	218	6	224
	Type III	332207	4015	336222	336267	22	6	27
	Type IV	332237	3753	**335990**	**336070**	259	6	264
1st-order nb	Type I	332222	3965	336187	336210	12	20	32
	Type II	332236	3780	336016	336093	535	20	555
	Type III	332399	3775	336174	336233	32	20	52
	Type IV	332323	3614	**335937**	**336020**	625	20	644

- Spatio-temporal models with BYM2 conditional autoregressive prior for space, first order random walk prior for time and the four types of space-time interactions.
- For the scalable model proposals, we divide the data into $D = 47$ sub-domains using the provinces of Spain to define a purely spatial partition ($D_t = 1$)

Data analysis: lung cancer mortality risks

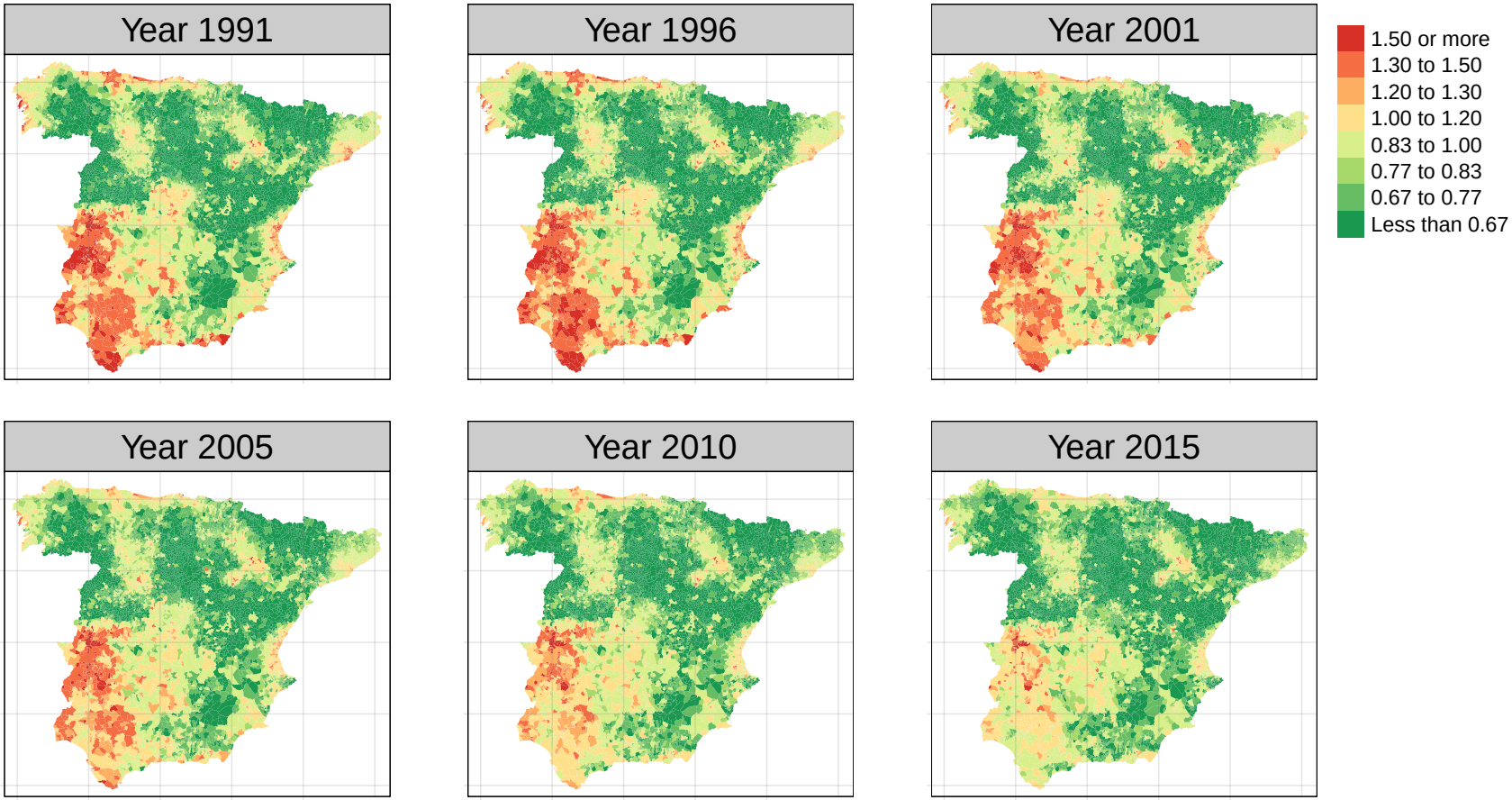


Figure 4: Posterior median estimates of relative risks r_{it} for the 1st-order neighbourhood model considering a Type IV interaction.

Data analysis: lung cancer mortality risks

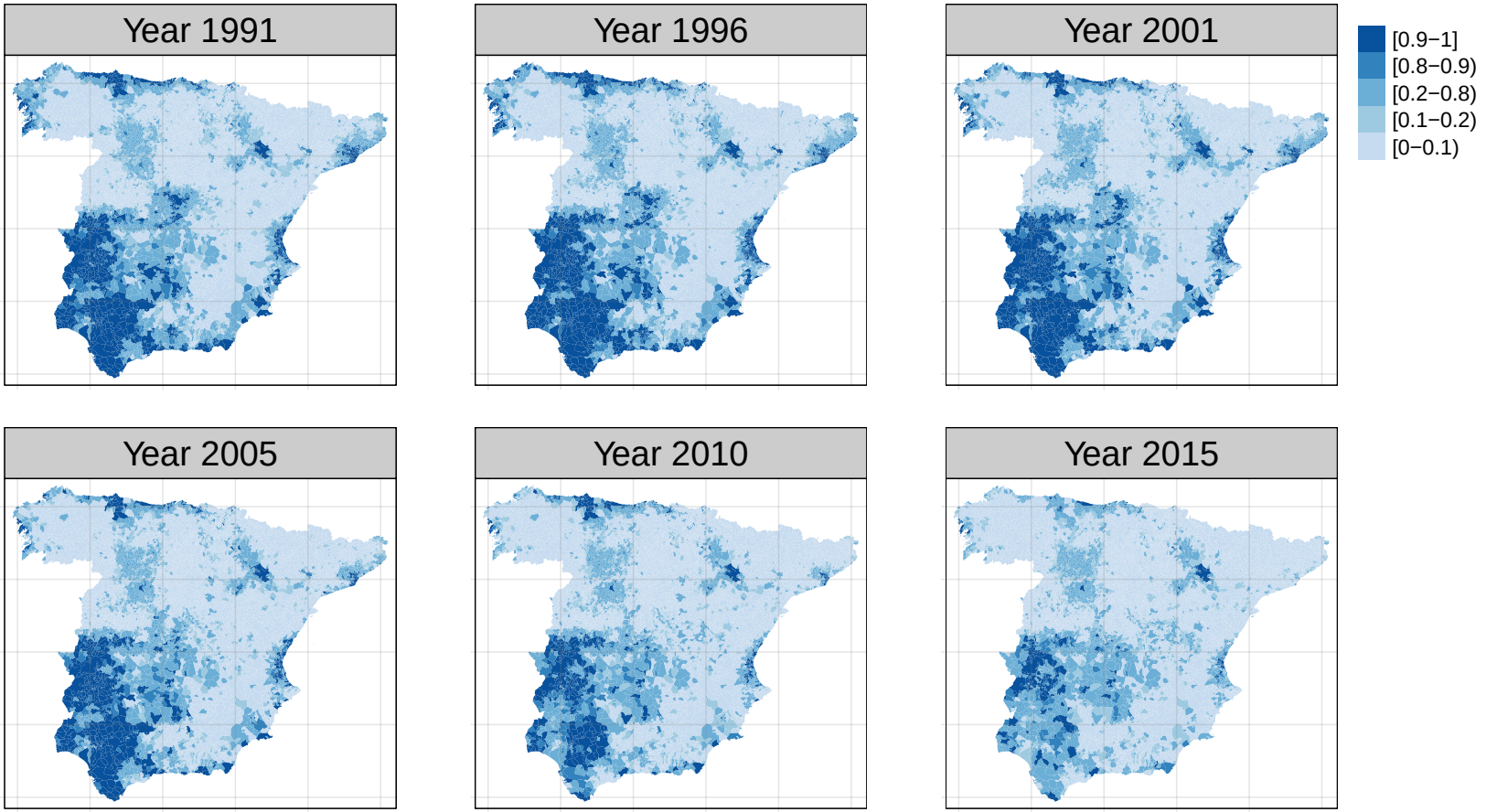


Figure 5: Posterior exceedence probabilities $P(r_{it} > 1|\mathbf{O})$ for 1st-order neighbourhood and Type IV interaction model.

Simulation study

- A simulation study has been conducted to compare the performance of our model's proposals over the almost 8 000 municipalities of continental Spain and $T = 25$ time periods.
- A smooth risk surface is generated by sampling from a three-dimensional P-spline with 20 equally spaced knots for longitude and latitude, and 6 equally spaced knots for time.
- Then, simulate counts for each municipality and time point using a Poisson distribution with mean $\mu_{it} = E_{it} r_{ir}$, where the number of expected cases E_{it} are fixed at value 10.
- A total of 50 simulations have been generated.

Simulation study

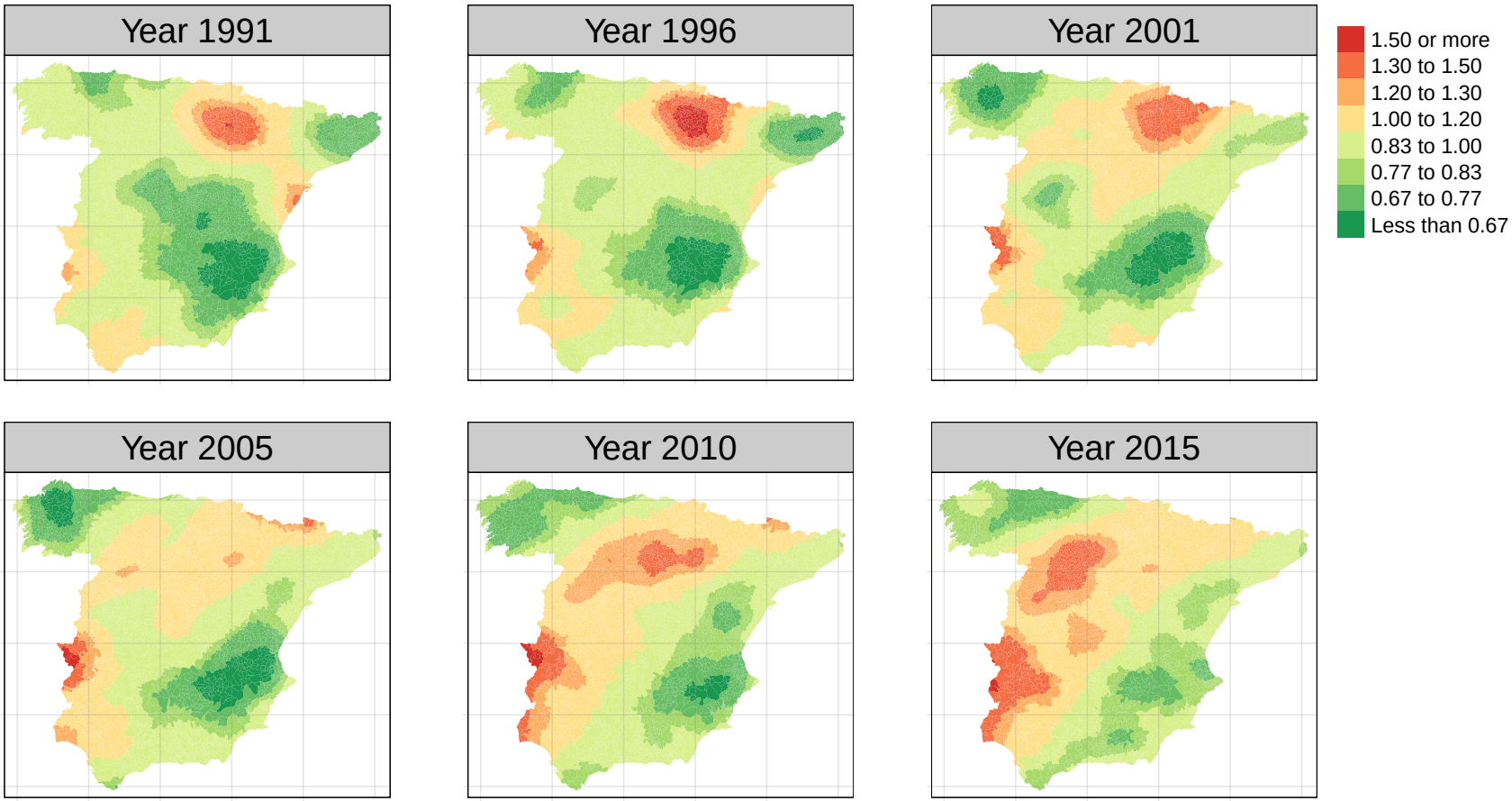


Figure 6: True risk surfaces for the simulation study.

Simulation study

Table 2: Average values of model selection criteria, mean absolute relative bias (MARB) and mean relative root mean square error (MRRMSE).

Model	Interaction	DIC	WAIC	MARB	MRRMSE
Global	Type I	217393	217832	0.0684	0.0782
	Type II	–	–	–	–
	Type III	204930	204666	0.0165	0.0387
	Type IV	–	–	–	–
Disjoint	Type I	206536	206516	0.0322	0.0434
	Type II	205934	205965	0.0281	0.0419
	Type III	204929	204829	0.0203	0.0377
	Type IV	204162	204151	0.0153	0.0331
1st-order nb	Type I	206028	205972	0.0303	0.0416
	Type II	205556	205560	0.0261	0.0404
	Type III	204451	204314	0.0173	0.0352
	Type IV	**203856**	**203833**	**0.0133**	**0.0311**
2nd-order nb	Type I	206165	206093	0.0311	0.0423
	Type II	205717	205706	0.0266	0.0413
	Type III	204534	204370	0.0173	0.0356
	Type IV	**203883**	**203856**	**0.0134**	**0.0312**

Simulation study

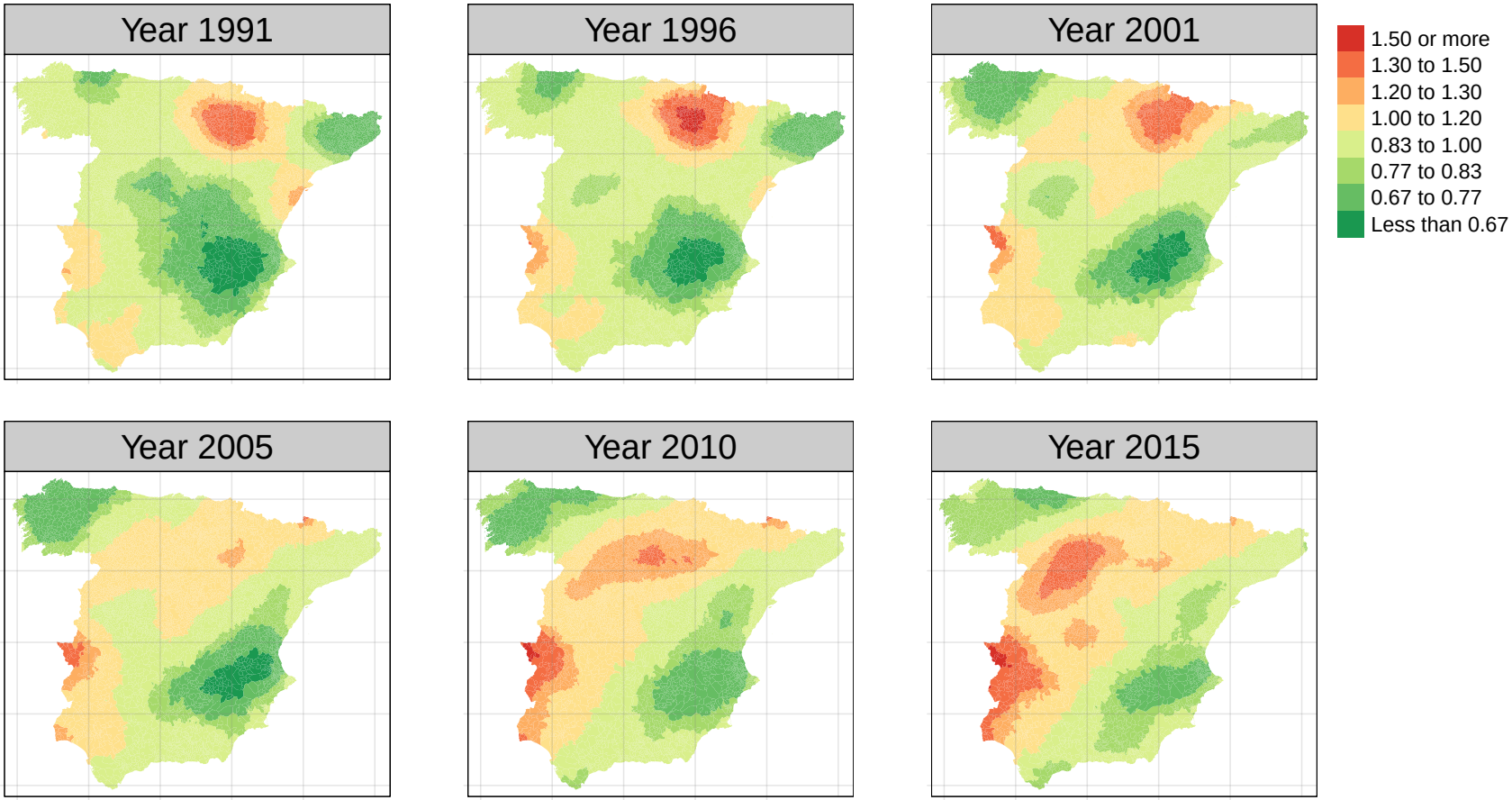


Figure 7: Average values of posterior median estimates of relative risks for 1st-order neighbourhood and Type IV interaction model.

Conclusions and further work

- The "divide-and-conquer" strategy has been extensively used to analyse big data in other contexts such as machine learning, usually using a Bayesian approach to compute tractable posterior distributions (posterior samples if MCMC methods are considered).

- Adapting this idea to the context of disease mapping seems to be very appropriate in practice, since CAR models induce spatial and/or temporal local smoothness by means of neighbouring areas and time points.

- Our scalable methodology proposal provides reliable risk estimates with a substantial reduction in computational time when fitting Bayesian hierarchical spatio-temporal models for high-dimensional data.

- In our data analysis, purely spatial partitions have been considered ($D_t = 1$) but spatio-temporal partitions could be also considered when analyzing large-scale temporal data.

Conclusions and future work

- Although the methodology described here uses the INLA estimation strategy, it could also be adapted to other Bayesian fitting techniques.

- The methods and algorithms proposed in this work are being implemented in the `R` package `bigDM` available at

  `https://github.com/spatialstatisticsupna/bigDM`

- Currently we are working on the development of scalable ecological regression models taking into account the spatial and/or spatio-temporal confounding issues between fixed and random effects (Adin et al., 2021).

- We would like to further investigate other model approaches and computational strategies to deal with large spatio-temporal datasets in disease mapping.

References

Adin, A., Goicoa, T., Hodges, J. S., Schnell, P., and Ugarte, M. D. (2021). Alleviating confounding in spatio-temporal areal models with an application on crimes against women in India. *Statistical Modelling (online first)*.

Bengtsson, H. (2020). A Unifying Framework for Parallel and Distributed Processing in R using Futures.

Goicoa, T., Adin, A., Ugarte, M. D., and Hodges, J. S. (2018). In spatio-temporal disease mapping models, identifiability constraints affect PQL and INLA results. *Stochastic Environmental Research and Risk Assessment*, 32(3):749–770.

Knorr-Held, L. (2000). Bayesian modelling of inseparable space-time variation in disease risk. *Statistics in Medicine*, 19(17-18):2555–2567.

Orozco-Acosta, E., Adin, A., and Ugarte, M. D. (2021). Scalable Bayesian modeling for smoothing disease mapping risks in large spatial data sets using INLA. *Spatial Statistics*, 41:100496.

Pettit, L. (1990). The conditional predictive ordinate for the normal distribution. *Journal of the Royal Statistical Society: Series B (Methodological)*, 52(1):175–184.

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.

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

Ugarte, M. D., Adin, A., and Goicoa, T. (2017). 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*, 22(2):451–468.

Ugarte, M. D., Adin, A., Goicoa, T., and Militino, A. F. (2014). On fitting spatio-temporal disease mapping models using approximate Bayesian inference. *Statistical Methods in Medical Research*, 23(6):507–530.

Acknowledgements

- This research has been supported by Project MTM2017-82553-R (AEI/FEDER, UE) and Project PID2020-113125RB-I00/MCIN/AEI/10.13039/501100011033. It has also been partially funded by the Public University of Navarra (project PJUPNA20001).

- We would like to thank the Spanish National Epidemiology Center (area of Environmental Epidemiology and Cancer) for providing the Spanish data.

Spatio-temporal CAR models

The following prior distribution is assumed for the spatio-temporal interaction random effect

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

where $\mathbf{R}_\delta$ is the space-time structure matrix of dimension $ST \times ST$ obtained as the Kronecker product of the corresponding spatial and temporal structure matrices.

Table 3: Specification for the four possible types of space-time interaction.

Interaction	Structure matrix	Spatial correlation	Temporal correlation
Type I	$\mathbf{R}_\delta = \mathbf{I}_T \otimes \mathbf{I}_S$	–	–
Type II	$\mathbf{R}_\delta = \mathbf{R}_\gamma \otimes \mathbf{I}_S$	–	✓
Type III	$\mathbf{R}_\delta = \mathbf{I}_T \otimes \mathbf{R}_\xi$	✓	–
Type IV	$\mathbf{R}_\delta = \mathbf{R}_\gamma \otimes \mathbf{R}_\xi$	✓	✓

Identifiability constraints

Identifiability constraints for the different types of space-time interaction effects in CAR models Goicoa et al. (2018).

Type I ($\mathbf{R}_\delta = \mathbf{I}_T \otimes \mathbf{I}_S$) :	$\sum_{i=1}^{S} \xi_i = 0,\ \sum_{t=1}^{T} \gamma_t = 0,\ \text{and}\ \sum_{i=1}^{S}\sum_{t=1}^{T} \delta_{it} = 0.$
Type II ($\mathbf{R}_\delta = \mathbf{R}_\gamma \otimes \mathbf{I}_S$) :	$\sum_{i=1}^{S} \xi_i = 0,\ \sum_{t=1}^{T} \gamma_t = 0,\ \text{and}\ \sum_{t=1}^{T} \delta_{it} = 0,\ \text{for}\ i = 1, \ldots, S.$
Type III ($\mathbf{R}_\delta = \mathbf{I}_T \otimes \mathbf{R}_\xi$) :	$\sum_{i=1}^{S} \xi_i = 0,\ \sum_{t=1}^{T} \gamma_t = 0,\ \text{and}\ \sum_{i=1}^{S} \delta_{it} = 0,\ \text{for}\ t = 1, \ldots, T.$
Type IV ($\mathbf{R}_\delta = \mathbf{R}_\gamma \otimes \mathbf{R}_\xi$) :	$\sum_{i=1}^{S} \xi_i = 0,\ \sum_{t=1}^{T} \gamma_t = 0,\ \text{and}\ \sum_{t=1}^{T} \delta_{it} = 0,\ \text{for}\ i = 1, \ldots, S,$ $\sum_{i=1}^{S} \delta_{it} = 0,\ \text{for}\ t = 1, \ldots, T.$

Simulation study

We evaluate models' performance in terms of relative risk estimates by computing the mean absolute relative bias (MARB) and mean relative root mean square error (MRRMSE), defined as

$$\text{MARB} = \frac{1}{ST}\sum_{i=1}^{S}\sum_{t=1}^{T}\frac{1}{100}\left|\sum_{l=1}^{100}\frac{\hat{r}_{it}^{l}-r_{it}}{r_{it}}\right|,$$

$$\text{MRRMSE} = \frac{1}{ST}\sum_{i=1}^{S}\sum_{t=1}^{T}\sqrt{\frac{1}{100}\sum_{l=1}^{100}\left(\frac{\hat{r}_{it}^{l}-r_{it}}{r_{it}}\right)^{2}},$$

where r_{it} is the true generated risk, and $\hat{r}_{it}^{l}$ is the posterior median estimate of the relative risk for arean unit i and time period t in the l-th simulation.