# Spatio-temporal models including time-varying shared space-time interactions to analyze rare cancer sites

*Garazi Retegui*[*,1,2], *Jaione Etxeberria*[1,2], *María Dolores Ugarte*[1,2]

[*]garazi.retegui@unavarra.es

[1] Department of Statistics, Computer Science and Mathematics, Public University of Navarre (UPNA)

[2] Institute for Advanced Materials and Mathematics (INAMAT[2]), Public University of Navarre (UPNA)

**Abstract**

Rare cancers are usually excluded from general analysis as data scarcity leads to imprecise estimates when using standard methods. In this work we propose the use of multivariate spatio-temporal models with different shared interaction terms to analyze jointly incidence and mortality. Results show that for rare cancer sites, multivariate spatio-temporal models with shared interaction perform better than the usual multivariate spatio-temporal models with independent interactions.

**Keywords:** cancer, disease mapping, shared component models.

## 1. Introduction

Disease mapping has a long history in epidemiology and public health. It helps to visualise the geographical distribution of a disease, to monitor changes in disease incidence or prevalence over time, and to evaluate the effectiveness of public health interventions. It can also be used to identify potential risk factors affecting the disease and to help public health authorities plan resource allocation and identify areas for priority action. This is important because the global cost of diseases such as cancer has a major impact on health budgets and society. For a correct allocation of health resources aimed at cancer prevention and control, different indicators such as cancer incidence and mortality rates are usually calculated. However, most of the literature provides incidence and mortality estimates for total cancer or for the most common cancer locations, such as breast or lung cancer. Less common cancers, such as brain or pancreatic cancer, are usually excluded as data scarcity leads to imprecise estimates when standard methods such as univariate spatio-temporal models are used. Therefore, estimating rare or less common cancers becomes a methodological challenge. One approach to overcome this drawback is to use multivariate spatio-temporal models to jointly analyze disease incidence and mortality, sharing information and obtaining better estimates. When using shared multivariate spatio-temporal models researchers usually consider independent spatio-temporal interactions. However, we believe that when the health outcomes of interest have low rates, sharing interactions could improve estimates as the amount of information shared is greater. Therefore, in this work we propose the use of multivariate spatio-temporal models with

shared interaction terms. Our proposal arises by combining ideas from the well-known shared component models [1] and considering the four types of spatio-temporal interactions defined by Knorr-Held [2]. In particular, we define spatio-temporal models that include interactions with a time-varying scaling parameter. To illustrate the new models, we analyze both pancreatic cancer and leukaemia incidence and mortality for males in 142 small areas of Great Britain over nine biennial periods. Model fitting and inference has been carried out with INLA.

## 2. Multivariate spatio-temporal models with time-varying shared interactions

Let $O_{itd}$, $n_{itd}$ and $r_{itd}$ be the observed number of cases, the population at risk and the rates in each area $i$, $i = 1, \ldots, A$ at time $t$, $t = 1, \ldots, T$ for $d = I$, incidence, and $d = M$, mortality. Then, conditional on the rates $r_{itd}$, the number of observed incidence or mortality cases in each area and time, $O_{itd}$, are assumed to follow a Poisson distribution with mean $\mu_{itd} = n_{itd} r_{itd}$, i.e.

$$\begin{aligned} O_{itI}|r_{itI} &\sim Poisson(\mu_{itI} = n_{itI} r_{itI}), \qquad \log \mu_{itI} = \log n_{itI} + \log r_{itI}, \\ O_{itM}|r_{itM} &\sim Poisson(\mu_{itM} = n_{itM} r_{itM}), \qquad \log \mu_{itM} = \log n_{itM} + \log r_{itM}. \end{aligned}$$

To model the log rates, $\log r_{itd}$, we propose two models with time-varying shared interactions. We define Model 1 with a fixed effect for each health-outcome, a shared spatial component, a time effect, and a shared component model for the interactions. In Model 2 we add a spatially unstructured random effect for mortality. Namely

$$\text{Model 1:} \quad \begin{aligned} \log r_{itI} &= \alpha_I + \delta\kappa_i + \gamma_{tI} + \varrho_t \chi_{it}, \\ \log r_{itM} &= \alpha_M + \frac{1}{\delta}\kappa_i + \gamma_{tM} + \frac{1}{\varrho_t}\chi_{it}, \end{aligned} \qquad \text{Model 2:} \quad \begin{aligned} \log r_{itI} &= \alpha_I + \delta\kappa_i + \gamma_{tI} + \varrho_t \chi_{it}, \\ \log r_{itM} &= \alpha_M + \frac{1}{\delta}\kappa_i + u_i + \gamma_{tM} + \frac{1}{\varrho_t}\chi_{it}, \end{aligned}$$

where $\alpha_d$ is a health outcome-specific intercept, $\delta$ is a scaling parameter, $\kappa_i$ represents the shared spatial component, $u_i$ is a spatially unstructured random effect, $\gamma_{td}$ represents the time effect specific for each health outcome $d$, $\varrho_t$ is a scaling parameter for each time $t$ and $\chi_{it}$ is the shared spatio-temporal interaction. The $\varrho_t$ scaling parameters do not need to be necessarily different for all times $t$, i.e., it is possible to define a total number of parameters $l \leq T$ and repeat them for certain periods. If this is the case, one defines

$$\boldsymbol{\varrho} = \text{diagonal}\left(\varrho_1 \mathbf{I}_{m_1}, \varrho_2 \mathbf{I}_{m_2}, \ldots, \varrho_l \mathbf{I}_{m_l}\right) \otimes \mathbf{I}_A \quad 1 \leq l \leq T,$$

where $\varrho_l$ are scaling parameters, $m_l$ is the number of years with the same scaling parameter $\varrho_l$ and $\mathbf{I}_{m_l}$ are identity matrices of size $m_l \times m_l$. We assume that the $\varrho_l$ scaling parameters are independent. The following prior distributions are used

$$\begin{aligned} &\alpha_d \sim N\left(0, 1/0.001\right), \quad d = I, M \qquad p(\mathbf{u}) \propto \exp\left(\frac{-\tau_u}{2}\mathbf{u}'\mathbf{I}_A\mathbf{u}\right), \qquad p(\boldsymbol{\chi}) \propto \exp\left(\frac{-\tau_\chi}{2}\boldsymbol{\chi}'\mathbf{Q}_\chi\boldsymbol{\chi}\right), \\ &\delta \sim \text{Gamma}(10, 10), \qquad p(\boldsymbol{\gamma}) \propto \exp\left(\frac{-\tau_\gamma}{2}\boldsymbol{\gamma}'\mathbf{R}_\gamma\boldsymbol{\gamma}\right), \\ &p(\boldsymbol{\kappa}) \propto \exp\left(\frac{-\tau_\kappa}{2}\boldsymbol{\kappa}'\mathbf{R}_\kappa\boldsymbol{\kappa}\right), \qquad \varrho_l \sim \text{Gamma}(10, 10), \quad 1 \leq l \leq T \end{aligned}$$

where $\mathbf{R}_\kappa$ is the well-known spatial neighbourhood structure matrix, $\mathbf{R}_\gamma$ is the temporal structure matrix of a first order random walk and $\mathbf{Q}_\chi$ represents any of the four spatio-temporal interaction types proposed by Knorr-Held [2]. For comparison purposes, we also consider multivariate spatio-temporal models with independent interactions, denoted by Model 0 and Model 0*. Namely

$$\text{Model 0:} \quad \begin{aligned} \log r_{itI} &= \alpha_I + \delta\kappa_i + \gamma_{tI} + \chi_{itI}, \\ \log r_{itM} &= \alpha_M + \frac{1}{\delta}\kappa_i + \gamma_{tM} + \chi_{itM}, \end{aligned} \qquad \text{Model 0}^*\text{:} \quad \begin{aligned} \log r_{itI} &= \alpha_I + \delta\kappa_i + \gamma_{tI} + \chi_{itI}, \\ \log r_{itM} &= \alpha_M + \frac{1}{\delta}\kappa_i + u_i + \gamma_{tM} + \chi_{itM}, \end{aligned}$$

where $\chi_{itd}$ are the spatio-temporal interactions specific for each health outcome $d$.

We remark that the new shared spatio-temporal models presented here are not directly available in INLA [3]. We have implemented them using the `rgeneric` model.

## 3. Illustration

In this work, we analyze jointly both pancreatic cancer and leukaemia incidence and mortality data in males during nine periods (2002-2003, . . . , 2018-2019) in 142 areas of Great Britain. For Model 1 and Model 2, we use different number of scaling parameters. First, we select the two extreme cases, i.e. the model with a single scaling parameter $l = 1$ (the most restrictive model) and the model with a different parameter for each time period $l = T$ (the most flexible model). Second, we select a specific number of scaling parameters depending on the cancer location. To select the number of scaling parameters we have performed an exploratory data analysis. Based on it, we take three different scaling parameters for pancreatic cancer, repeating each of them over three periods, and we consider seven scaling parameters for leukaemia, one for each period, except for the 5th and 6th period, and the 7th and 8th period where we define the same scaling parameter. We also consider the four interaction types. Although not shown here to conserve space, for pancreatic cancer type II interactions have been selected for Model 0, and type I interactions for Model 1 and Model 2, while for leukaemia, type III interactions have been selected as the best for all models. We have used the Deviance Information Criterion (DIC) [4], the Watanabe-Akaike Information Criterion (WAIC) [5] and the logarithmic score (LS) [6] criteria to select the best model among the different proposals.

Table 1 shows the model selection criteria values obtained by the best models at each cancer location. We can see that for pancreatic cancer Model 1 with a constant scaling parameter is the best model. For leukaemia similar results have been obtained using Model 2 with a different scaling parameter for each time period or with seven scaling parameters. After examining the values of the scaling parameters, we have selected Model 2 with seven scaling parameters as the best model.

Figure 1 shows the posterior mean of the rate estimates with Model 1 and a scaling parameter ($l = 1$) for pancreatic cancer incidence and mortality. To conserve space, we only present the time evolution of the geographical patterns of pancreatic cancer rate estimates. We observe an increase in pancreatic cancer incidence and mortality rates over the years. The increase in rates is first seen in the southern coast areas and this increase is spreading northwards. In 2018-2019 the areas with the lowest rates are located in central England.

Table 1: Model selection criteria.

| Pancreatic cancer | | | | | Leukaemia cancer | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | DIC | WAIC | LS | | | DIC | WAIC | LS |
| Model 0 | | 17070 | 17063 | 8545 | Model 0* | | 17161 | 17164 | 8661 |
| **Model 1** | $\boldsymbol{l = 1}$ | **16716** | **16523** | **8289** | Model 2 | $l = 1$ | 17098 | 17086 | 8619 |
| Model 1 | $l = T$ | 16791 | 16546 | 8298 | Model 2 | $l = T$ | 17086 | 17083 | 8610 |
| Model 1 | $l = 3$ | 16719 | 16524 | 8290 | **Model 2** | $\boldsymbol{l = 7}$ | **17088** | **17078** | **8609** |

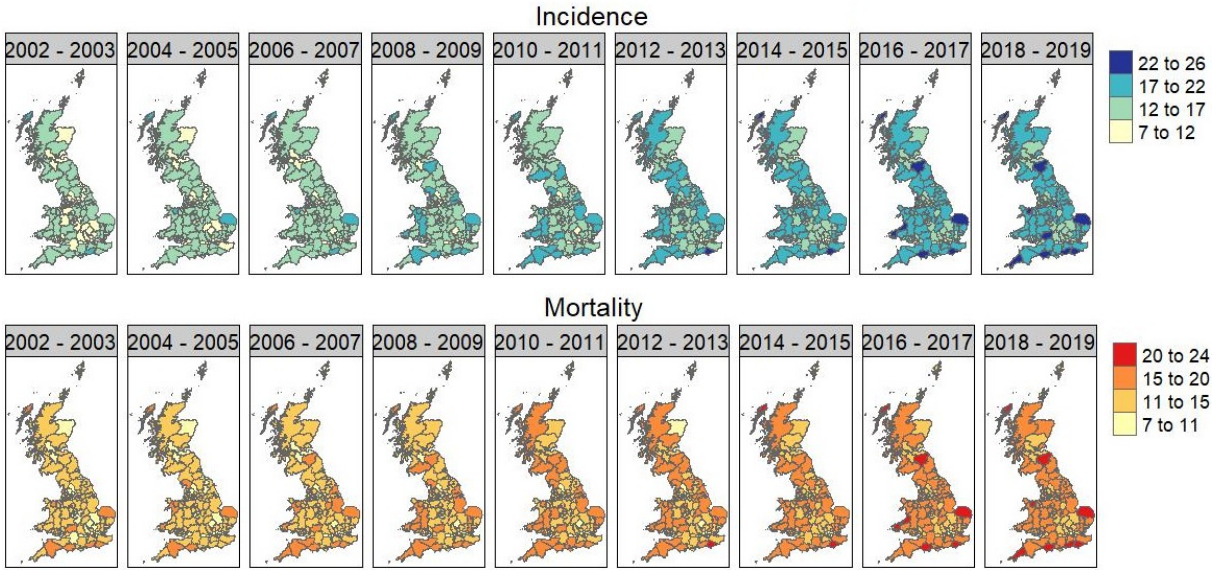


Figure 1: Temporal evolution of geographical patterns of pancreatic cancer rates with Model 1.

## 4. Conclusions

Estimating rare or less frequent cancer sites becomes a methodological challenge. We propose the use of multivariate spatio-temporal models with time-varying shared interaction terms to maximize the information coming from disease incidence and mortality. The results show that for rare cancer sites, multivariate spatio-temporal models with shared interactions perform better than the usual multivariate spatio-temporal models with independent interactions.

## 5. Bibliography

[1] Held L., Natário I., Fenton S.E., Rue H., and Becker N. (2005). Towards joint disease mapping. *Statistical Methods in Medical Research, 14*(1), 61-82.

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

[3] 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.

[4] 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.

[5] Watanabe S., and Opper M. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research, 11*(12), 3571-3594.

[6] Gneiting T., and Raftery A.E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association, 102*(477), 359-378.