Received: 21 July 2019 | Revised: 20 December 2019 | Accepted: 3 January 2020

DOI: 10.1002/bimj.201900229

Biometrical Journal

RESEARCH PAPER

A model with space-varying regression coefficients for clustering multivariate spatial count data

Francesco Lagona[1,2] | **Monia Ranalli**[3] | **Elisabetta Barbi**[3]

[1]Department of Political Sciences, Roma Tre University, Rome, Italy

[2]Department of Mathematics, University of Bergen, Bergen, Norway

[3]Department of Statistical Sciences, Sapienza University of Rome, Rome, Italy

Correspondene

Francesco Lagona, Department of Political Sciences, Roma Tre University, Via G. Chiabrera 199, 00145 Rome, Italy.

Email: francesco.lagona@uniroma3.it

Funding information

MIUR, Grant/Award Number: PRIN2015

This article has earned an open data badge "**Reproducible Research**" for making publicly available the code necessary to reproduce the reported results. The results in this paper could not be fully reproduced due to insufficient transparency of the submitted code and their computational cost.

Abstract

Multivariate spatial count data are often segmented by unobserved space-varying factors that vary across space. In this setting, regression models that assume space-constant covariate effects could be too restrictive. Motivated by the analysis of cause-specific mortality data, we propose to estimate space-varying effects by exploiting a multivariate hidden Markov field. It models the data by a battery of Poisson regressions with spatially correlated regression coefficients, which are driven by an unobserved spatial multinomial process. It parsimoniously describes multivariate count data by means of a finite number of latent classes. Parameter estimation is carried out by composite likelihood methods, that we specifically develop for the proposed model. In a case study of cause-specific mortality data in Italy, the model was capable to capture the spatial variation of gender differences and age effects.

KEYWORDS

cause-specific mortality, composite likelihood, hidden Markov field, model-based clustering, Potts model

1 | INTRODUCTION

Multivariate spatial count data arise when multiple counts are observed across the partitioning of an area of interest. These data are routinely collected across many disciplines such as ecology, epidemiology, demography, and geography. A popular example is provided by multivariate spatial series of mortality data, obtained by collecting cause-specific death counts across administrative geographical units and often tabulated according to different demographic variables, such as age and gender.

The analysis of spatial data often involves the estimation of the effects of space-constant covariates under space-varying conditions. When, for example, death counts are observed across the administrative units of an area of interest, mortality rates are often spatially segmented according to space-varying conditions that reflect environmental and socioeconomic differences across the area under study. In these cases, there is interest in assessing whether the fixed effect of a space-constant covariate (e.g., gender differences) varies with these conditions.

If these conditions are satisfactorily captured by a battery of observable geo-referenced covariates, then covariate effects under different conditions can be efficiently estimated by specifying a regression model for counts that includes terms of interaction between the covariates of interest and the observed conditions. However, it is hardly plausible that all the factors acting on the underlying cause-specific mortality risk can be identified or measured at the required geographic level. Thus, there often remains residual heterogeneity in the death event rate, which moreover is likely to have a spatial structure inherited from some of the unmeasured or undiscovered risk factors for mortality.

In this case, a natural strategy relies on assuming that the regression coefficients are drawn from a latent spatial process. In a univariate setting, for example, Assunção (2003) introduces a Bayesian Poisson model for a spatial series of counts, where the regression coefficients are drawn from a Gaussian Markov random field, extending the popular model by Besag, York, and Mollié (1991) where only intercept values are allowed to vary across the area under study. Multivariate extensions of this approach have been developed within a Bayesian setting by assuming that the regression coefficients are drawn from multivariate Markov random fields (Gelfand & Vounatsou, 2003). These models are often referred to as hierarchical space-varying coefficients regression models. A key feature of these models is the spatial Gaussian process that generates the regression coefficients, which, therefore, vary smoothly across space. On the one side, assuming that covariates' effects vary smoothly in space can be helpful to smooth extreme mortality rates that are associated with the less populated areas, with no relationship to the underlying risk. On the other side, the parameters characterizing the spatial dependence of the regression coefficients are typically constant across the entire study region. As a result, there is the potential risk of oversmoothing and masking of local discontinuities due to the global effect of the parameters. When the interest is on local discontinuities of covariates' effect, a natural approach relies on replacing a Gaussian random field by a multinomial random field with a finite number of classes. This allows to segment the effects of covariates across space.

Replacing a Gaussian process by a discrete multinomial process brings two advantages. First, it provides a nonparametric alternative to parametric random effects models. Parametric assumptions on the random effects are convenient but they are often unverifiable and may be a source of model misspecification. A discrete process allows to estimate the nonparametric distribution of the regression coefficients, avoiding parametric assumptions. A second advantage is the parsimonious representation of the random effects. Under a traditional Gaussian approach, each observation is associated to a random effect, which can be predicted using the model estimates. In a spatial setting, the predicted random effects are geo-referenced and they are typically displayed in a choropleth map. This is done by clustering the predicted random effects according to a finite number of grey levels or colors. Therefore, although the random effects take different values, they are conveniently (but arbitrarily) clustered within classes, to facilitate visual inspection and interpretation. Under a discrete approach, random effects are automatically clustered in an optimal number of latent classes that are jointly estimated with the model parameters.

In a univariate setting, latent multinomial processes have been exploited by Green and Richardson (2002), who introduced a Bayesian Poisson regression model, whose intercept values are allowed to vary across the area of interest according to a Potts model, that is, a parametric spatial multinomial process. Models such as these are often referred to as hidden Markov field models and are typically associated with a numerically intractable likelihood function that complicates parameter estimation and inference. In a Bayesian setting, an intractable likelihood typically leads to an intractable posterior distribution of the parameters. Bayesian methods therefore provide a natural estimation strategy only if an efficient algorithm is available for sampling from the posterior distribution. In general, sampling from the posterior parameter distribution under a hidden Markov field is not straightforward (Green & Richardson, 2002). Algorithms depend on an empirical calibration of the hyperparameters of the prior distributions. In addition, special care is required to account for the high correlation a posteriori between the parameters. Finally, numeric approximations are often necessary to handle intractable normalizing constants. In this setting, a frequentist approach is a viable alternative to Bayesian methods if the likelihood can be efficiently approximated by an optimization function (e.g., a pseudo-likelihood) that can be easily maximized (Alfò, Nieddu, & Vicari, 2009; Klauenberg & Lagona, 2007). This approach is usually pursued by estimating the model parameters by EM-type algorithms, obtained by a mean-field approximation of the complete-data log-likelihood function (Lagona & Picone, 2016). Although this method is computationally feasible, little is known about its distributional properties.

Hidden Markov random fields for mortality count data are typically specified by including a random intercept that varies across space, while additional regression coefficients are kept fixed. We are, however, motivated by a case study of Italian cause-specific mortality data, where the interest is in the spatial segmentation of age and gender effects according to a finite number of latent classes. We therefore propose a hidden Markov model where all the regression coefficients are allowed to vary across space. In a longitudinal setting, hidden Markov models with time-varying regression coefficients have been recently proposed by Lagona, Jdanov, and Shkolnikova (2014). The approach taken in this paper can be then viewed as a spatial extension of such proposal. In the temporal setting, computational methods for maximum likelihood estimation of hidden Markov models are well known, but extending these methods to the spatial case is not obvious. We therefore propose a computationally feasible EM algorithm to estimate the parameters of the proposed model, by relying on composite likelihood (CL) methods that have been recently developed for hidden Markov fields (Ranalli, Lagona, Picone, & Zambianchi, 2018).

The rest of the paper is organized as follows. Section 2 describes the mortality data that motivated this work. Section 3 is instead devoted to the proposed model. The estimation procedure is detailed in Section 4. Section 5 illustrates the results that we obtain on the motivating case study. Finally, Section 6 summarizes some relevant points of discussion.

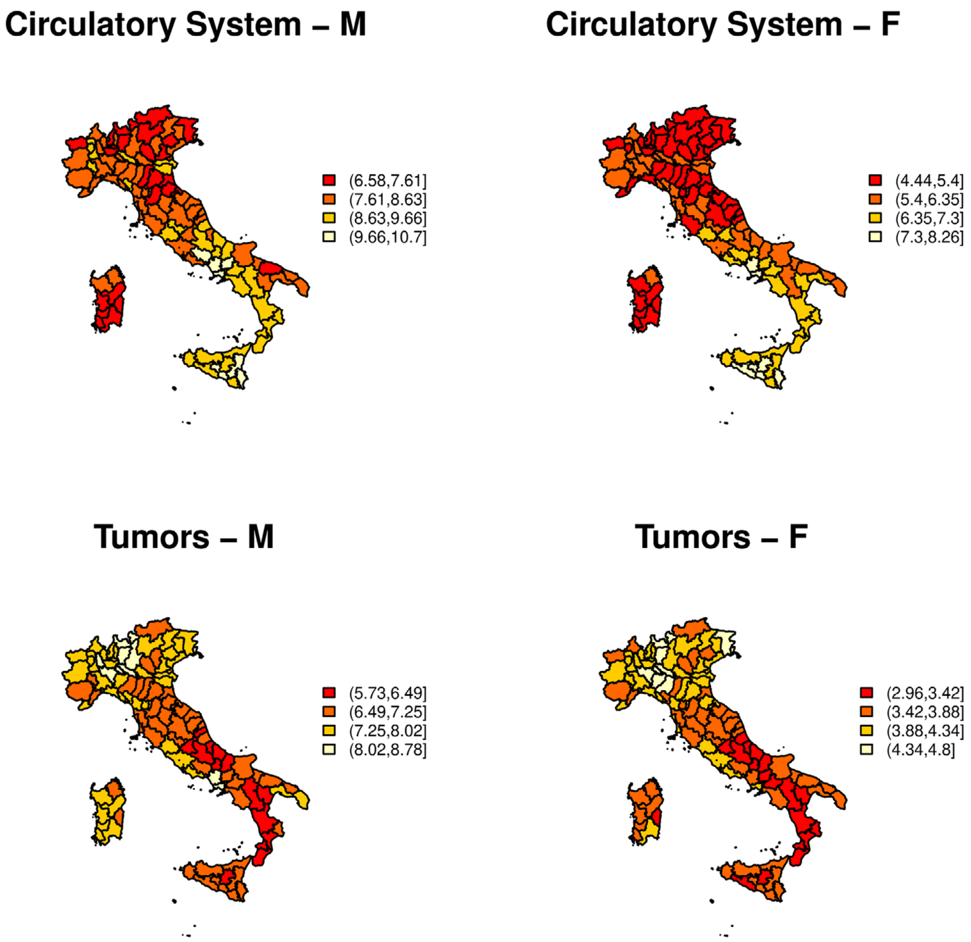


FIGURE 1 Top: SMRs (×1,000) due to circulatory diseases for males (left) and females (right). Bottom: SMRs (×1,000) due to tumor for males (left) and females (right)

2 | CAUSE-SPECIFIC MORTALITY IN ITALY

The data that motivated this paper are spatial, cause-specific mortality data observed in Italy in the period 2010–2013. During this period, Italy was divided into 110 administrative units, known as provinces. For each province, we obtained the number of cause-specific deaths for single year of age (from 40 to 100) and separately for women and men. Death counts were clustered according to the main four groups of causes: diseases of the circulatory system, malignant neoplasms (tumors), diseases of the respiratory system, and diseases of the digestive system. Cause-specific mortality data specifically refer to the X International Classification of Diseases (ICD10). Diseases of the circulatory system refer to ICD10: I000–I990; malignant neoplasms refer to ICD10: C000–D489; diseases of the respiratory system refer to ICD10: J000–J998; diseases of the digestive system refer to ICD10: K000–K938. We also obtained exposure data, in the form of person years, counted in each age–gender stratum of the population and each administrative unit, during the observation period.

The availability of cause-specific mortality data at a provincial level provides an opportunity for comparing the geographical variation of mortality risks between different causes, by computing standardized mortality rates (SMRs) that adjust for the confounding effects of age and gender. To enhance the readability, the SMRs have been multiplied by 1,000.

Figure 1 displays the gender-specific SMRs due to cancers and cardiovascular diseases, which are the two leading causes of death in Italy (about two thirds of the total deaths in Italy are due to these two causes). The geographical patterns of these two leading groups of causes show a quite accentuate but opposite north–south gradient. In the case of malignant neoplasms, the excess of mortality is recorded, for both women and men, primarily in the northern regions, reflecting a reverse relationship between socioeconomic development and mortality. The diseases of the circulatory system show a totally inverted geographical profile: the excess of mortality for both sexes is recorded mostly in the southern regions, which have long been economically and socially disadvantaged with respect to the northern ones. The geographical patterns of these two leading causes of death are similar for the two sexes, but the intensity of each cause is very different for women and men: for men, mortality from cancer

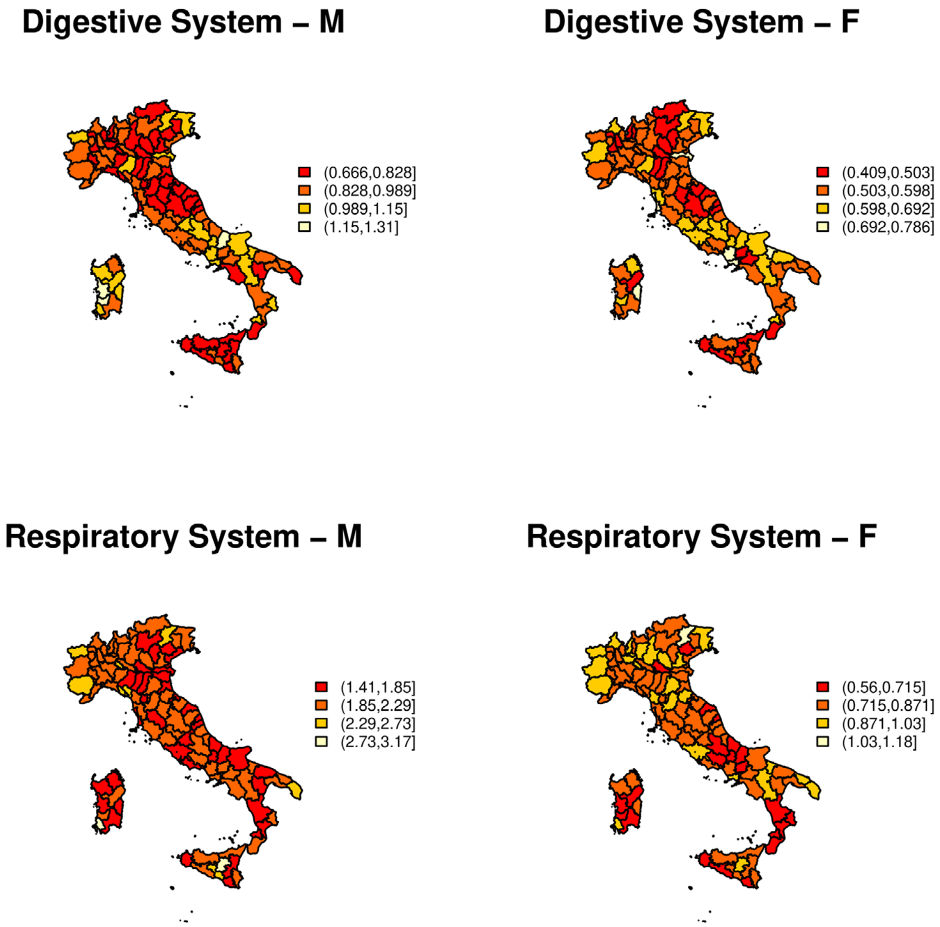


FIGURE 2 Top: SMRs (×1,000) due to digestive diseases for males (left) and females (right). Bottom: SMRs (×1,000) due to respiratory diseases for males (left) and females (right)

can be twice the level observed for women, while male and female levels of mortality from cardiovascular diseases are similar (Lagona & Barbi, 2019). Thus, for men in the north, the disadvantage is mainly due to mortality from cancers whose risk factors are linked to the occupations and life styles (high alcohol consumption, early spread of smoking, etc.) typical of more developed and industrialized provinces. For women in the south, the disadvantage is due to mortality from cardiovascular disease, which is prevalent among women in the southern provinces where health care is generally less adequate and women are generally less educated and economically less independent (Barbi, Casacchia, & Racioppi, 2018).

Figure 2 displays the SMRs due to diseases of the digestive system (top panel) and the respiratory system (bottom panel). Deaths due to diseases of the digestive system show a more mixed spatial picture, especially for women, with areas of higher mortality scattered through the Italian territory.

The geographical distribution of mortality due to respiratory diseases shows instead a clear north–south gradient, for both sexes but especially for women. Such a geographical pattern is somehow expected as the northern provinces are more exposed to pollution and heavy tobacco consumption.

SMRs are helpful to explore the geographical variation of mortality risk across administrative units. Further insights can be however extracted by looking at the geographical variation of age and gender effects on mortality. Accordingly, in the next section we introduce a regression model with space-varying regression coefficients that can be exploited for exploring the geographical variation of covariate effects on spatial multivariate count data.

3 | A MULTIVARIATE HIDDEN MARKOV RANDOM FIELD MODEL

Our proposal can be formally introduced by describing multivariate spatial count data as a multidimensional array $(y_{shi}, s = 1, \dots, S; h = 1, \dots, H; i = 1, \dots, I)$, where y_{shi} indicates the number of events (e.g., deaths) due to the hth cause and occurred within the sth stratum of the population in the ith administrative unit of an area of interest. These data are associated with the array $(n_{si}, s = 1, \dots, S; i = 1, \dots, I)$, where n_{si} indicates the number of person-years in the sth stratum and in the ith administrative unit. In our application, strata are specified by combining age and gender groups. By allowing for a generic stratum, however, we are able to simplify the following notation and to extend our proposal to general stratification structures. Our proposal integrates a battery of Poisson regressions with a spatial multinomial process (the Potts model), which are, respectively, described in the following two subsections.

3.1 | Poisson regression

By viewing deaths as events arising from a Poisson process, y_{shi} can be interpreted as the number of events that have occurred during an exposure period of n_{si} person-years, in which the occurrence rate is assumed to be λ_{shi}. Under this setting, y_{shi} is Poisson distributed with rate $n_{si}\lambda_{shi}$ (Armitage, 1966; Brillinger, 1986).

Differences between cause-specific mortality rates across strata are usually examined by modulating Poisson rates in terms of a suitable $S \times (P+1)$ design matrix $\mathbf{X}$, namely,

$$\log \lambda_{shi} = \log n_{si} + \beta_{0h} + x_{s1}\beta_{1h} +, \dots, + x_{sP}\beta_{Ph} \quad s = 1, \dots, S; \quad h = 1, \dots, H; \quad i = 1, \dots, I, \tag{1}$$

where $(1, x_{s1}, \dots, x_{sP})^{\mathsf{T}}$ is the sth row of $\mathbf{X}$ and $(\beta_{0h}, \beta_{1h}, \dots, \beta_{ph}, \dots, \beta_{Ph})$ is a vector of cause-specific regression coefficients. For example, if $P = 2$ and, respectively, x_{s1} and x_{s2} indicate the age group and the gender of stratum s, then model (1) reduces to a proportional-hazard Gompertz model, where the mortality rate exponentially increases with age, while females and males share the same rate of aging.

Model (1) can be extended by allowing the regression coefficients to vary according to the values taken by another variable, say r_{si}. In our context, a varying-coefficient model has the form

$$\log \lambda_{shi} = \log n_{si} + \beta_{0h}(r_{si}) + x_{s1}\beta_{1h}(r_{si}) +, \dots, + x_{sP}\beta_{Ph}(r_{si}), \tag{2}$$

where $s = 1, \dots, S$, $h = 1, \dots, H$ and $i = 1, \dots, I$. The variable r_{si} indicates the conditions under which stratum s is observed in the ith administrative unit, playing the role of an "effect modifier" through the functions $\beta_{ph}(\cdot)$, $h = 1, \dots, H, p = 0, \dots, P$. Varying-coefficient models such as (2) have been introduced by Hastie and Tibshirani (1993). The unrestricted nonparametric estimation of functions $\beta_{ph}(\cdot)$ involves a number of parameters that increases with the sample size, and it is hence not recommended. This issue is typically addressed by imposing suitable constraints, pursuing strategies in a parametric or nonparametric framework. On the one side, in a parametric framework, functions β can be assumed known up to a finite number of parameters to be estimated, using, for example, a piecewise constant model or a more complex functional form. On the other side, in a nonparametric framework, functions β can be assumed unknown but smooth. The latter option flexibly accommodates nonlinear effects without assuming a specific parametric form of the predictors.

3.2 | The Potts model

Model (2) is formulated by assuming that the effect modifier r_{si} is observed. In our case study, however, the effect modifier is not available, because we do not have a variable that proxies the environmental conditions under which the data are observed. We therefore propose to replace the effect modifier r_{si} by an unobserved (latent) multinomial random variable with one trial and K classes, say $\mathbf{u}_{si} = (u_{si1}, \dots, u_{siK})$, where u_{sik} is a binary variable indicating class membership. Specifically, we assume that

$$\beta_{ph}(\mathbf{u}_{si}) = \sum_{k=1}^{K} u_{sik}\beta_{phk}, \quad h = 1, \dots, H; \quad p = 0, \dots, P; \quad i = 1, \dots, I. \tag{3}$$

Equation (3) segments the effect of the pth covariate (specific of the hth cause, the sth stratum and the ith administrative unit) according to K unknown levels $\beta_{ph1}, \dots, \beta_{phK}$, to be estimated. Under this setting, covariate effects are clustered according to K latent classes of regression coefficients, which represent covariate effects under unobserved conditions.

We could ignore the spatial structure of the data and assume that the multinomial variables $\mathbf{u}_{si}$ are independent, in keeping with the usual assumptions of latent class analysis. However, it is quite possible that covariate effects in neighboring sites are similar, because of the spatial autocorrelation between latent conditions shared by adjacent areas. Spatial autocorrelation introduces redundant information that should be accounted for when clustering regression coefficients.

We allow for spatial autocorrelation by assuming that the random variables $\mathbf{u}_{s1}, \dots \mathbf{u}_{sI}$ are the I dependent components of a spatial array $\mathbf{u}_s$, drawn from the joint distribution of a homogeneous Markov field with K classes. Array $\mathbf{u}_s$ can be interpreted as the spatial pattern of the conditions under which the sth stratum of the sample is observed.

Markov random fields can be specified in several ways (Gaetan & Guyon, 2010). In the present paper, we concentrate on the Potts model (Strauss, 1977), a spatial allocation model originally exploited in image processing applications (Tjelmeland & Besag, 1998). The Potts model has been successfully implemented in disease mapping studies (Alfò et al., 2009; Green & Richardson, 2002). The model depends on a neighborhood structure among the administrative units, obtained by associating each administrative unit with the set $N(i)$ of its neighbors. In our application, two administrative units are neighbors if they share a boundary, but other definitions are possible. The Potts model also depends on a battery of sufficient statistics. By taking the first class as a reference, we associate each sample $\mathbf{u}_s$ with K sufficient statistics: the number of neighboring sites that share the same class $k \neq 1$

$$n(\mathbf{u}_s) = \sum_{i=1}^{n} \sum_{j>i:j\in N(i)} \sum_{k=2}^{K} u_{sik}u_{sjk},$$

and $K - 1$ sufficient statistics

$$n_k(\mathbf{u}_s) = \sum_{i=1}^{n} u_{sik} \quad k = 2, \dots, K,$$

which indicate the number of neighboring sites that are associated with latent class k. Under a Potts model, the probability of a specific sample $\mathbf{u}_s$ is known up to K parameters $\alpha_2 \dots \alpha_K, \rho$ and it is given by

$$p(\mathbf{u}_s; \boldsymbol{\alpha}, \rho) = \frac{\exp\left(\sum_{k=2}^{K} n_k(\mathbf{u}_s)\alpha_k + n(\mathbf{u}_s)\rho\right)}{W(\alpha, \rho)}, \tag{4}$$

where $W(\boldsymbol{\alpha}, \rho)$ is the normalizing constant. The parameter ρ is an autocorrelation parameter: if it is positive (negative) then it penalizes spatial patterns with a few concordant (discordant) neighbors. This parameter is often referred to as a regularization parameter, given that large values of ρ are associated with patterns where areas with the same label are geometrically regular. Each parameter α_k penalizes patterns with a few sites that belong to class k. When $\rho = 0$, then (4) reduces to a multinomial distribution where the parameters α are class-specific log-odds:

$$\alpha_k = \log \frac{P(u_{sik} = 1)}{P(u_{si1} = 1)}, \quad k = 2, \dots, K.$$

Under the Potts model (4), the conditional distribution of each site depends only on the labels taken by the neighboring sites, namely,

$$p(u_{sik} = 1 \mid \mathbf{u}_{s1}, \dots \mathbf{u}_{s,i-1}, \mathbf{u}_{s,i+1}, \dots \mathbf{u}_{sI}) = \frac{\exp\left(\alpha_k + \rho n_k(\mathbf{u}_{s,N^+(i)})\right)}{1 + \sum_{k=2}^{K} \exp\left(\alpha_k + \rho n_k(\mathbf{u}_{s,N^+(i)})\right)}, \tag{5}$$

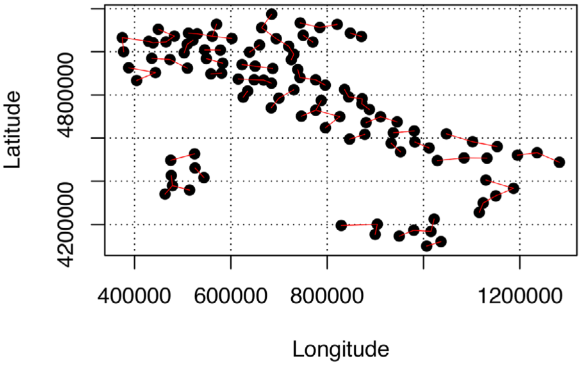
where $n_k(\mathbf{u}_{s,N^+(i)}) = u_{sik} \sum_{j\in N(i)} u_{sjk}$ is the number of sites in the neighborhood of i that are labeled by k, being $N^+(i) = N(i) \cup \{i\}$. Accordingly, the model is a Markov random field with respect to the neighborhood structure, and the autocorrelation coefficient ρ can be viewed as an autoregression coefficient that is associated with the spatially lagged outcome $n_k(\mathbf{u}_{s,N^+(i)})$.

3.3 | A hidden Markov model for multivariate counts

The specification of the hidden Markov field is completed by assuming that the observed death counts within each stratum of the sample are conditionally independent given the spatial latent process. Precisely, let $\mathbf{y}_s = (\mathbf{y}_{si}, i = 1, \dots I)$ be the multivariate spatial series of death counts, observed within the stratum s of the sample, where $\mathbf{y}_{si} = (y_{shi}, h = 1, \dots, H)$ includes the deaths counts due to each cause of death. Under this setting, the joint distribution of the multivariate spatial series $\mathbf{y}_s$ is given by

$$p(\mathbf{y}_s) = \sum_{\mathbf{u}_s} p(\mathbf{u}_s) \prod_{i=1}^{I} \prod_{k=1}^{K} \left(p(\mathbf{y}_{si} \mid u_{sik})\right)^{u_{sik}}, \tag{6}$$

FIGURE 3 The pairs of neighboring sites, on which the composite likelihood is based. *Note*: Points indicate the centroids of the Italian provinces



where

$$p(\mathbf{y}_{si} \mid u_{sik}) = \prod_{h=1}^{H} p(y_{shi} \mid u_{sik}),$$

whereas $p(y_{shi} \mid u_{sik})$ is a Poisson distribution with mean λ_{shik} and

$$\log \lambda_{shik} = \log n_{si} + \beta_{0hk} + x_{s1}\beta_{1hk} +, \ldots, +x_{sP}\beta_{Phk} \tag{7}$$

with $s = 1, \ldots, S$, $h = 1, \ldots, H$, $i = 1, \ldots, I$, and $k = 1, \ldots, K$.

The likelihood function of the proposed model is given by

$$L(\boldsymbol{\theta}) = \prod_{s=1}^{S} p(\mathbf{y}_s; \boldsymbol{\theta}) = \prod_{s=1}^{S} \sum_{\mathbf{u}_s} p(\mathbf{u}_s; \boldsymbol{\alpha}, \rho) \prod_{i=1}^{I} \prod_{k=1}^{K} \left(p(\mathbf{y}_{is} \mid u_{isk}; \boldsymbol{\beta})\right)^{u_{isk}}, \tag{8}$$

where $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \rho, \boldsymbol{\beta})$.

This likelihood function can be viewed as a spatial extension of the likelihood function of a hidden Markov model with time-varying coefficients that evolve according to a latent Markov chain (Lagona et al., 2014). The likelihood function of hidden Markov models is numerically intractable in both the temporal and spatial setting. However, in the temporal setting, hidden Markov models can be estimated by exploiting specific recursive methods in order to avoid unpractical summations over the state space of latent Markov chains and numerical underflows and overflows. These recursive methods cannot be extended to the spatial setting, and, as a result, alternative methods are needed.

4 | ESTIMATION

4.1 | Composite likelihood methods

Direct maximization of the likelihood function (8) is unfeasible, as it requires a summation over K^I latent segmentations and $p(\mathbf{u}_s; \boldsymbol{\alpha}, \rho)$ depends on an intractable normalizing constant, which is a function of unknown parameters. To obtain parameter estimates, we propose to maximize a surrogate function, namely, a CL function (Lindsay, 1988). Following Ranalli et al. (2018) and Ameijeiras-Alonso, Lagona, Ranalli, and Crujeiras (2019), our proposal relies on covering the set $\{1 \ldots I\}$ of the administrative units by all the pairs $N_2 = \{i, j\}$ of neighbors, that is, such that $i \in N(j)$ and $j \in N(i)$. Figure 3 shows the neighboring pairs that we have chosen in our application.

For each subset $\{i, j\}$, we define

$$L_{N_2}(\boldsymbol{\theta}) = \prod_{s=1}^{S} p(\mathbf{y}_s(N_2); \boldsymbol{\theta}) = \prod_{s=1}^{S} \sum_{\mathbf{u}_s(N_2)} p(\mathbf{u}_s(N_2); \boldsymbol{\alpha}, \rho) \prod_{i \in N_2} \prod_{k=1}^{K} \left(p(\mathbf{y}_{is} \mid u_{isk}; \boldsymbol{\beta})\right)^{u_{isk}}$$

as the contribution of the data in N_2 to the CL function, where $\mathbf{u}_s(N_2) = \{\mathbf{u}_{is} : i \in N_2\}$ and $\mathbf{y}_s(N_2) = \{\mathbf{y}_{is} : i \in N_2\}$, while

$$p(\mathbf{u}_s(N_2); \alpha, \rho) = \frac{\exp\left(\sum_{k=2}^{K} n_k(\mathbf{u}_s(N_2))\alpha_k + n(\mathbf{u}_s(N_2))\rho\right)}{W_2(\alpha, \rho)},$$

with $W_2(\alpha, \rho)$ being the normalizing constant, is a two-site nonhomogeneous Potts model. The normalizing constant of these probabilities is now numerically tractable, as it involves a summation over two sites.

We propose to estimate the parameters by maximizing the following composite log-likelihood function

$$c\ell(\theta) = \sum_{N_2} \log L_{N_2}(\theta). \tag{9}$$

4.2 | EM algorithm

To carry out the parameter estimates, we use an EM algorithm that iteratively generates a sequence of parameter values by alternating an E step and an M step, until convergence. During the tth E step, it computes the expected value of the complete-data composite log-likelihood with respect to the predictive distribution of the segmentation. The tth step reduces to the computation of the following expected value for each pair of sites

$$\mathbb{E} \log L_{N_2}(\boldsymbol{\theta}) = \sum_{s=1}^{S} \sum_{i \in N_2} \sum_{k=1}^{K} \hat{u}_{isk}^{(t)} \log p(\mathbf{y}_{is} \mid u_{isk}; \boldsymbol{\beta}) + \sum_{s=1}^{S} \sum_{\mathbf{u}_s(N_2)}^{(t)} \hat{\mathbf{u}}_s^{(t)}(N_2) \log p(\mathbf{u}_s(N_2); \boldsymbol{\alpha}, \rho),$$

where the $K \times K$ predictive probabilities computed at the tth step, $\hat{\mathbf{u}}_s^{(t)}(N_2)$ are given by

$$\begin{aligned} \hat{\mathbf{u}}_s^{(t)}(N_2) =& p(\mathbf{u}_s^{(t)}(N_2) \mid \mathbf{y}_s(N_2), \hat{\boldsymbol{\beta}}^{(t-1)}, \hat{\boldsymbol{\alpha}}^{(t-1)}, \hat{\rho}^{(t-1)}) \\ =& \frac{p(\mathbf{u}_s^{(t)}(N_2); \hat{\boldsymbol{\alpha}}^{(t-1)}, \hat{\rho}^{(t-1)}) p(\mathbf{y}_s(N_2) \mid \mathbf{u}_s^{(t)}(N_2); \hat{\boldsymbol{\beta}}^{(t-1)})}{\sum_{\mathbf{u}_s^{(t)}(N_2)} p(\mathbf{u}_s^{(t)}(N_2); \hat{\boldsymbol{\alpha}}^{(t-1)}, \hat{\rho}^{(t-1)}) p(\mathbf{y}_s(N_2) \mid \mathbf{u}_s^{(t)}(N_2); \hat{\boldsymbol{\beta}}^{(t-1)})}, \end{aligned} \tag{10}$$

whereas $\hat{\boldsymbol{\theta}}^{(t-1)}$, $\hat{\boldsymbol{\alpha}}^{(t-1)}$ and $\hat{\rho}^{(t-1)}$ are the parameter values that were available from the previous step of the algorithm. Suitable marginalization of (10) provides the univariate probabilities for the stratum sth

$$\hat{\mathbf{u}}_{is}^{(t)} = p(\mathbf{u}_{is} \mid \mathbf{y}_s(N_2), \hat{\boldsymbol{\theta}}^{(t-1)}, \hat{\boldsymbol{\alpha}}^{(t-1)}, \hat{\rho}^{(t-1)}), \text{ with } i = 1, \dots, I. \tag{11}$$

During the M step, the algorithm maximizes the expected complete-data composite log-likelihood with respect to the unknown parameters. Because this function is the sum of two components that depend on different sets of parameters, the M-step reduces to the separate maximization of two functions, namely,

$$Q(\boldsymbol{\beta}) = \sum_{s=1}^{S} \sum_{N_2} \sum_{i \in N_2} \sum_{k=1}^{K} \hat{u}_{isk} \log p(y_{is} \mid u_{isk}; \boldsymbol{\beta}_k) \tag{12}$$

$$Q(\boldsymbol{\alpha}, \rho) = \sum_{s=1}^{S} \sum_{N_2} \sum_{\mathbf{u}_s(N_2)} \hat{\mathbf{u}}_s(N_2) \log p(\mathbf{u}_{s,N_2}; \boldsymbol{\alpha}, \rho). \tag{13}$$

Maximization of both $Q(\boldsymbol{\beta})$ and $Q(\boldsymbol{\alpha}, \rho)$ can be carried out by a standard optimization routine, such as the quasi-Newton method in the `optim` command in `R`. The source code can be found online in the Supporting Information.

4.3 | Additional details

The proposed CL has been defined by covering the study area with pairs of neighboring sites. Covering the area by a larger subset might have been an option. However, the numerical tractability of the EM algorithm dramatically decreases with the cardinality of the largest subset of the cover. A cover that includes subsets with two elements is therefore a natural strategy. When the cover includes all the subsets of two elements, Equation (9) reduces to the pairwise likelihood function (Varin, Reid, & Firth, 2011). In

TABLE 1 Model selection

K	$c\ell$ value	cCLC
2	−3,263,277	6,520,767
3	−3,259,609	6,506,486
4	−3,262,882	6,511,278

a spatial setting, a pairwise likelihood can be further simplified by discarding all the pairs $\{i, j\}$ that do not include neighboring sites. Ranalli et al. (2018) provide an extensive simulation study that shows that this choice provides a computationally efficient EM algorithm, without a relevant loss in statistical efficiency.

It is well known that the EM algorithm suffers from two drawbacks: it is sensitive to the choice of starting points and it may converge to local maxima. These two aspects are strictly linked to each other. To avoid local maxima we follow a short-run strategy, by running the EM algorithm from 50 random initializations, and stopping the algorithm without waiting for full convergence, that is, when the relative increase in two consecutive composite log-likelihoods is less than 10^{-2}. The best solution is taken as starting point to run the EM algorithm until full convergence, that is when the difference in two consecutive composite log-likelihoods is less than 10^{-5}.

Standard errors could in principle be obtained by numerically approximating the observed Godambe (1960) matrix, which is, however, known to present numerical instability. This computation requires both the numerical approximation of variability and sensitivity matrices, and the inversion of the variability matrix (i.e., the covariance of the CL score), that it is usually a large-size matrix. A feasible alternative can be found in parametric bootstrap methods, to obtain the standard errors of the estimates. In this paper, we re-fitted the model to $R = 500$ bootstrap samples, which were simulated from the estimated model parameters. We then computed the standard deviation of the empirical distribution of each bootstrap estimate.

The number K of latent classes was chosen by selecting the model minimizing the composite integrated classification likelihood (cCLC; Ranalli & Rocci, 2016). The cCLC statistic extends the integrated classification likelihood by Biernacki and Govaert (1997) to the CL framework, and it reduces to the composite log-likelihood $c\ell(\hat{\boldsymbol{\theta}})$, evaluated at the maximum CL estimates, and penalized by the entropy $E(\hat{\mathbf{u}})$ of the class membership probabilities $\hat{u}$, as follows:

$$cCLC = -2c\ell(\hat{\boldsymbol{\theta}}) + 2E(\hat{\mathbf{u}}). \tag{14}$$

According to Table 1, a model with three components attains the minimum cCLC value.

5 | RESULTS

The Italian mortality data described in Section 2 have been fitted by a mixture of conditionally independent, cause-specific regressions whose parameters depend on the levels taken by latent Markov field with $K = 3$ states. The number of latent states has been selected by cCLC (see Table 1). Each regression model includes effects of age and gender and an age/gender interaction effect. The state-specific parameter estimates are displayed in Table 2.

As expected, the effect of age on mortality is always positive, with noticeable differences across causes of death. Age has a strong effect on mortality due to diseases in the circulatory and respiratory system, while this effect is less pronounced for deaths due to tumors and digestive diseases. However, age effects are not constant across space and, accordingly, they are clustered by the model into three latent classes, associated with the states of the estimated hidden Markov field. Class 1 is characterized by strong and significant effects of age, while under class 3 such effects are moderate (and even nonsignificant in the case of tumors). State 2 can be viewed as an intermediate state, with a non-significant age effect for mortality due to cancer, to be compared with the significant effects of age under the remaining three causes of death.

Compared to the estimated effects of age, the pattern of gender effects is more diversified across causes of deaths and latent classes. Under class 2, the advantage of the females over the males is always strong and significant, although modulated across causes of death. This advantage is attenuated and not significant in the case of cancer mortality. Classes 1 and 3 are instead characterized by distinct patterns of gender differences. Under class 1, the mortality risk among females is significantly lower than males in the case of deaths due to diseases in the digestive and the circulatory systems, while gender differences are not significant for tumors and respiratory diseases. Under class 3, the pattern is reversed: males have a lower mortality risk than females. However, the significantly negative estimate of α_3 indicates that only a small portion of data are clustered under this class, which should be therefore considered as a class of outliers.

TABLE 2 Estimated parameters and standard errors under a hidden Markov field with three states

		Latent states		
	Estimate	$k = 1$	$k = 2$	$k = 3$
Tumors	Intercept	−12.021 (1.113)	−9.295 (2.724)	−9.021 (3.107)
	Age	0.104 (0.016)	0.067 (0.037)	0.054 (0.040)
	Gender (ref: male)	1.783 (1.01)	−1.184 (1.658)	7.561 (1.78)
	Interaction	−0.035 (0.012)	0.007 (0.021)	−0.086 (0.021)
Circulatory system	Intercept	−13.126 (0.607)	−14.393 (1.389)	−12.188 (1.620)
	Age	0.112 (0.007)	0.132 (0.022)	0.103 (0.023)
	Gender (ref: male)	−2.260 (0.430)	−3.775 (0.813)	6.724 (0.878)
	Interaction	0.030 (0.004)	0.039 (0.012)	−0.070 (0.013)
Respiratory system	Intercept	−17.500 (1.562)	−16.889 (0.859)	−15.092 (0.665)
	Age	0.150 (0.021)	0.145 (0.008)	0.118 (0.008)
	Gender (ref: male)	0.055 (0.973)	−2.929 (1.111)	3.020 (1.500)
	Interaction	−0.008 (0.012)	0.025 (0.012)	−0.039 (0.017)
Digestive system	Intercept	−12.595 (1.222)	−13.238 (0.831)	−11.199 (0.758)
	Age	0.078 (0.019)	0.090 (0.016)	0.061 (0.013)
	Gender (ref: male)	−2.088 (0.692)	−3.331 (0.663)	4.635 (0.433)
	Interaction	0.026 (0.008)	0.036 (0.010)	−0.049 (0.007)
			α_2	α_3
			0.071 (0.048)	−0.615 (0.055)
Spatial dependence (ρ)	0.945 (0.023)			
State probability		0.463	0.493	0.044

Most of the age/gender interaction effects are significant across latent classes and causes of death. As expected, this indicates significant differences of age effects between males and females. The model correctly displays the plasticity of such interaction, by clustering the estimates according to three latent classes.

The model therefore summarizes the data by eight triplets of state-specific regression lines for each gender and cause of death. Figure 4 displays these regression lines, overlapped on the observations (crude mortality rates on the log scale). The mortality rates have been colored according to the maximum class membership probability $\hat{\mathbf{u}}_{is}$, defined in (11). We remark that these probabilities account for the spatial autocorrelation, which is strongly positive and significant (see the last row of Table 2). According to this classification criterion, most points are clustered within classes 1 and 2 (respectively, associated with black and red), while the third class clusters outlying mortality rates, occurring at the most advanced ages. The distribution of the clusters is associated with the estimated values of the parameters α_2 and α_3: while the parameter α_2, associated with latent class 2, is not significantly different of the reference latent class 1, the parameter α_3 is significantly less than the reference. The last row of Table 2 displays the estimated probabilities of each latent class.

The estimates of Table 2 and the class membership probabilities can be exploited to capture the average effect of age on cause-specific mortality, in the form of a smoothed regression line. More precisely, the state-specific regression lines can be weighted by the proportions of the observations that have been associated to each class, in order to obtain the average effect of age for each stratum s and each cause of death h, namely,

$$\hat{y}_{sh} = \sum_{p} \hat{\beta}_{ph} x_{sp},$$

where

$$\hat{\beta}_{ph} = \sum_{k=1}^{K} w_{hk} \beta_{phk},$$

whereas w_{hk} indicates the proportion of strata assigned to class k in Figure 4. The resulting average effect of age on (log-scaled) mortality is displayed by Figure 5. In keeping with what is already known in the epidemiological literature, we obtain curves that are essentially linear and depict a clear advantage of women over men, regardless of the cause of death. In addition, the proposed

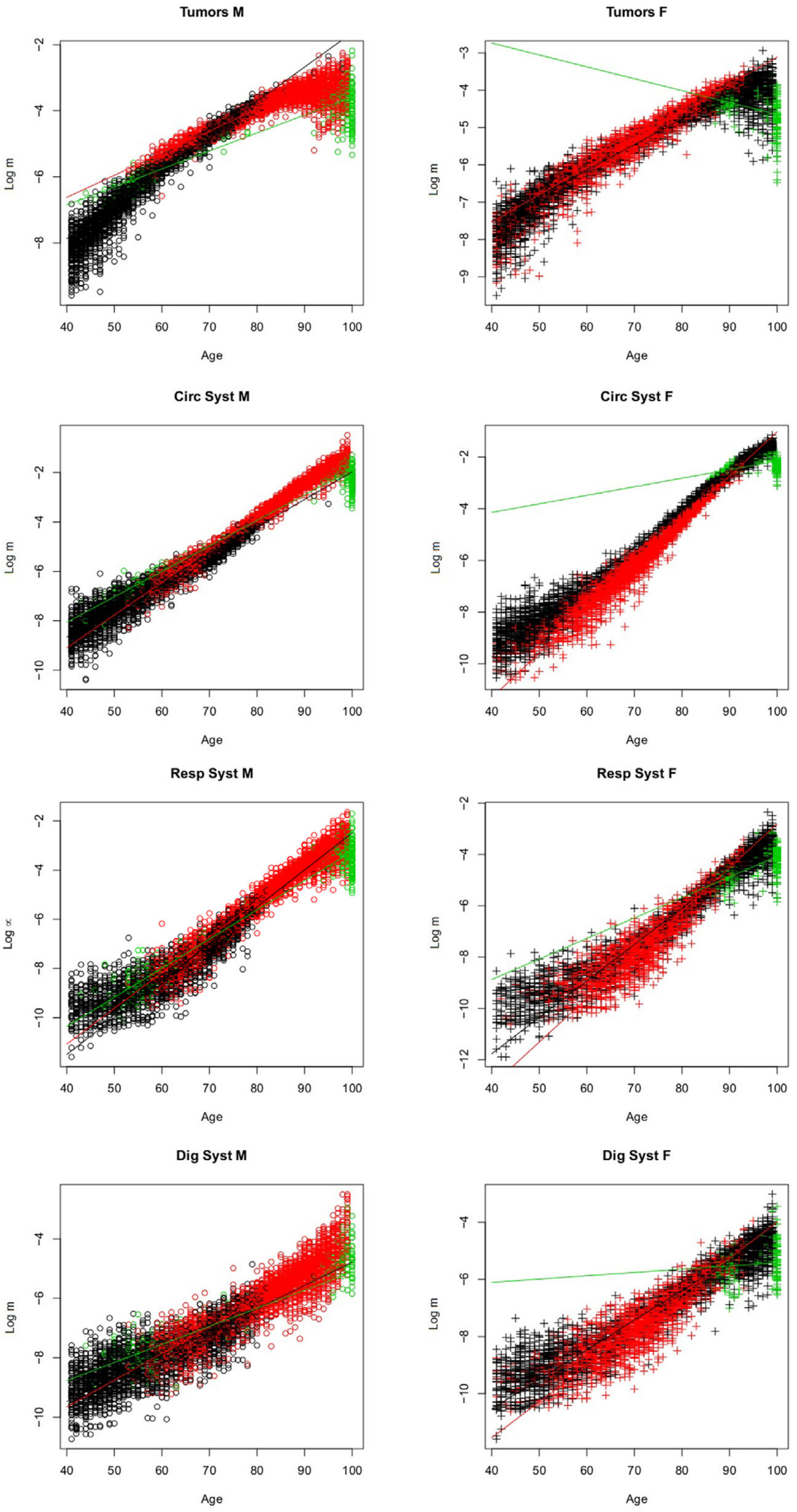


FIGURE 4 Scatterplots of mortality log-rates versus age, by gender and cause of death. *Note*: Points are colored according to the maximum a posterior estimation of class membership (black, red, and green, respectively, indicate classes 1–3). Lines indicate the class-specific regressions

model allows to interpret these curves as a combination of three regression curves, each associated with a specific latent class. That mortality curves should be always interpreted as mixtures of multiple trajectories because of latent heterogeneity is a well-known fact, since the seminal paper by Manton, Stallard, and Vaupel (1986). In our model, latent heterogeneity is parsimoniously represented by three latent classes that are associated with three different effects of age on mortality. Furthermore, all the curves of Figure 5 share a concave shape at advanced age, showing a decreasing rate of aging. This result is consistent with recent

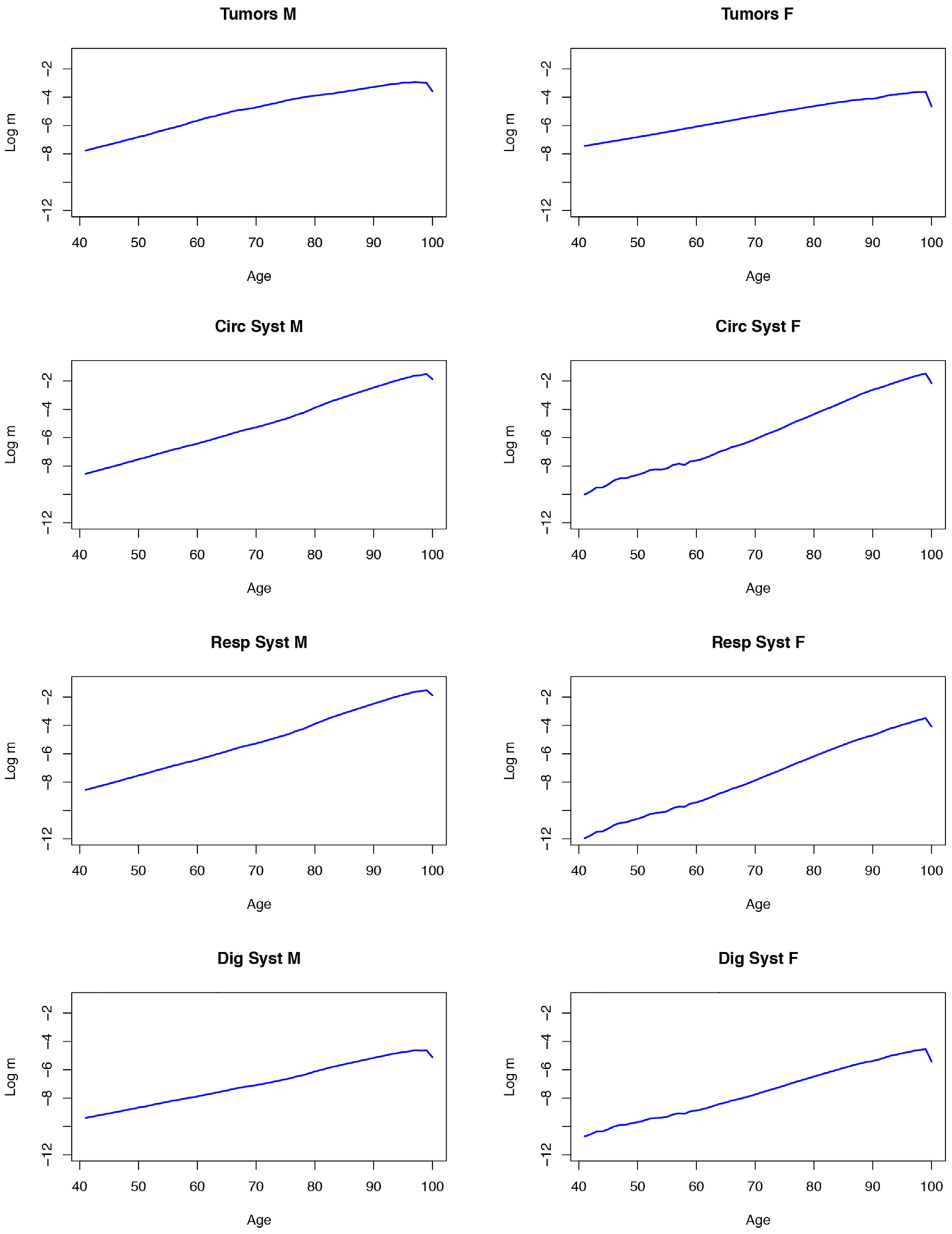


FIGURE 5 The average effect of age on mortality, as obtained by a hidden Markov field with three latent states

findings that demonstrate that the mortality curve reaches a plateau at the most advanced ages (Barbi, Lagona, Marsili, Vaupel, & Wachter, 2018). The age effect seems more important for those causes that are related with the senescence process, that is diseases of the cardiovascular and respiratory systems. For cancer, which is associated with risks not necessarily dependent on age, this effect is less pronounced.

By following the same method, the model allows to estimate the average effect of age, gender and their interaction across space.

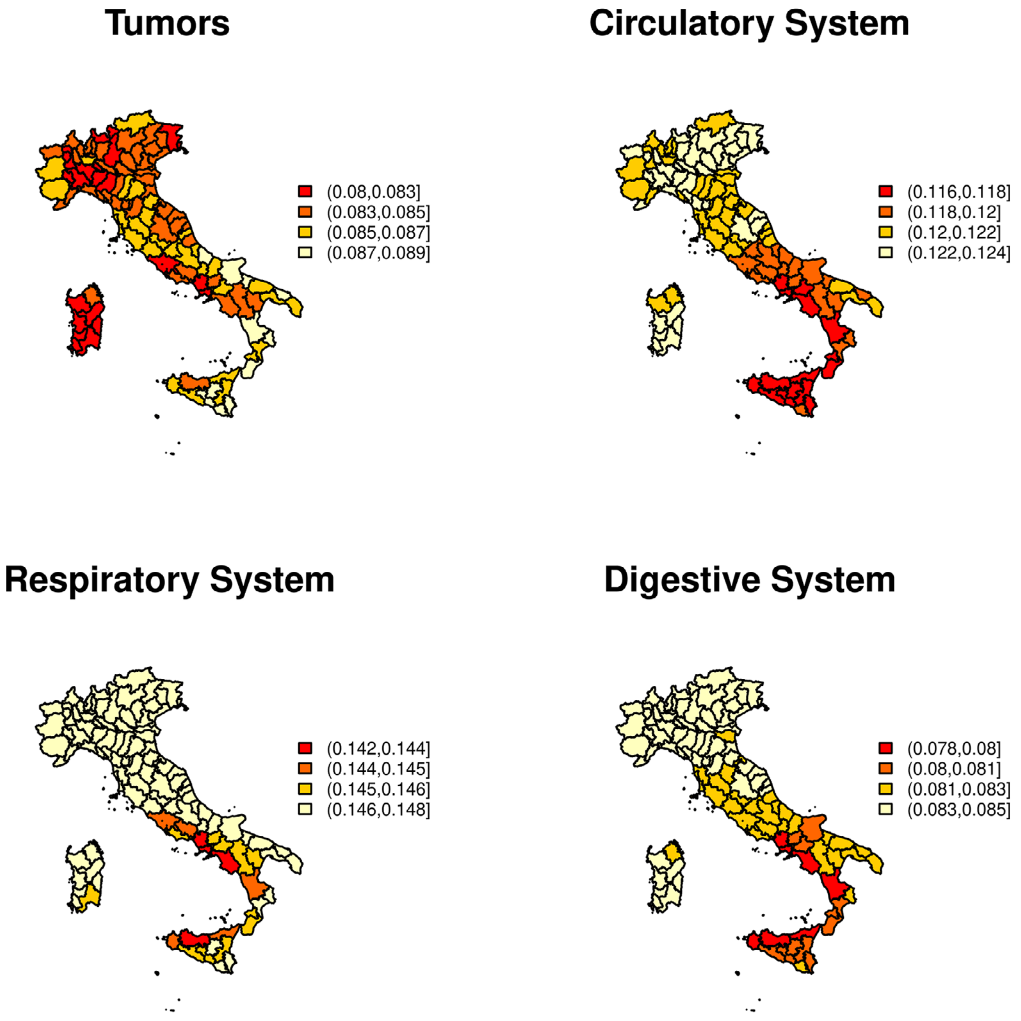


FIGURE 6 The spatial variation of the average age effect on cause-specific mortality, as estimated by a hidden Markov field with three latent states

Specifically, Figures 6–8 display the geographical variation of the available effects on cause-specific mortality across the Italian provinces. In order to interpret these results, we recall that our proposal was motivated by issues of mortality classification and, accordingly, we show the results that we obtained with the model with three latent classes that was picked up by the cCLC criterion. Should a higher degree of smoothness be desired, then a model with a larger number of states could be fitted.

Figure 6 shows the geographical patterns of the age effect for each cause of death. Remarkable differences can be observed between different causes of death. On one side, mortality due to tumors and cardiovascular diseases is characterized by a cluster of similar age effects in northern and central Italy. On the other side, the age effect on mortality due to respiratory and digestive disorders seems uniformly distributed across the whole peninsula.

Gender differences, instead (Figure 7), show similar spatial patterns, with three distinct clusters, respectively, associated with northern, central, and southern Italy. In our study, the reference level of gender is male, therefore gender effect indicates the effect of being male. We observe that the average effect of gender is always negative in the case of mortality due to circulatory, respiratory and digestive diseases, indicating an expected female advantage. However, the model spots specific provinces where gender differences are limited. In the case of cancer mortality, instead, the average gender effect is positive. This result should be interpreted by looking at the top-left panel of Figure 8, which displays the average interaction age/gender effect. The negative values taken by these estimates indicate that the rate of aging among women is less than the rate of aging of men. In summary, our findings indicate that while the risk of cancer mortality among men is less than the mortality risk among women at young ages, this inequality is reversed at older ages. For the other causes of death, instead, the average interaction effect takes positive values. As a result, women always experience lower risks of mortality than men, but gender differences decrease with age in most Italian provinces.

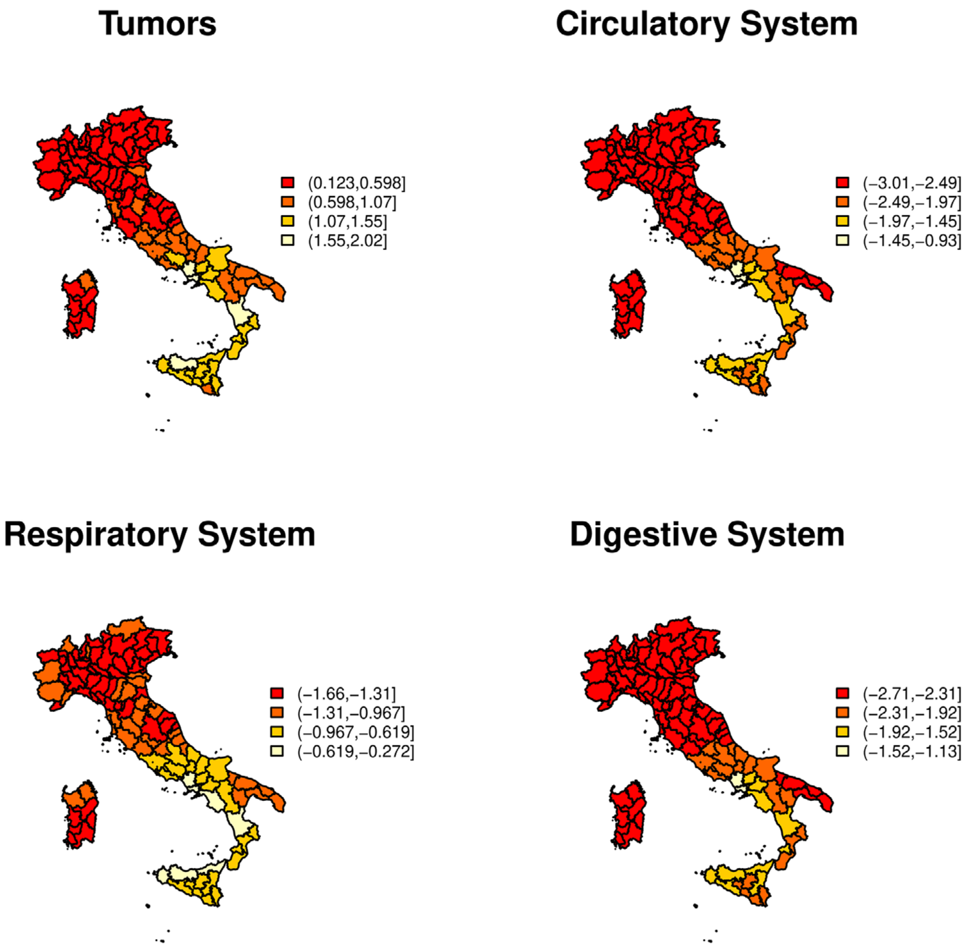


FIGURE 7 The spatial variation of gender differences (reference is male) on cause-specific mortality, as estimated by a hidden Markov field with three latent classes

6 | DISCUSSION

Generalized linear models can be extended in a number of ways. A popular extension is provided by generalized additive models, which are obtained by replacing some or all of the linear and parametric functions of the regressors by smooth nonparametric functions. Another, perhaps less developed, approach toward the flexibility of generalized linear models is provided by a model where the regressors are linear but the regression coefficients are allowed to vary. This leads to the class of the varying-coefficients models. Such an extension was introduced by Hastie and Tibshirani (1993), who assumed a regression model whose coefficients are functions of a specific covariate, known as an “effect-modifier.” Our proposal can be viewed as a varying-coefficients model where the effect modifier is a latent factor, representing unobserved conditions under which the data are observed. A specific feature of our proposal is that the latent factors are driven by a spatial multinomial process, to account for the spatial autocorrelation of the unobserved conditions.

The level of autocorrelation is parsimoniously captured by a single parameter ρ. When $\rho = 0$, our proposal reduces to a finite mixture of regressions. The use of a single autocorrelation parameter could be in principle seen as an unnecessary limitation. In principle, multiple autocorrelation parameters would enhance model flexibility, by capturing local autocorrelation levels that vary across space (Tjelmeland & Besag, 1998). From a practical viewpoint, however, multiple autocorrelation levels can be easily distinguished only when the number of spatial units is extremely large. In our study of Italian mortality, the sample size (110 provinces) does not seem adequately large to allow for more than one autocorrelation parameter.

Under the proposed model, death counts are conditionally Poisson-distributed, given spatially correlated random effects driven by a Potts model. Random effects should capture unobserved heterogeneity and simultaneously account for overdispersion. However, the data could be affected by supplementary sources of overdispersion that may require further modeling of the

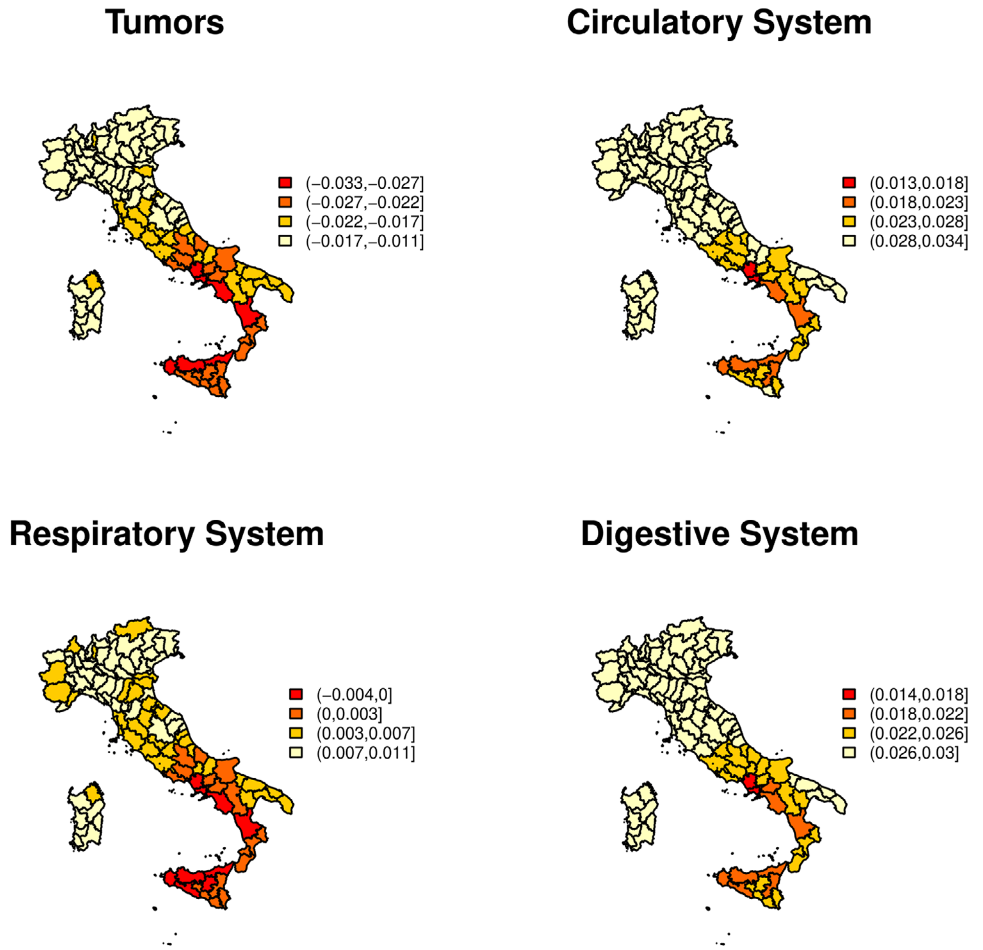


FIGURE 8 The spatial variation of the age/gender interaction effect on cause-specific mortality, as estimated by a hidden Markov field model with three latent classes

variance function. In this paper, we focused on flexible regression functions with space-varying coefficients. Our proposal could, however, be extended to allow for flexible variance functions that vary across space.

A limit of the model is the intractable likelihood function, which complicates parameter estimation. This issue was addressed by taking a CL approach, which is based on the definition of a class of subsets that covers the study area. The numerical tractability of this approach depends on the size of the largest covering subset. This suggested the use of pairs of neighboring sites as covering subsets. This method, however, depends on a spatial neighborhood structure, namely, the nearest neighborhood structure, which is assumed a priori on the spatial lattice and which is not necessarily the best choice for defining the CL function.

Our proposal was conceptualized to segment mortality data at a given point in time, and, as such, it does not include the dynamics of the mortality process. Hidden Markov models for the analysis of cause-specific mortality data over time have been recently proposed by Lagona and Barbi (2019) and our proposal can be seen as the spatial extension of these models. Integrating the two approaches to model space–time multivariate mortality data is a quite natural, although not obvious, idea. A possible approach in this direction, for example, is the specification of a sequence of conditionally independent hidden Markov fields, whose parameters evolve according to the states of a latent Markov chain. Further research is, however, needed to explore ways to integrate CL methods with the forward–backward estimation methods that are routinely exploited in models that involve latent Markov chains.

ACKNOWLEDGMENTS

This work is developed under the PRIN2015 supported-project “Environmental processes and human activities: capturing their interactions via statistical methods (EPHASTAT)” funded by MIUR (Italian Ministry of Education, University and Scientific Research).

CONFLICT OF INTEREST

The authors have declared no conflict of interest.

OPEN RESEARCH BADGES

This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available in the Supporting Information section.

This article has earned an open data badge "**Reproducible Research**" for making publicly available the code necessary to reproduce the reported results. The results in this paper could not be fully reproduced due to insufficient transparency of the submitted code and their computational cost.

ORCID

Monia Ranalli https://orcid.org/0000-0001-7193-8803

REFERENCES

Alfò, M., Nieddu, L., & Vicari, D. (2009). Finite mixture models for mapping spatially dependent disease counts. *Biometrical Journal*, *51*(1), 84–97.

Ameijeiras-Alonso, J., Lagona, F., Ranalli, M., & Crujeiras, R. M. (2019). A circular nonhomogeneous hidden Markov field for the spatial segmentation of wildfire occurrences. *Environmetrics*, *30*(2), e2501.

Armitage, P. (1966). The chi-square test for heterogeneity of proportions, after adjustment for stratification. *Journal of the Royal Statistical Society B*, *28*, 150–163.

Assunção, R. M. (2003). Space varying coefficient models for small area data. *Environmetrics*, *14*(5), 453–473.

Barbi, E., Casacchia, O., & Racioppi, F. (2018). Cause-specific mortality as a sentinel indicator of current socioeconomic conditions in Italy. *Demographic Research*, *39*(21), 635–646.

Barbi, E., Lagona, F., Marsili, M., Vaupel, J. W., & Wachter, K. W. (2018). The plateau of human mortality: Demography of longevity pioneers. *Science*, *360*(6396), 1459–1461.

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

Biernacki, C., & Govaert, G. (1997). Using the classification likelihood to choose the number of clusters. *Computing Science and Statistics*, *29*, 451–457.

Brillinger, D. R. (1986). The natural variability of vital rates and associated statistics. *Biometrics*, *42*, 693–734.

Gaetan, C., & Guyon, X. (2010). *Spatial statistics and modeling*. Berlin: Springer.

Gelfand, A. E., & Vounatsou, P. (2003). Proper multivariate conditional autoregressive models for spatial data analysis. *Biostatistics*, *4*(1), 11–15.

Godambe, V. P. (1960). An optimum property of regular maximum likelihood estimation. *Annals of Mathematical Statistics*, *31*(4), 1208–1211.

Green, P. J., & Richardson, S. (2002). Hidden Markov models and disease mapping. *Journal of the American Statistical Association*, *97*(460), 1055–1070.

Hastie, T., & Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society. Series B (Methodological)*, *55*(4), 757–796.

Klauenberg, K., & Lagona, F. (2007). Hidden Markov random field models for TCA image analysis. *Computational Statistics and Data Analysis*, *52*(2), 855–868.

Lagona, F., & Barbi, E. (2019). Segmentation of mortality surfaces by hidden markov models. *Statistical Modelling*, *19*(3), 276–298.

Lagona, F., Jdanov, D., & Shkolnikova, M. (2014). Latent time-varying factors in longitudinal analysis: A linear mixed hidden Markov model for heart rates. *Statistics in Medicine*, *33*(23), 4116–4134.

Lagona, F., & Picone, M. (2016). Model-based segmentation of spatial cylindrical data. *Journal of Statistical Computation and Simulation*, *86*(13), 2598–2610.

Lindsay, B. (1988). Composite likelihood methods. *Contemporary Mathematics*, *80*, 221–239.

Manton, K., Stallard, E., & Vaupel, J. W. (1986). Alternative models for the heterogeneity of mortality risks among the aged. *Journal of the American Statistical Association*, *81*(395), 635–644.

Ranalli, M., Lagona, F., Picone, M., & Zambianchi, E. (2018). Segmentation of sea current fields by cylindrical hidden markov models: A composite likelihood approach. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *67*(3), 575–598.

Ranalli, M., & Rocci, R. (2016). Standard and novel model selection criteria in the pairwise likelihood estimation of a mixture model for ordinal data. In *Analysis of large and complex data* (pp. 53–68). Cham: Springer.

Strauss, D. (1977). Clustering on coloured lattices. *Journal of Applied Probability*, *14*, 135–143.

Tjelmeland, H., & Besag, J. (1998). Markov random fields with higher-order interactions. *Scandinavian Journal of Statistics*, *25*(3), 415–433.

Varin, C., Reid, N., & Firth, D., (2011). An overview of composite likelihood methods. *Statistica Sinica*, *21*(1), 1–41.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Lagona F, Ranalli M, Barbi E. A model with space-varying regression coefficients for clustering multivariate spatial count data. *Biometrical Journal*. 2020;1–17. https://doi.org/10.1002/bimj.201900229