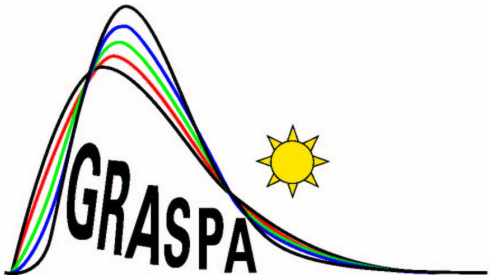
ISSN: 2037-7738
GRASPA Working Papers
[online]

Web Working Papers
by
The Italian Group of Environmental Statistics



Gruppo di Ricerca per le Applicazioni della Statistica
ai Problemi Ambientali

www.graspa.org

Spatial Variation of Multiple Diseases in Relation to an Environmental Risk Source

Crescenza Calculli, Alessio Pollice, Lucia Bisceglia

GRASPA Working paper n.38, June 2010

Spatial Variation of Multiple Diseases in Relation to an Environmental Risk Source

Crescenza Calculli[*1], **Alessio Pollice**[1], and **Lucia Bisceglia**[2]

[1] Department of Statistics C. Cecchi, University of Bari, Via C. Rosalba 53, 70124, Bari, Italy
[2] ARPA Puglia, Corso Trieste 27, 70126, Bari, Italy

Summary

The analysis of the spatial variation of disease risk is crucial in Environmental Epidemiology studies. In this context, the effects of the presence of a source of pollution on the population health can be evaluated by models that consider distance from the source as a possible risk factor. We introduce an hierarchical Bayesian model in order to investigate the association between the risk of multiple pathologies and the presence of the risk source in the context of spatial case-control studies. Our approach extends some previous works based on spatial point patterns, concerning the risk variation of a single pathology and provides the possibility to incorporate spatial effects and other confounding factors within a logistic regression model. Moreover, spatial effects are decomposed into the sum of a disease-specific parametric component accounting for the distance from the point source and a common semi-parametric component that can be interpreted as a residual spatial variation. The proposed model is estimated by MCMC and is applied to data from a spatial case-control study in order to evaluate the association of the incidence of different cancers typologies with the residential location in the neighbourhood of a petrochemical plant in the Brindisi area (South-eastern Italy).

Key words: Multiple Diseases, Environmental Risk Source, Thin-plate Splines

1 Introduction

In the last few years the application of methods based on spatial point processes has caught on the epidemiological field more and more, specifically with respect to spatial distribution of diseases in relation with suspect sources of environmental risk. Whereas the large availability of data concerning counts of disease cases within territorial partitions, such as districts, towns, regions, or other small areas, allowed a huge enrichment of the literature about modeling and graphic representation of epidemiological data at the area level (disease mapping), point data mapping methods represent an investigation field still open and in development. Though in the last two decades, methods based on point processes allowed an improvement in studies of the spatial variation of risk in the presence of environmental sources of pollution. Yet we observed a lack of analysis performed referring to more diseases with respect to one or more point sources of pollution (Diggle *et al.*, 1997; Lawson, 1993). The simultaneous treatment of data referred to several pathologies seems to be hardly touched by the point data literature, while it has been object of a deeper discourse in the disease mapping field (Knorr-Held and Best, 2001; Held *et al.*, 2005). In this paper the possibility of analyzing the simultaneous spatial behavior of several pathologies with respect to an environmental source of pollution has been pursued through a modeling proposal inspired by several works more generally concerned with other areas of spatial statistical analysis.

As a starting point we consider a logistic regression model with a GAM-type predictor including the spatial effect of the distance from the source and the non-spatial effects due to other exposure-related risk factors (Diggle *et al.*, 2000). Diggle and Zheng (2005) suggest to extend this type of predictor in order to incorporate a residual spatial component independent on the source, in the form of a Gaussian random

* Corresponding author: e-mail: e.calculli@dss.uniba.it, Phone: +39 080 5049243

field (GFR). Such a residual component could be probably due to the socio-demographic, geo-morphologic and economic structure of the area under consideration. Also Hughes-Oliver *et al.* (2008) decompose the random part of a second-order non stationary spatial model into the sum of an AutoRegressive Point Source model (ARPS) and a residual spatial conditional autoregressive model. As an alternative approach Crainiceanu *et al.* (2008) suggest a bivariate extension of a binomial geo-statistical model to map the spatial variation of disease prevalence. The authors model the spatially varying log-odds as a linear predictor including the effects of covariates and a zero-mean spatial stationary process that represents the residual spatial variation. The latter, is represented through a thin-plate splines interpolator that is an efficient alternative to the use of methods based on Gaussian stationary processes.

This paper is structured in four sections. The first shows the characteristics of the case study and types of data used in the subsequent application; the second section is focused on the extension of the models for the evaluation of the relative risk in the presence of a environmental source of pollution, to the simultaneous analysis of the risk of multiple pathologies. The third section concerns instead the application of the proposed model to the case-study. Finally the fourth section is dedicated to the presentation and to the discussion of the main results.

2 Data

The study area is characterized by high environmental risk due to the presence of many industrial sites. We focus on a petrochemical plant in the city of Brindisi and considered the city and three neighbouring municipalities (Carovigno, San Pietro Vernotico and Torchiarolo) to define the area at risk. Cases are 403 subjects resident in the study area in 1999-2001 with histologically confirmed lung cancer, pleura neoplasm, bladder cancer and lymphohematopoietic malignancies, retrieved from the Apulia Cancer Registry. Controls are 1694 subjects resident in the same area in 1999-2001, randomly selected and matched to the cases by sex, date of birth and residential municipality. The residential location of cases was defined as the address at the diagnosis, while addresses in the regional Sanitary Registry were used for controls. All the subjects were geographically positioned according to GPS standards (Fig. 1). Addresses were checked directly and the distance between each subject residential location and the petrochemical plant was calculated.

3 Model Formulation

In the case of a single pathology assume that cases and controls are realizations of two independent inhomogeneous Poisson processes, with intensity functions $\lambda_1(s)$ and $\lambda_2(s)$ respectively. Given a generic location s, a function $f(s)$ can be defined so that the two intensities result in being bound by the relation:

$$\lambda_1(s) = \widetilde{\mu}\lambda_2(s)f(s) \tag{1}$$

Barring the constant $\widetilde{\mu}$, representing the ratio of the number of cases to controls, $f(s)$ is obtained as the ratio of the two intensity functions and can be interpreted as the relative risk surface, with s varying over the whole area (Kelsall and Diggle, 1995). Conditioning with respect to the sample size and the spatial location of subjects, the case/control label of an individual living at location s, is characterized by a Bernoulli distribution: $Y(s) \sim \text{Bernoulli}\{\pi(s)\}$, where $\pi(s)$ represents the conditional probability of observing a case at localization s. Given (1), $\pi(s)$ can be expressed as

$$\pi(s) = \frac{\lambda_1(s)}{\lambda_1(s) + \lambda_2(s)} = \frac{\widetilde{\mu}f(s)}{1 + \widetilde{\mu}f(s)} \tag{2}$$

The estimation of the disease odds and of the relative risk $f(s)$ can thus be achieved through logistic regression modeling, coding cases and controls as 1 and 0. This approach enables to consider a predictor

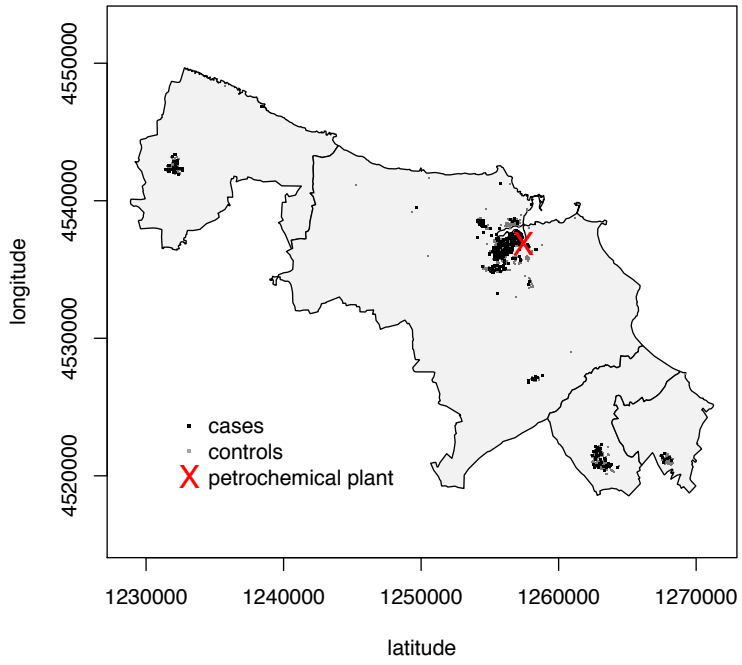


Figure 1 Case and control locations in relation to the petrochemical plant location in the Brindisi risk area

including a set of terms that measure the effects of covariates and spatial variation. Suppose that we have a putative point source of pollution at location s_0 and that we are interested to know how the risk surface varies in relation to this source. This can be achieved by an explicit consideration of the location of the source in the specification of the relative risk function $f(s) = f(d(s, s_0), \boldsymbol{\theta}_k)$, describing its variation according to the distance $d(s, s_0)$ from s_0 and to the parameter vector $\boldsymbol{\theta}_k$.

In order to consider an extension of the model to the risk of a set of K pathologies, let $Y_k(s)$ be the case/control label of the k-th pathology for an individual living at location s ($k = 1, \ldots, K$). As before $Y_k(s)$ is conditionally distributed as a Bernoulli with parameter $\pi_k(s)$ representing the probability to observe a case of the k-th pathology at location s. The locations of cases and controls for the k-th pathology are $s_1 \ldots s_{n_k}$, where $N = \sum_{k=1}^{k} n_k$ is the total number of locations of cases and controls. A mixed-effects linear predictor is specified as follows:

$$\text{logit}\{\pi_k(s)\} = \mu_k + \mathbf{X}(s)\boldsymbol{\beta}_k + f\{d(s, s_0); \boldsymbol{\theta}_k\} + S(s) + \varepsilon(s) \tag{3}$$

The mixed-effects linear predictor in (3) is characterized by a fixed trend component and a random (similar specifications for different data contexts are contained in Crainiceanu *et al.* (2008), Hughes-Oliver *et al.* (2008)). Fixed effects include the average effects of each disease μ_k and the product of the s-th row of matrix $\mathbf{X}$ containing p individual covariates corresponding to non-spatial risk factors and a p-dimensional vector of disease-specific coefficients $\boldsymbol{\beta}_k$. Also, the function $f\{d(s, s_0); \boldsymbol{\theta}_k\}$ in (3) represents the disease-specific fixed effects of the distance from the source ruled by parameter vector $\boldsymbol{\theta}_k$. The model specification is indeed very much affected by the choice of the function that describes the changes in the intensity due

to the distance from the source s_0. This function can be specified according to different parametric forms, considering effects caused by directional components, by the distance from the source or eventual peak effects. Therefore the elements of the parameter vector $\boldsymbol{\theta}_k$ depend on the chosen functional form.

The random effect $S(\cdot)$ represents a residual spatial component of the model, common to all pathologies and not depending on the presence of the source (probably due to the socio-demographic, geo-morphologic and economic structure of the area under consideration). The standard approach would be to model $S(\cdot)$ as a stationary Gaussian random field, as proposed by Diggle *et al.* (1998) in a different context. In the work of Hughes-Oliver *et al.* (2008), concerning data on a regular lattice, $S(\cdot)$ is modeled through an ARPS process, that is to say an auto-regressive process whose specification depends on the proximity structure between points. In the case of non-regular spatial data, the application of such a model, other than being less convenient, brings also a considerable computational complication because of the high number of parameters involved in the evaluation. As an alternative the Gaussian process $S(\cdot)$ may be conveniently substituted by a semi-parametric structure based on a thin-plate splines interpolator which can be given a linear low-rank approximation and offers substantial computational advantages (Crainiceanu *et al.*, 2008). Given a set of T spatial nodes representative of the N spatial locations, the low-rank representation of $S(s)$ takes the following linear form:

$$S(s) = \mathbf{Z}(s)\mathbf{b} \tag{4}$$

where $\mathbf{b}$ is a T-dimensional vector of random coefficients that control the total amount of spatial smoothing and $\mathbf{Z}(s)$ is the s-th row of the design matrix:

$$\mathbf{Z} = \mathbf{Z}_T \boldsymbol{\Omega}_T^{-1/2} \tag{5}$$

where matrix $\mathbf{Z}_T$ and $\boldsymbol{\Omega}_T$ respectively represent the correlation matrix between points and nodes and that among nodes. Both $\mathbf{Z}_T$ and $\boldsymbol{\Omega}_T$ can be built by an isotropic spatial correlation function $C(\cdot)$ so that $\mathbf{Z}_T = [C\{d(s,t)\}]$ and $\boldsymbol{\Omega}_T = [C\{d(t,t')\}]$ where $t, t' = 1, \ldots, T$.

In fact, in the presence of a high number of sampling locations, the estimates don't depend anymore on them, but rather are evaluated by a reduced number of nodes specified for the smooth function.

Finally in (3) another non-structured error term $\varepsilon(s)$ is considered, in addition to the casual component of the model.

3.1 Bayesian Specification

The logistic regression model with the predictor in (3) and (4) is estimated in the hierarchical Bayesian framework specifying the prior distributions of fixed and random effects. Gaussian priors are assumed for fixed covariate and random spatial effects:

$$\mu_k \sim \mathrm{N}(0, \tau_\mu), \quad \boldsymbol{\beta}_k \sim \mathrm{N}_p(\mathbf{0}_p, \tau_\beta \mathbf{I}_p), \quad \mathbf{b} \sim \mathrm{N}_T(\mathbf{0}_T, \tau_b \mathbf{I}_T), \quad \varepsilon \sim \mathrm{N}(0, \tau_\varepsilon) \tag{6}$$

where $\mathbf{0}_m$ and $\mathbf{I}_m$ stand for the m-dimensional null vector and identity matrix respectively. The parameter vector $\boldsymbol{\theta}_k$ depends on the choice of the functional form of the distance function. In this paper we refer to the "distance-decline" isotropic semi-Gaussian model: $f\{d(s,s_0), \boldsymbol{\theta}_k\} = 1 + \alpha_k \exp\{-\phi_k d(s,s_0)^2\}$ (Diggle and Rowlingson, 1994). For the parameter vector $\boldsymbol{\theta}_k = (\alpha_k; \phi_k)$ we propose the following prior structure:

$$\alpha_k \sim \mathrm{Gamma}(a, b), \quad \phi_k \sim \mathrm{U}(0, \phi_{max}) \tag{7}$$

The choice of these two priors is bound to the meaning of the corresponding parameters in the "distance-decline model": α_k represents the excess of relative risk at the source location for the k-th disease and

ϕ_k can be interpreted as the decrease rate of the risk of the k-th disease as a function of the distance from the source. The choice of the Gamma distribution α_k is widely used in the literature (Dreassi *et al.*, 2008; Wakefield and Morris, 2001) and allows incorporating epidemiological knowledge of the case-study specifying appropriate values of the hyperparameters a and b. Also the choice of the hyperparameter value ϕ_{max} in the Uniform prior for ϕ_k reflects the distance within which the effect of the source is assumed to disappear. Diggle *et al.* (2000) suggest to set $\phi_{max} = 0.2 \times d_{max}$ or $\phi_{max} = 0.5 \times d_{max}$.

Finally a non-informative approach is adopted in the specification of the Gamma priors for precision hyperparameters τ_μ, τ_β, τ_b and τ_ε in (6). This is still a discussed issue as this choice corresponds to very small values of Gamma shape and scale parameters and leads to prior distributions concentrated at zero and with heavy tails (Gustafson *et al.*, 2006). Crainiceanu *et al.* (2008) suggest that generally the choice of Gamma$(.001, .001)$ does not affect posterior estimates, though this choice causes slow convergence and poor mixing in MCMC estimation algorithms, when information conveyed by the data on precision hyperparameters is insufficient. In the proposed model this is particularly true for the spatial effect precision τ_b.

4 The Brindisi Case-study and Some Results

The model in (3)-(4) was applied to the data of the Brindisi case-control study. The 2097 case and control subjects were arranged in five different cancers' classes as reported in the Table 1:

Table 1 Cases and controls diseases classification

	Cases	Controls
Lung	169	688
Bladder	145	616
Non-Hodgkin limphoma	49	217
Leukaemias	19	81
Others	21	92

For $k = 1, \ldots, 5$, the linear mixed predictor in our model includes a spatially varying mean $\mu_k + f\{d(s, s_0); \boldsymbol{\theta}_k\}$ and a residual spatial variation, $S(s)$. The disease-specific effects of covariates, $\mathbf{X}(s)\boldsymbol{\beta}_k$, can be ignored in this case-control study that concerns cases and controls matching by sex, age and residential location.

The prior distribution for the model parameters is as follows. The intercept μ_k representing the overall mean has a non informative flat distribution. The function $f\{d(s, s_0); \boldsymbol{\theta}_k\}$ has a "distance-decline" isotropic semi-Gaussian specification and the choice of priors for the vector parameter $\boldsymbol{\theta}_k = (\alpha_k; \phi_k)$ is a delicate issue as discussed in section 3.1. We suggest for the excess of relative risk at the source, $\alpha_k \sim$ Gamma$(0.5, 0.5)$ and for the decrease rate of risk $\phi_k \sim$ U$(0, 13)$. For the hyperparameters in (6) a non informative approach was used and specified as $\tau_\mu, \tau_\beta, \tau_b, \tau_\epsilon \sim$ Gamma$(.001, .001)$.

The Bayesian estimate of the model above was implemented using MCMC simulation through the use of the software `WinBUGS` (Spiegelhalter *et al.*, 2004). Posterior distributions of unknown parameters were obtained using 110,000 iterations discarding the first 10,000 corresponding to a burn-in phase of the algorithm. Initial values were obtained as the results of an earlier simulation run of two chains with overdispersed starting values. The convergence of chains has been evaluated through graphical inspection of the trace plots and convergence statistics provided by the software. The high number of parameters involved and the complexity of the model cause some critical factors to emerge in the simulations, such as poor mixing (slow convergence) of the chains and long updating times (the average time for a single update is about 1,2 seconds). The fixed parameters show the best performances in terms of convergence. Especially α_k and ϕ_k namely the risk at source and the risk decrease rate getting far from the source itself,

Table 2 Summaries of posterior densities for several parameters of interest

	Mean	SD	MC error
$\mu_{(lung)}$	0.0724	0.4845	0.0187
$\mu_{(bladder)}$	0.0839	0.4910	0.0189
$\mu_{(limphNH)}$	0.0867	0.5027	0.0194
$\mu_{(leuk)}$	0.0816	0.4962	0.0189
$\mu_{(others)}$	0.0844	0.5032	0.0192
$\alpha_{(lung)}$	0.9206	1.3000	0.0069
$\alpha_{(bladder)}$	0.8954	1.295	0.0065
$\alpha_{(limphNH)}$	0.9480	1.355	0.0069
$\alpha_{(leuk)}$	0.9704	1.3740	0.0067
$\alpha_{(others)}$	0.9612	1.3580	0.0069
$\phi_{(lung)}$	6.5540	3.7050	0.0122
$\phi_{(bladder)}$	6.8090	3.6670	0.0129
$\phi_{(limphNH)}$	6.6590	3.6920	0.0101
$\phi_{(leuk)}$	6.5760	3.7110	0.0095
$\phi_{(others)}$	6.5570	3.7230	0.0096
τ_b	0.0126	0.0202	8.06E-4
τ_ϵ	0.2373	0.8877	0.0374

reach convergence after a small number of iterations. Problems of convergence instead occur for the spatial random effects **b** and their precision τ_b for which an elevated number of iterations is needed to converge. This kind of problems also occurs in other spatial models based on Markov Random Fields or Gaussian precesses where precision parameters are known to threaten the convergence of iterative estimation algorithms (Knorr-Held and Rue, 2002).

The main posterior statistics concerning the more relevant parameters are reported in Table 2. We can notice the absence of relevant differences of the posterior means of distance function parameters α_k and ϕ_k among the five pathologies. The likelihood function of this kind of models is known to be quite flat and to produce estimates characterized by elevated standard errors for any choice of the distance function Dreassi *et al.* (2008). As a consequence, the choice of the prior distribution highly affects the distance function parameter estimates. With the data at hand, distance function parameter estimates are very close to the average values specified by their prior distributions and are characterized by high variability. The distance functions estimated for the five pathologies are reported in Fig. 2, showing similar values of the risk at the source for the five diseases and close to 2. For all pathologies the effect of the risk source seems to vanish for distances higher than 1 km. The estimated spatial effect common to all neoplasias $S(s)$ is reported in Fig. 3. Here the contours interpolate the predictions of the random effect **b** corresponding to $T = 50$ spatial nodes obtained by the `clara` space-filling algorithm in the `R` package `SemiPar` (Ganguli and Wand, 2005). The level curves show an increase of the common risk towards North-West of the study area. As we expected, this spatial effect largely reproduces the distribution of the population at risk, due its substantial overlapping with the location of cases. On the whole, the mixed effects hierarchical model

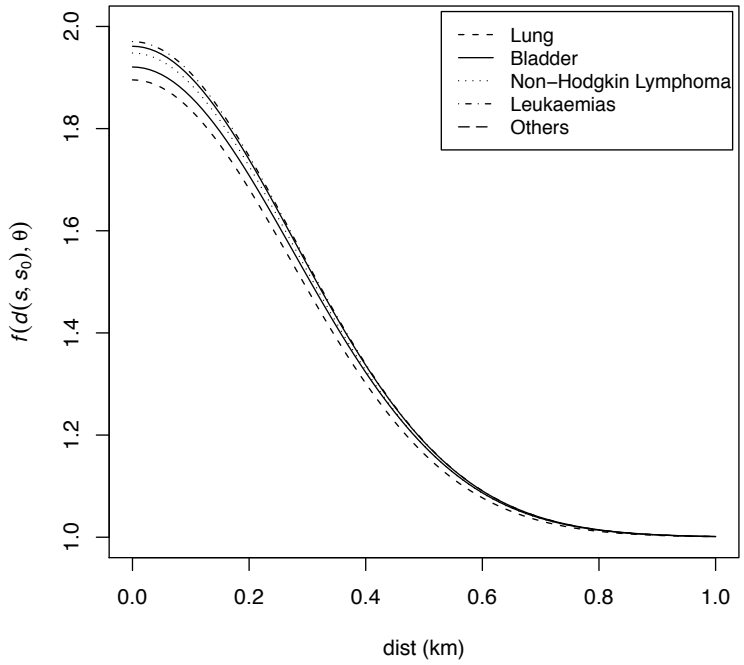


Figure 2 Curves of $f(d(s, s_0); \boldsymbol{\theta}_k)$ for different pathologies resulting from estimates of α_k and ϕ_k

applied to the data from the case-control study succeeds in catching the effects of the pollution source and the residual spatial component at the same time though it fails in discriminating between considered pathologies. In the absence of substantial epidemiological information about data, or in a non-informative Bayesian framework, the proposed methodology allows to estimate effects due to the presence of an environmental pollution source on the risk of various diseases, properly taking a common residual spatial pattern into account.

4.1 Concluding Remarks

The paper focuses on the application of a point process model for the analysis of the spatial distribution of the incidence of cases of five pathologies in the presence of an environmental risk source. We propose a hierarchical Bayesian GAM-type model that includes the possibility to incorporate disease-specific source-related spatial effects and a residual spatial effect common to all pathologies. The spatial component related to the specific effects of the different diseases is specified by a function that describes the changes of intensity due to the distance from the source, while the common residual spatial component approximates a Gaussian random field by a low-rank representation of a thin-plate spline interpolator. Such a representation constitutes an efficient alternative to the use of parametric covariance structures, common in geostatistic, above all in terms of computational advantages. The Bayesian approach and the complex hierarchical structure show a few limitations particularly bound to priors specification. Yet, the possibility to properly deal with all those features with a non informative approach, makes the model suitable to be used for the analysis of complex real cases such as the one concerning the case-control study. The Brindisi area, at risk of environmental crisis, is characterized by a strong anthropical pressure and by the presence

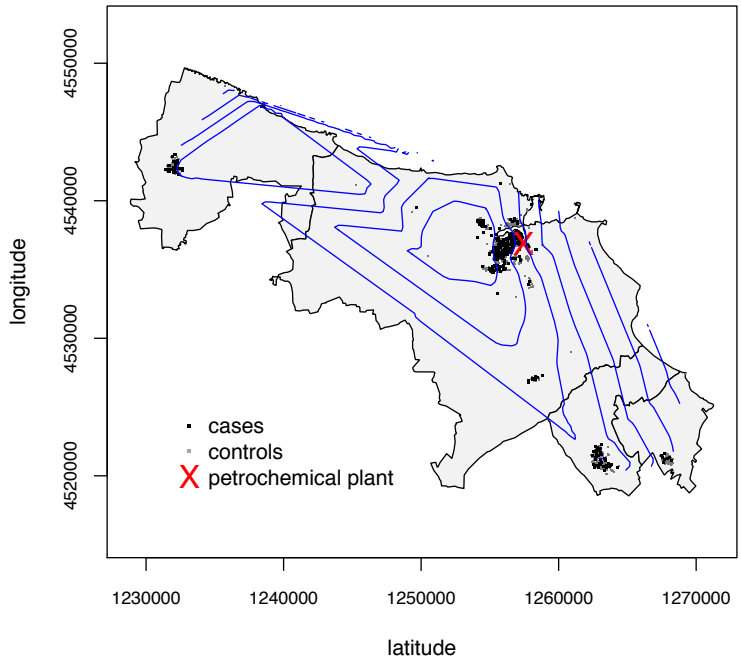


Figure 3 Map of the low-rank predicted GRF, $S(s)$

of elevated risk industrial facilities and has been object of several epidemiological studies which indeed pointed out an excess of risk in relation with some cancer pathologies and lung diseases. The case-control study presented in this paper has the aim to evaluate the eventual association of some cancer pathologies with the distance of the residence to a petrochemical plant. With respect to this objective, the application of the proposed model allowed to point out an excess of risk related to all the analyzed pathologies. Another relevant aspect concerns common spatial effect not depending on the source location of the risk source. The estimated model allowed to point out an excess of risk towards north-west with respect to whole risk area. The proposed model offers the advantage of efficiently considering a set of factors that affects the estimate of risk of multiple diseases related to the presence of an environmental pollution source, integrating epidemiological and modeling aspects.

References

Crainiceanu, C. M., Diggle, P. J., and Rowlingson, B. (2008). Bivariate binomial spatial modeling of Loa loa prevalence in tropical Africa (with discussion). *Journal of the American Statistical Association*, **103**, 21–47.

Diggle, P. and Rowlingson, B. (1994). A conditional approach to point process modelling of elevated risk. *Journal of the Royal Statistical Society, Series A*, **153**, 349–362.

Diggle, P., Tawn, J., and Moyeed, R. (1998). Model-based geostatistics (with discussion). *Applied Statistics*, **47**, 299–350.

Diggle, P. J. and Zheng, P. (2005). Nonparametric estimation of spatial segregation in a multivariate point process: bovine tubercolosis in Cornwall, UK. *Applied Statistics*, **54**, 645–658.

Diggle, P. J., Morris, S., Elliott, P., and Shaddick, G. (1997). Regression modelling of disease risk in relation to point sources. *Journal of the Royal Statistical Society, Series A*, **160**, 491–505.

Diggle, P. J., Morris, S. E., and Wakefield, J. C. (2000). Point-source modelling using matched case-control data. *Biostatistics*, **1**, 89–105.

Dreassi, E., Lagazio, C., Maule, M. M., Magnani, C., and Biggeri, A. (2008). Sensitivity analysis of the relathionship between disease occurence and distance from a putative source of pollution. *Geospatial Health*, **2**, 263–271.

Ganguli, B. and Wand, M. P. (2005). SemiPar 1.0. *Users' Manual*. `http://www.maths.unsw.edu.au/wand/papers.html`.

Gustafson, P., Hossain, S., and MacNab, Y. C. (2006). Conservative prior distributions for variance parameters in hierarchical models. *Canadian Journal of Statistics*, **34**, 377–390.

Held, L., Natario, I., Fenton, S., Rue, H., and Becker, N. (2005). Towards joint disease mapping. *Statistical methods in medical research*, **14**, 61–82.

Hughes-Oliver, J. M., Heo, T., and Ghosh, S. K. (2008). An autoregressive point source model for spatial processes. *Environmetrics*, **20**, 575–594.

Kelsall, J. and Diggle, P. (1995). Kernel estimation of relative risk. *Bernoulli*, **1**, 3–16.

Knorr-Held, L. and Best, N. (2001). A shared component model for detecting joint and selective clustering of two diseases. *Journal of the Royal Statistical Society, Series A*, **164**, 73–85.

Knorr-Held, L. and Rue, H. (2002). On block updating in markov random field models for disease mapping. *Scadinavian journal of statistics*, **29**, 597–614.

Lawson, A. (1993). On the analysis of mortality events associated with a prespecified fixed point. *Journal of the Royal Statistical Society, Series A*, **156**, 363–377.

Spiegelhalter, D., Thomas, A., Best, N., and Lunn, D. (2004). *WinBUGS User Manual, Version 1.4.1.* `http://www.mrc-bsu.cam.ac.uk/bugs/`.

Wakefield, J. C. and Morris, S. E. (2001). The bayesin modeling of disease risk in relation to a point source. *Journal of the American Statistical Association*, **96**, 77–91.