Received: 27 April 2011, Revised: 13 September 2011, Accepted: 15 September 2011, Published online in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/env.1138

Spatial analysis of the risk of multiple cancers in relation to a petrochemical plant

Crescenza Calculli[a]*, Alessio Pollice[a] and Maria Serinelli[b]

In Environmental Epidemiology studies, the effects of the presence of a source of pollution on the population health can be evaluated by models that consider the distance from the source as a possible risk factor. We introduce a hierarchical Bayesian model in order to investigate the association between the risk of multiple pathologies and the presence of a single pollution source. Our approach provides the possibility to incorporate spatial effects and other confounding factors within a logistic regression model. 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 model is applied to data from a spatial case–control study to evaluate the association of the incidence of different cancers with the residential location in the neighborhood of a petrochemical plant in the Brindisi area (Italy). Copyright © 2011 John Wiley & Sons, Ltd.

Keywords: multiple disease risk; environmental pollution sources; thin-plate splines; spatial variation; MCMC

1. INTRODUCTION

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 under development. In the last two decades, the application of methods based on spatial point processes has caught on the epidemiological field more and more, specifically with respect to the spatial distribution of diseases in relation with suspect sources of environmental risk. Analyses including more than one disease outcome are still rare: the simultaneous treatment of data referred to several pathologies seems to be hardly touched by the point data literature (Diggle *et al.*, 1997; Lawson, 1993), while it has been the subject of a deeper debate in the disease mapping field (Knorr-Held and Best, 2001; Held *et al*., 2005; Tzala and Best, 2008). In this paper, the possibility of analyzing the simultaneous spatial behavior of several pathologies with respect to an environmental point source of pollution has been pursued through a modeling proposal inspired by several works more generally concerned with other areas of spatial statistics.

As a starting point, we consider a logistic regression model with a generalized additive model (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). Although distance may not be a particularly accurate surrogate for exposure, it can be easily measured for this case study, where increasing distance is considered as a proxy for decreasing risk with decreasing exposure to the pollution source. Diggle and Zheng (2005) suggested to extend the former type of predictor in order to incorporate a residual spatial component in the form of a Gaussian random field (GRF). In the point data literature, the explicit consideration of residual spatial variation was addressed in few other cases. Hughes-Oliver *et al.* (2008) decomposed the random part of a second-order non-stationary spatial model into the sum of an autoregressive point source model and a residual spatial conditional autoregressive model. Crainiceanu *et al.* (2008) suggested a bivariate extension of a binomial geostatistical model to map the spatial variation of disease prevalence. In their model, the spatially varying log-odds are given as a linear predictor including the effects of covariates and also a zero-mean stationary spatial process that represents the residual spatial variation given in the form of a thin-plate spline interpolator, as an efficient alternative to the use of stationary Gaussian processes. In spatial epidemiological studies, the possibility of residual spatial variation always exists, perhaps because of an infectious agent or spatially varying unmeasured risk factors (Diggle *et al*., 2000). In the present application, the inclusion of a residual spatial term in the predictor acknowledges for confounders related to the sociodemographic, geomorphologic, and economic structure of the area under consideration. This spatial effect, collecting residual variation across all diseases, reflects the underlying shared risk pattern and is especially suitable for rare diseases and for those with a less clear etiology.

This paper is structured in five sections. Sections 2 and 3 are, respectively, focused on the extension of models for the evaluation of the relative risk in the presence of an environmental source of pollution to the case of multiple pathologies and to the proposed model prior

* *Correspondence to: Crescenza Calculli, Dipartimento di Scienze Statistiche, "Carlo Cecchi" Università degli Studi di Bari "Aldo Moro" Bari, Italy. E-mail: e.calculli@dss.uniba.it*

[a] *Dipartimento di Scienze Statistiche "Carlo Cecchi", Università degli Studi di Bari "Aldo Moro", Bari, Italy*

[b] *ARPA Puglia, Bari, Italy*

specification in the Bayesian framework. Section 4 describes the characteristics of the case study. Section 5 is dedicated to the presentation and discussion of the main results of the application of a suite of alternative models to the data under consideration.

2. MODEL FORMULATION

In this section, we propose an extension of the conditional logistic regression model for case–control data (Diggle and Rowlingson, 1994) to the consideration of multiple pathologies. In the case of a single pathology, it can be assumed 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 $\rho(s)$ can be defined so that the two intensities result in being bound by the relation

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

Barring the constant $\widetilde{\mu}$, representing the ratio of the number of cases to controls, $\rho(s)$ 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 probability of observing a case at location s. Given (1), $\pi(s)$ can be expressed as

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

The estimation of the disease odds and of the relative risk $\rho(s)$ can thus be achieved through logistic regression modeling, coding cases and controls as 1 and 0 (Diggle *et al*., 1997). This approach enables us to consider a predictor including a set of terms that measure the effects of covariates and the spatial variation. In the presence of a putative point source of pollution, the specification of the predictor can explicitly consider the source location s_0 describing the risk variation at s according to the distance $d(s, s_0)$ (Diggle and Rowlingson, 1994). In order to consider an extension of the previous model to the risk of a set of K pathologies, let $Y_k(s)$ be the case–control label of the kth pathology for an individual living at location s (with $s = s_1, \ldots, s_{n_k}$ and $N = \sum_{k=1}^{K} n_k$). As before, we assume that $Y_k(s)$ is conditionally distributed as a Bernoulli with parameter $\pi_k(s)$ representing the probability to observe a case of the kth pathology at location s. We specify a mixed effects linear predictor as follows:

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

The predictor in (3) is characterized by a fixed trend component and a random term (as in Crainiceanu *et al*., 2008, and Hughes-Oliver *et al*., 2008, for different data contexts). Disease-specific fixed effects include the averages μ_k and the effects $\boldsymbol{\beta}_k$ of p individual covariates corresponding to as many non-spatial risk factors in $\mathbf{X}$. The distance function $f\{d(s, s_0); \boldsymbol{\theta}_k\}$ in (3) represents the disease-specific fixed effect of the distance from the source ruled by the parameter vector $\boldsymbol{\theta}_k$ ($k = 1, \ldots, K$). This function can be specified according to different parametric forms, considering the distance from the source together with effects caused by directional components and eventual peak effects (Lawson, 1993). As in Diggle and Rowlingson (1994), we consider an isotropic semi-Gaussian form

$$f\{d(s, s_0), \boldsymbol{\theta}_k\} = 1 + \alpha_k \exp\{-\phi_k d(s, s_0)^2\} \tag{4}$$

where α_k represents the excess of relative risk of the kth disease at the source location and ϕ_k can be interpreted as the decreasing rate of the same risk as a function of the distance from the source. For environmental exposures measured by residential distance, the choice of the isotropic semi-Gaussian specification ensures that the influence of the source decays and becomes null at infinite distance, as the distance function is not increasing and tends asymptotically to zero.

The random effect $S(s)$ represents a residual spatial component of the model, collecting residual variation across all pathologies and not depending on the presence of the source. This term averages residual spatial variation over pathologies and reflects all the variation not accounted for by the covariates, including the distance to the pollution source. Notice that $S(s)$ is the only element of the predictor common to the K diseases, so that neglecting it corresponds to considering K separate logistic regression models (each one for one of the K diseases). The standard approach would be to model $S(\cdot)$ as a stationary GRF, as proposed by Diggle *et al.* (1998) in a different context. For data on a regular lattice, Hughes-Oliver *et al.* (2008) model $S(\cdot)$ by an autoregressive point source process, an autoregressive spatial process whose specification depends on the proximity structure between points. In the case of non-regular spatial disease data, the application of such a model, other than being less convenient, brings a considerable computational complication because of the high number of parameters involved. As an alternative, the Gaussian process $S(\cdot)$ may be conveniently represented by a semi-parametric term 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 residential locations, the low-rank representation of $S(s)$ takes the following linear form:

$$S(s) = \mathbf{Z}(s)\mathbf{b} \tag{5}$$

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 sth row of the design matrix

$$\mathbf{Z} = \mathbf{Z}_T \boldsymbol{\Omega}_T^{-1/2} \tag{6}$$

where matrices $\mathbf{Z}_T$ and $\boldsymbol{\Omega}_T$ are, respectively, the spatial correlation matrix between the N subjects' residential locations and the T nodes and that among the nodes. Both $\mathbf{Z}_T$ and $\boldsymbol{\Omega}_T$ are based on an isotropic spatial correlation function in the form of the radial basis function $C(r) = r^2 \log(r)$, where r is the Euclidean distance. Then, $\mathbf{Z}_T = [C\{d(s,t)\}]$ and $\boldsymbol{\Omega}_T = \left[C\{d(t,t')\}\right]$, with $t, t' = 1, \ldots, t_T$, lead to a thin-plate spline representation for (5) and to a logistic GAM.

In the presence of a high number of sampling locations, the computational advantage implied by the previous low-rank representation is due to the fact that the smooth function $C(\cdot)$ is evaluated for a reduced number of nodes.

3. PRIOR SPECIFICATION

The logistic regression model with the predictor in (3) and (5) is estimated in the hierarchical Bayesian framework, and the prior distributions of fixed and random effects are fully specified in this section. Independent Gaussian vague priors $\mathrm{N}(0, 10^{-3})$ (where we report the mean and the precision of each Gaussian distribution) are assumed for fixed disease-specific effects μ_k and $\boldsymbol{\beta}_k$. For the parameter vector $\boldsymbol{\theta}_k = (\alpha_k; \phi_k)$, we propose the following prior structure with $k = 1, \ldots, K$:

$$\alpha_k \overset{\text{i.i.d.}}{\sim} \mathrm{Gamma}(a, c), \quad \phi_k \overset{\text{i.i.d.}}{\sim} \mathrm{U}(0, \phi_{\max}) \tag{7}$$

The Gamma distribution is widely used as a prior for α_k 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 c. Also, the choice of the hyperparameter $\phi_{\max}$ 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) suggested setting $\phi_{\max} = 0.2 \times d_{\max}$ or $\phi_{\max} = 0.5 \times d_{\max}$, where $d_{\max}$ is the maximum distance observed between two subjects' locations. In the following, we treat the disease-specific parameters α_k and ϕ_k as random effects through the following hyperprior structure:

$$a \sim \mathrm{Gamma}(e, f), \quad c \sim \mathrm{Gamma}(g, h), \quad \phi_{\max} \sim \mathrm{N}(\mu_\phi, \tau_\phi) \tag{8}$$

where the hyperparameters e, f, g, h, μ_ϕ, and τ_ϕ are set as to make Gamma and Normal distributions characterized by very high variability and scarcely informative with respect to model inferences. Finally, the random spatial effect $\mathbf{b}$ is given the following prior distribution:

$$\mathbf{b} \sim \mathrm{N}_T(\mathbf{0}_T, \tau_b \mathbf{I}_T) \tag{9}$$

where $\mathbf{0}_T$ and $\mathbf{I}_T$ stand for the T-dimensional null vector and identity matrix, respectively. The precision hyperparameter τ_b of the spatial residual component has a Gamma prior distribution. A general non-informative approach would assign small values of the Gamma shape and scale parameters (Gustafson *et al*., 2006). Crainiceanu *et al.* (2008) suggested that the choice of $\mathrm{Gamma}(0.001, 0.001)$ does not affect posterior estimates, although it causes slow convergence and poor mixing in Markov chain Monte Carlo estimation algorithms when the information conveyed by the data on precision hyperparameters is not sufficient. In the proposed model, this is particularly true for the spatial effect precision τ_b. For the application proposed in Sections 4 and 5, these considerations lead us to choose, after a fine-tuning phase, an informative specification of the Gamma prior symmetrical around a mean value moved away from zero and with small variance, as reported in Section 5 together with the whole specification of the prior structure.

4. THE BRINDISI CASE STUDY

A number of epidemiological studies of varying designs have investigated the environmental risk in the Brindisi area (southeast of Italy) because of the presence of a large petrochemical plant and industrial sites (power plants and chemical, pharmaceutical, metallurgical, and manufacturing plants). In the last WHO report (Martuzzi *et al.*, 2002), significant excesses in the mortality from all causes, cancers, and respiratory diseases were found for the Brindisi area in 1990–1994. Belli *et al.* (2004) reported a moderate increase in the risk of mortality for lung, bladder, and lymphohematopoietic neoplasms in the population resident within 2 km from the Brindisi petrochemical plant in 1996–1997. Based on the application of standardized criteria, the SENTIERI project (Studio Epidemiologico Nazionale dei Territori e degli Insediamenti Esposti a Rischio da Inquinamento, Pirastu *et al*., 2010), a national mortality study of residents in Italian polluted sites, assessed the association between 63 causes of death and selected environmental exposures for some sites of national interest for environmental remediation, including Brindisi. According to this study, cancers for the lung and trachea have limited or suggestive evidence of an association with the petrochemical plant exposure, while for bladder, non-Hodgkin, and leukemias, there is inadequate or insufficient evidence to determine whether an association exists. A recent geographical study confirms the results of the previous analysis and shows significant excesses in the mortality for all cancers, stomach cancers, and lung cancers among males (Gianicolo *et al*., 2008).

In order to further investigate the potential association between the environmental exposure to petrochemical plant emissions and some cancers, a case–control study based on incidence data was carried out. The area at risk is defined considering the city and three neighboring municipalities (Carovigno, San Pietro Vernotico, and Torchiarolo). 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 Cancer Registry of the Puglia region. Controls are 1694 subjects resident in the same area in 1999–2001, randomly selected and matched to the cases by disease, sex, date of birth and residential municipality. Although residential addresses from the Health Information System of the Puglia region were used for controls, no residential histories were available for the cases, so addresses at diagnosis (last residence or current residence) were used as a proxy of the relevant exposure. This contrasts with a previous study (Belli *et al*., 2004) that typically used the longest held residence with exclusion of the last 10 years of the complete residential history of each subject, weighted by the length of stay at each residence. All the addresses were geocoded to latitude and longitude (Figure 1), and coordinates were validated by visual inspection of the map. To account for the influence of the petrochemical plant, considered here as the main environmental risk source, we calculated the distance between the plant and each subject's residential location. The 2097 case and control subjects are rearranged into five different representative cancer classes as reported in Table 1.

The logistic regression model in (3)–(5) is applied to the data from the previously illustrated case–control study with the aim of simultaneously modeling the incidence of the five cancer diseases among subjects resident in the neighborhood of the petrochemical plant.

5. RESULTS

In applying the model in (3)–(5) with the prior specifications outlined in Section 3, the disease-specific effects of sex, and age were ignored because of cases and controls being matched by disease, sex, age and residential municipality in the sample design. In order to check the suitability of the proposed model for the data at hand, we compared the performance of a suite of alternative specifications using the posterior mean deviance ($\bar{D}$) and the Bayesian *Deviance Information Criterion* (DIC; Spiegelhalter *et al*., 2002). Starting from the previously specified prior structure, alternative models were in turn obtained by modifying one of the main characterizing features:

1. No distance function, common spatial component
2. Common distance function, no spatial component
3. Common distance function, common spatial component
4. Disease-specific distance function, no spatial component
5. Disease-specific distance function, common spatial component

The Bayesian Markov chain Monte Carlo estimation of the five specifications of the logistic model was implemented through the use of the software WinBUGS (Spiegelhalter *et al*., 2004). For each specification, posterior distributions of unknown parameters were obtained using 150,000 iterations, discarding the first 50,000 corresponding to a burn-in phase of the estimation algorithm. Initial values were obtained as the results of an earlier simulation run of a chain with overdispersed starting values. The convergence of chains was evaluated through the graphical inspection of the trace plots and Gelman–Rubin convergence statistics as modified by Brooks and Gelman (1998), provided by

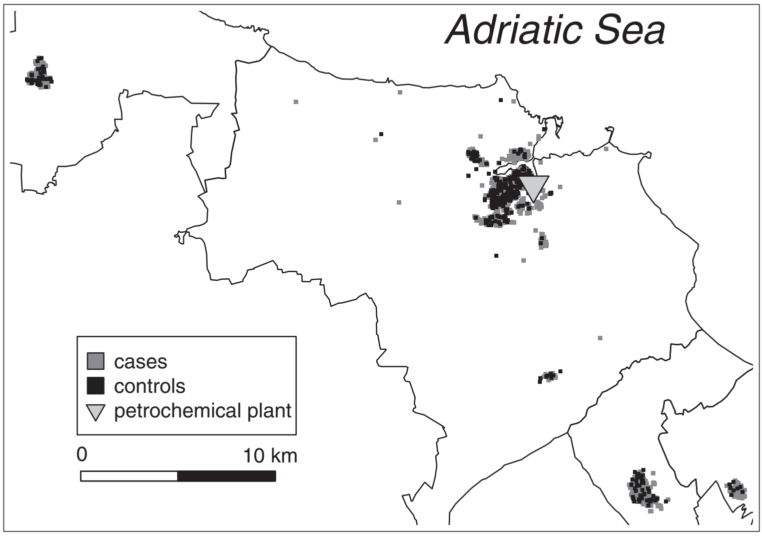

Figure 1. Case and control locations in the Brindisi area at risk

Table 1. Cases and controls classified by cancer types

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

Table 2. Comparison of alternative model specifications based on posterior mean deviance ($\bar{D}$), effective number of parameters (p_D), and deviance information criterion (DIC)

	Distance function					
Model	Common	Disease-specific	Spatial component	$\bar{D}$	p_D	DIC
1			X	2051.020	7.890	2058.910
2	X			2057.950	5.190	2063.140
3	X		X	2051.620	8.138	2059.760
4		X		2059.350	6.168	2065.510
5		X	X	2052.780	9.052	2061.830

the software. Model comparison was carried out via the posterior mean deviance ($\bar{D}$) and the DIC (Spiegelhalter *et al.*, 2002) reported in Table 2. Notice that the two measures of model fit vary accordingly and thus provide the same ordering for the five models.

In Table 2, we observe an expected behavior of the effective number of parameters (p_D) for the five models, with lower values when the semi-parametric spatial component is not included and higher values corresponding to more complex predictor structures. Notice that differences between the DIC values are very slight and less than 7 so they can be considered negligible as Spiegelhalter *et al.* (2002) suggest. As a consequence, the alternative specifications are substantially all equal in terms of model fit. The best performing predictor in model 1 does not contain the distance function but only the semi-parametric spatial relative risk surface. Models 3 and 5, obtained introducing the common and disease-specific effects of the distance from the petrochemical plant, respectively, show a slightly worse fit when compared to model 1. In order to verify whether the semi-parametric spatial component masks the spatial variability induced by the distance from the pollution source, we only considered the common and disease-specific distance functions in the predictors of models 2 and 4, obtaining the higher DIC values (worst fit). In both cases, it is clear from the values of the DIC that adding the residual spatial component improves the results in terms of model fit. Thus, if the aim is to verify the significance of the association between the proposed distance-decay function and risk, the assumption of the presence of a spatial component, collecting residual variation across all pathologies and not depending on the presence of the source, should be taken into account. Finally, the consideration of a common rather than disease-specific distance function proves to be more justified in terms of model fit either with or without the residual spatial component.

From the model comparison point of view, we get the overall conclusion that the isotropic semi-Gaussian distance-decay function provides a vague contribution to the variation of the disease risk. However, notice that differences between the DIC values are very slight and can be considered negligible; as a consequence, the alternative specifications are substantially all equal in terms of model fit. Then, in order to point out the main inferences obtainable by the proposed modeling strategy, we choose to show the results for the best model with disease-specific effects (model 5). For this model, the prior distributions of hyperparameters are specified as follows: $\alpha_k \sim \text{Gamma}(2, 1)$ for the excess of relative risk at the source, $\phi_k \sim \text{U}(0, 20)$ for the risk decrease rate, and $\tau_b \sim \text{Gamma}(9, 3)$ as suggested in Section 3. Posterior statistics concerning the main disease-specific parameters are reported in Table 3, where we can notice slight differences in the posterior medians 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 high standard errors for any choice of the distance function (Dreassi *et al.*, 2008). Consequently, distance function parameter estimates are close to their prior mean values and are characterized by high variability. Estimates of the distance functions are reported in Figure 2 for the five pathologies, showing similar values of the risk at the source (close to 2) and a decay vanishing for distances higher than 1 km for all pathologies. Credibility intervals of the disease-specific α parameters clearly overlap (Table 4); nevertheless, a higher risk at the source is estimated for bladder and lung cancer in line with the findings in Belli *et al.* (2004),

Table 3. Posterior medians and standard deviation of relevant disease-specific parameters for model 5

	μ		α		ϕ	
	Median	SD	Median	SD	Median	SD
Bladder	−2.546	0.321	1.042	1.928	11.930	5.372
Lung	−2.608	0.306	0.889	1.729	11.330	5.507
Non-Hodgkin lymphoma	−2.580	0.239	0.648	1.217	11.900	5.487
Leukemias	−2.487	0.194	0.435	0.655	11.510	5.333
Others	−2.528	0.195	0.294	0.609	12.150	5.642

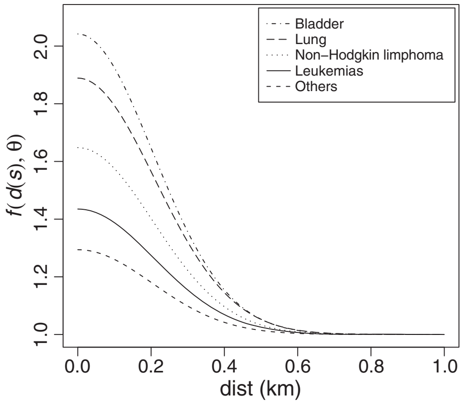

Figure 2. Curves of $f(d(s, s_0); \boldsymbol{\theta}_k)$ for different pathologies resulting from estimates of $\boldsymbol{\alpha}_k$ and $\boldsymbol{\phi}_k$

Table 4. 95% posterior credibility intervals of disease-specific α parameter for model 5

	α	
	2.5%	97.5%
Bladder	0.001	7.006
Lung	0.004	6.115
Non-Hodgkin lymphoma	<0.000	4.402
Leukemias	<0.000	2.347
Others	<0.000	2.157

who reported the value 3.1 to estimate the odds ratio of lung cancer mortality for residents within 2 km from the petrochemical plant, after adjusting for age, sex, smoking, occupation, and education (despite its magnitude, the odds ratio does not reach the statistical significance because of the scarcity of lung cancer cases). The estimated spatial effect common to all neoplasias $S(s)$ is reported in Figure 3, where we plot the predictions of the random effects **b** corresponding to $T = 50$ spatial nodes obtained by the `clara` space-filling algorithm with the R package `SemiPar` (Ganguli and Wand, 2005). The darkest points show an increase of the common risk at a distance of about 2500 m from the petrochemical plant. As we expected, this spatial effect largely reproduces the distribution of the population at risk, because of its

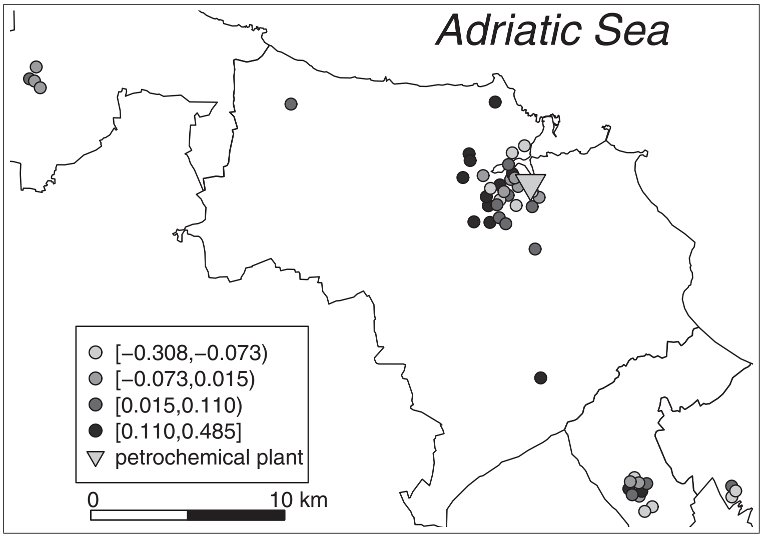

Figure 3. Map of the low-rank predicted Gaussian random field, $S(s)$ based on 50 spatial nodes

substantial overlapping with the location of subjects. Yet, notice that the estimated common spatial component adds to the estimated effects of the distance from the source. Therefore, it seems quite reasonable to expect an increased common spatial effect where the effects of the distance function fade out.

The mixed effects hierarchical model allows estimation of the effects caused by the presence of an environmental pollution source on the risk of various diseases, properly taking a common residual spatial pattern into account. When applied to the data from the present case study, the proposed methodology succeeds in catching the effects of the pollution source and the residual spatial component at the same time, although it fails in discriminating between the considered pathologies.

6. CONCLUDING REMARKS

This paper focuses on the simultaneous 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 logistic regression model that includes the possibility to incorporate disease-specific source-related spatial effects and a common residual spatial effect across all pathologies. The novelty of this approach is twofold: it enables extension of known models for spatial case–control data to the case of multiple pathologies and allows one to consider a common residual spatial component that can be given an epidemiological interpretation. The latter component is given in the form of a GRF approximated by the 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 geostatistics, above all in terms of computational advantages. The spatial effect specific to each disease is specified by a function that describes the changes in its relative risk due to the distance from the source. The Bayesian approach and the hierarchical model structure provide the possibility to properly deal with all the previous features and make the proposed model suitable to be used for the analysis of complex real cases such as the one concerning the Brindisi case study. The Brindisi area, at risk of environmental crisis and characterized by a strong anthropical pressure, has been the object of several epidemiological studies that pointed out an excess of risk in relation with some cancer pathologies and lung diseases. The present case–control study aims to evaluate the eventual association of some cancer pathologies with the distance of the residential location to a petrochemical plant. With respect to this objective, the application of the proposed model allows the estimation of, although with scarce significance, the excess of risk related to all the analyzed pathologies. The results obtained are generally consistent with the findings of Belli *et al.* (2004) who reported a high (although not statistically significant) risk of bladder (odds ratio, 3.9) and lung (odds ratio, 3.1) cancer mortality in the proximity of the petrochemical plant. The investigation of the residential proximity to specific industrial sites was addressed in a number of studies characterized by very different designs: increased incidences of cancers have been generally reported, although it is difficult to compare the associations observed for several reasons, including the sensitivity to the modeling choice, the confounding of variables taken into account as possible risk factors, the sample size, and the definition of the distance from the putative source. The study of Simonsen *et al.* (2010) was carried out to investigate the potential links between environmental exposure to emissions from a petrochemical site and lung cancer. Sans *et al.* (1995) studied the incidence of various diseases (all cancers, cancer of larynx, and leukemias) and mortality and reported an excess of the incidence for all and larynx cancers within 5–7 km from petrochemical works. The authors did not obtain evidence of the decline in risk with the distance from the plant for the incidence of leukemia. Also, no significant evidence of excess or decline in the risk of death for larynx, liver, and lung cancers was found in relation to the position of the plant. As mentioned above, in the Brindisi case study, the environmental exposure to emissions from the petrochemical plant was indirectly measured by an exposure surrogate based on the distance between the plant and the subjects' residential locations, using addresses at diagnosis for the cases. This is a limitation of the present study, where we do not account for subject mobility or length of stay at each residence. These are important issues, especially when the diseases are characterized by different latency times, ranging from over 30 years for mesothelioma to 10–20 years, for lymphohematopoietic malignancies and bladder cancer (Belli *et al*., 2004). A further major drawback is the inadequate characterization of air pollution exposure implied by the definition of residence associated to the use of distance as a proxy, raising concern for measurement error and increased uncertainties in risk estimates (Katsouyanni and Pershagen, 1997). Another limitation of the proposed analysis is that exposure to hazardous occupation, possibly altering the association between industrial site proximity and cancers, was not taken into account. However, the proposed model offers the advantage of efficiently considering a set of factors affecting the estimate of the risk of multiple diseases related to the presence of an environmental pollution source, integrating epidemiological and modeling aspects. Possible directions for future work include a thorough study of the properties of the proposed model by analytic computations and simulation experiments.

REFERENCES

Belli S, Benedetti M, Comba P, Vigotti MA, Portaluri M. 2004. Case–control study on cancer risk associated to residence in the neighbourhoods of petrochemical plant. *European Journal of Epidemiology* **19**(1): 49–54.

Brooks SP, Gelman A. 1998. Alternative methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics* **7**: 434–455.

Crainiceanu CM, Diggle PJ, 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 PJ, Morris S, Elliott P, 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 PJ, Morris SE, Wakefield JC. 2000. Point-source modelling using matched case–control data. *Biostatistics* **1**: 89–105.

Diggle PJ, Rowlingson BS. 1994. A conditional approach to point process modelling of elevated risk. *Journal of the Royal Statistical Society, Series A* **153**: 349–362.

Diggle PJ, Tawn JA, Moyeed RA. 1998. Model-based geostatistics (with discussion). *Applied Statistics* **47**: 299–350.

Diggle PJ, Zheng P. 2005. Nonparametric estimation of spatial segregation in a multivariate point process: bovine tuberculosis in Cornwall, UK. *Applied Statistics* **54**: 645–658.

Dreassi E, Lagazio C, Maule MM, Magnani C, Biggeri A. 2008. Sensitivity analysis of the relationship between disease occurrence and distance from a putative source of pollution. *Geospatial Health* **2**: 263–271.

Ganguli B, Wand MP. 2005. SemiPar 1.0—User Manual. http://www.maths.unsw.edu.au/wand/papers.html.

Gianicolo E, Serinelli M, Vigotti MA, Portaluri M. 2008. Mortality in the municipalities of Brindisi Province, 1981–2001. *Epidemiologia e Prevenzione* **32**(1): 49–57.

Gustafson P, Hossain S, MacNab YC. 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, Becker N. 2005. Towards joint disease mapping. *Statistical Methods in Medical Research* **14**: 61–82.

Hughes-Oliver JM, Heo T, Ghosh SK. 2008. An autoregressive point source model for spatial processes. *Environmetrics* **20**: 575–594.

Katsouyanni K, Pershagen G. 1997. Ambient air pollution exposure and cancer. *Cancer Causes and Control* **8**: 284–291.

Kelsall JE, Diggle PJ. 1995. Kernel estimation of relative risk. *Bernoulli* **1**: 3–16.

Knorr-Held L, 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.

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

Martuzzi M, Mitis F, Biggeri A, Terracini B, Bertollini R. 2002. Environment and health status of the population in areas with high risk of environmental crisis in Italy. *Epidemiologia e Prevenzione* **26**(6): 1–53.

Pirastu R, Ancona C, Iavarone I, Mitis F, Zona A, Comba P. 2010. S.E.N.T.I.E.R.I. Working Group. *Epidemiologia e Prevenzione* **34**: 1–2.

Sans S, Elliott P, Kleinschmidt I, Shaddick G, Pattenden S, Walls P, Grundy C, Dolk H. 1995. Cancer incidence and mortality near the Baglan Bay petrochemical works, South Wales. *Occupational and Environmental Medicine* **52**: 217–224.

Simonsen N, Scribner R, Su LJ, Williams D, Luckett B, Yang T, Fontham ETH. 2010. Environmental exposure to emissions from petrochemical sites and lung cancer: the Lower Mississippi Interagency Cancer Study. *Journal of Environmental and Public Health* **2010**: 1–10.

Spiegelhalter D, Best N, Carlin B, van der Linde A. 2002. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B* **64**: 583–639.

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

Tzala E, Best N. 2008. Bayesian latent variable modelling of multivariate spatio-temporal variation in cancer mortality. *Statistical Methods in Medical Research* **17**: 97–118.

Wakefield JC, Morris SE. 2001. The Bayesian modeling of disease risk in relation to a point source. *Journal of the American Statistical Association* **96**: 77–91.