Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
0(0) 1–20

DOI: 10.1177/0962280214536636
smm.sagepub.com
SAGE

Semiparametric M-quantile regression for count data

Emanuela Dreassi,[1] M Giovanna Ranalli[2] and Nicola Salvati[3]

Abstract

Lung cancer incidence over 2005–2010 for 326 Local Authority Districts in England is investigated by ecological regression. Motivated from mis-specification of a Negative Binomial additive model, a semiparametric Negative Binomial M-quantile regression model is introduced. The additive part relates to those univariate or bivariate smoothing components, which are included in the model to capture nonlinearities in the predictor or to account for spatial dependence. All such components are estimated using penalized splines. The results show the capability of the semiparametric Negative Binomial M-quantile regression model to handle data with a strong spatial structure.

Keywords

disease mapping, ecological regression, geoadditive models, Negative Binomial, penalized splines, robust method

1 Introduction

The aim of this paper is the ecological regression analysis for data on lung cancer incidence over 2005–2010 for 326 Local Authority Districts (LADs) in England. The data set is obtained from the UK Public Health Observatory. The total number of observed cases is 165,476 (with the highest number 3176 for Birmingham) for 313,404,270 person-years. The expected number of cases is defined using internal standardization based on age-specific rates. The Standardized Mortality Ratio (SMR) is defined as the ratio between observed and expected number of cases. The SMRs range from 0.54 to 1.90. Figure 2 (top right panel) shows the geographical distribution of their quintiles.

Lung cancer is the most common cancer worldwide and the second most prevalent in England. The latest analysis of lung cancer incidence throughout England reports the highest values in the north of England and lower values in the east, south-east and south-west, reflecting regional variation in smoking prevalence. Moreover, lung cancer incidence is strongly related to

[1]Dipartimento di Statistica, Informatica, Applicazioni, Università di Firenze, Firenze, Italy
[2]Dipartimento di Scienze Politiche, Università di Perugia, Perugia, Italy
[3]Dipartimento di Economia e Management, Università di Pisa, Pisa, Italy

Corresponding author:
Emanuela Dreassi, Dipartimento di Statistica, Informatica, Applicazioni, Università di Firenze, Firenze, Italy.
Email: dreassi@disia.unifi.it

deprivation and there is a clear trend of increasing values with increasing levels of deprivation. Hence, in this work, we consider the index of deprivation and the smoking habit as covariates.

Indices of Deprivation 2007 (ID 2007 hereafter) are the English Government's official measure of multiple deprivation at small area level. The Index of Multiple Deprivation 2007 (IMD), which forms part of the ID 2007, is based on the small area geography known as Lower Super Output Areas. In most cases, these are smaller than wards, thus allowing for the identification of small pockets of deprivation. IMD brings together 37 different indicators which cover specific aspects or dimensions of deprivation: education, income, employment, health and disability, skills and training, barriers to housing and services, living environment and crime. These are weighted and combined to create the overall IMD. The majority of the data underpinning the IMD refers to 2005 although some data cover a number of years, for example an average of the period 2003–2005. ID 2007 includes the following measures of deprivation at 2007: the IMD as described above; County Council summaries of the IMD; LAD summaries of the IMD; income deprivation affecting older people index and income deprivation affecting children index. Proportion of adults who smoke (SMOKE 2007 hereafter) refers to the population aged 16 and older who are cigarette smokers and is measured using data from the 2007 edition of the General Lifestyle Survey, formerly known as the General Household Survey. Both covariates are considered as score, so effect on disease risk has been considered for a score point. Figure 1 shows the relationship between the relative risk for lung cancer and these covariates in an ecological regression setting.

A Negative Binomial Generalized Linear Model (NBGLM) has been first fitted to these data. By introducing expected cases as an offset term and a log link, we model the logarithm of the disease relative risks instead of the count: log(relative risk) = intercept + ID 2007 + SMOKE 2007, model (a). This is a quite good model in which both covariates are highly significant with residual Deviance 318.1 on 321 degrees of freedom and Akaike Information Criteria (AIC) 3518.1. However, the right panel of Figure 1 may suggest a nonlinear relation with ID 2007. This leads to consider a Negative Binomial Generalized Additive Model (NBGAM,[1] Section 5.4) as a plausible model: log(relative risk) = intercept + s(ID 2007) + SMOKE 2007, model (b). This model has residual deviance 320.3 on

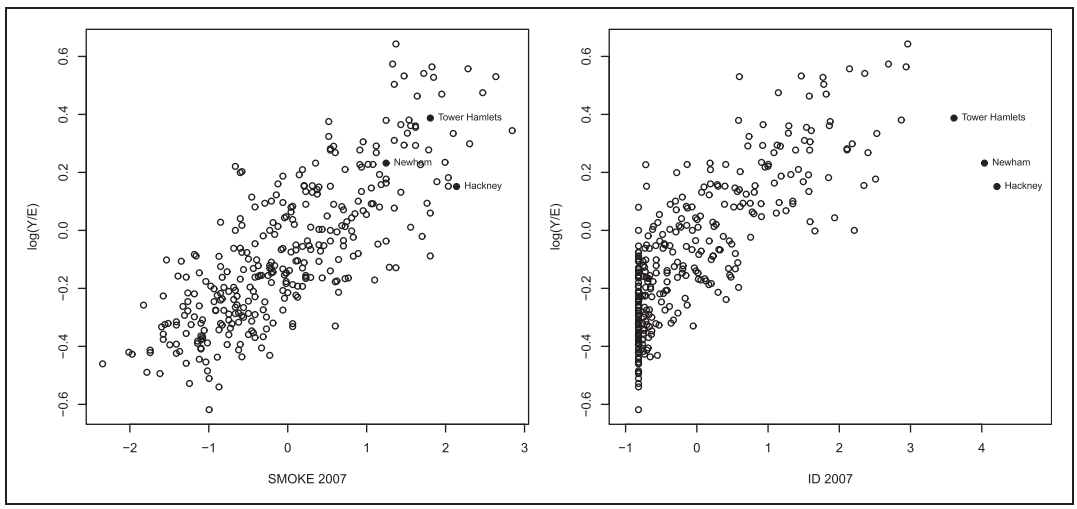

Figure 1. Scatter plots of the relative risk versus SMOKE 2007 (left) and versus ID 2007 (right).

314.1 effective degrees of freedom and a smaller AIC (3490.9) with respect to model (a). The nonlinear component in ID 2007 is tested to be significant. Figure 2 shows the maps for raw data (SMR), the covariates (SMOKE 2007 and ID 2007) and Pearson residuals (range between −2.33 and 3.18) from model (b) fitted using the mgcv package of R. All maps are strongly spatially structured and therefore suggest the inclusion of spatial information into the model. The latter consists of the geographical coordinates (latitude and longitude) of each centroids of the LADs normalized so that they take values in [0; 1].

The residual spatial structure of model (b) has been included in model (c) using a bivariate penalized spline: log(relative risk) = intercept + s(ID 2007) + SMOKE 2007 + s(lon, lat). The spatial component is highly significant. The model has a residual deviance 323.1 on 258.4 effective degrees of freedom and AIC is 3320.9, that is sensibly smaller than the others. Top panels of Figure 3 show both smoothing components of model (c) on the linear predictor scale: the left top plot exhibits the nonlinear relationship between the response variable and ID 2007; the right top plot illustrates the bivariate penalized spline. It can be noted that the latter spline modeling the spatial component gives the highest contribute in the north of England where the disease risk for lung cancer is indeed quite large. On the other hand, the nonlinear relationship with ID 2007 seems to be driven essentially by the behavior of three areas (Hackney, Newham, and Tower Hamlets, see the right panel of Figure 1), which is different from the others: these neighboring areas (at the North East of London) have high values for ID 2007 but moderate lung cancer mortality. Therefore, once the spatial structure has been included in the model, a linear relation between the logarithm of the relative risk and ID 2007 could be valid: the three areas are neighboring and the spatial spline could adjust for them. For this reason a reduced NBGAM has been fitted: log(relative risk) = intercept + ID 2007 + SMOKE 2007+s(lon, lat), model (d). This latter model has a residual deviance 319.7 on 263.9 effective degrees of freedom and AIC 3341.8.

Bottom panels of Figure 3 show the parametric component for ID 2007 and the spatial component for model (d). The spatial structure is very similar in complexity to that of model (c) and, in fact, uses a similar number of degrees of freedom. Model (c) shows a smaller AIC value than model (d), thereby suggesting that the nonlinear relationship with ID 2007 is significant. It should be noted, however, that since the estimate of the overdispersion parameter of the Negative Binomial (NB) model takes different values in these models, model comparison via AIC is to be taken with care. In addition, model (c) and model (d) fit, similarly, the data: Figure 4 (left panels) shows the distribution of the Pearson residuals obtained by models (c) and (d). For each LAD, the Pearson residual is the raw residual (the difference between the observed number of cases and its predicted value from the model) divided by the square root of the variance function.

The parametric effect of ID 2007 from model (d) seems to be mitigated by the values taken by the three areas highlighted in Figure 1. Indeed, the relatively smaller value for SMR shown by these three areas can be the effect of the activity of the North East London Cancer Network that in the last decade has promoted screening and knowledge of cares and prevention. Therefore, another approach can consider the three areas as possible outliers and use a robust approach. In addition, plots from Figure 4 indicate the presence of other potential outlier observations in the data, with a number of large Pearson residuals (absolute value greater than 2). This fact is also confirmed by the right panels of Figure 4 which plot the raw residuals against the fitted values from models (c) and (d).[1] The x-axis ranges between 0 and 1500 in order to show clearly over 99% of the observations. The plots suggest higher variability when the predicted value of the response variable is large. These diagnostics, showing the presence of potential model mis-specification, suggest that the use of an alternative to the models (c) and (d) may be justified in this case. Another arising problem is that the covariates are obtained from UK Public Health Observatory data using small area estimation

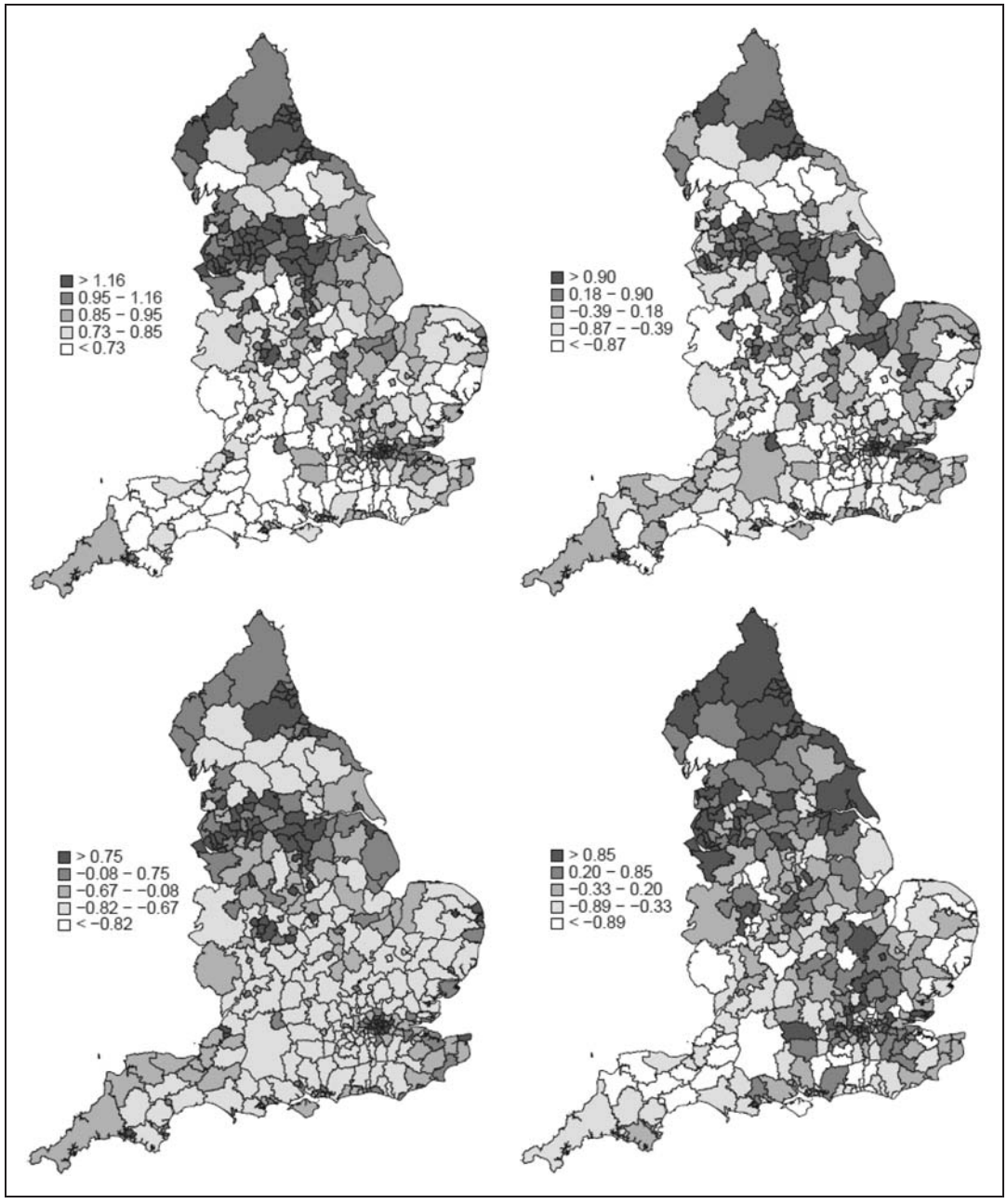

Figure 2. Maps of SMR (top left), SMOKE 2007 (top right), and ID 2007 (bottom left), Pearson residuals from model (b) (bottom right).

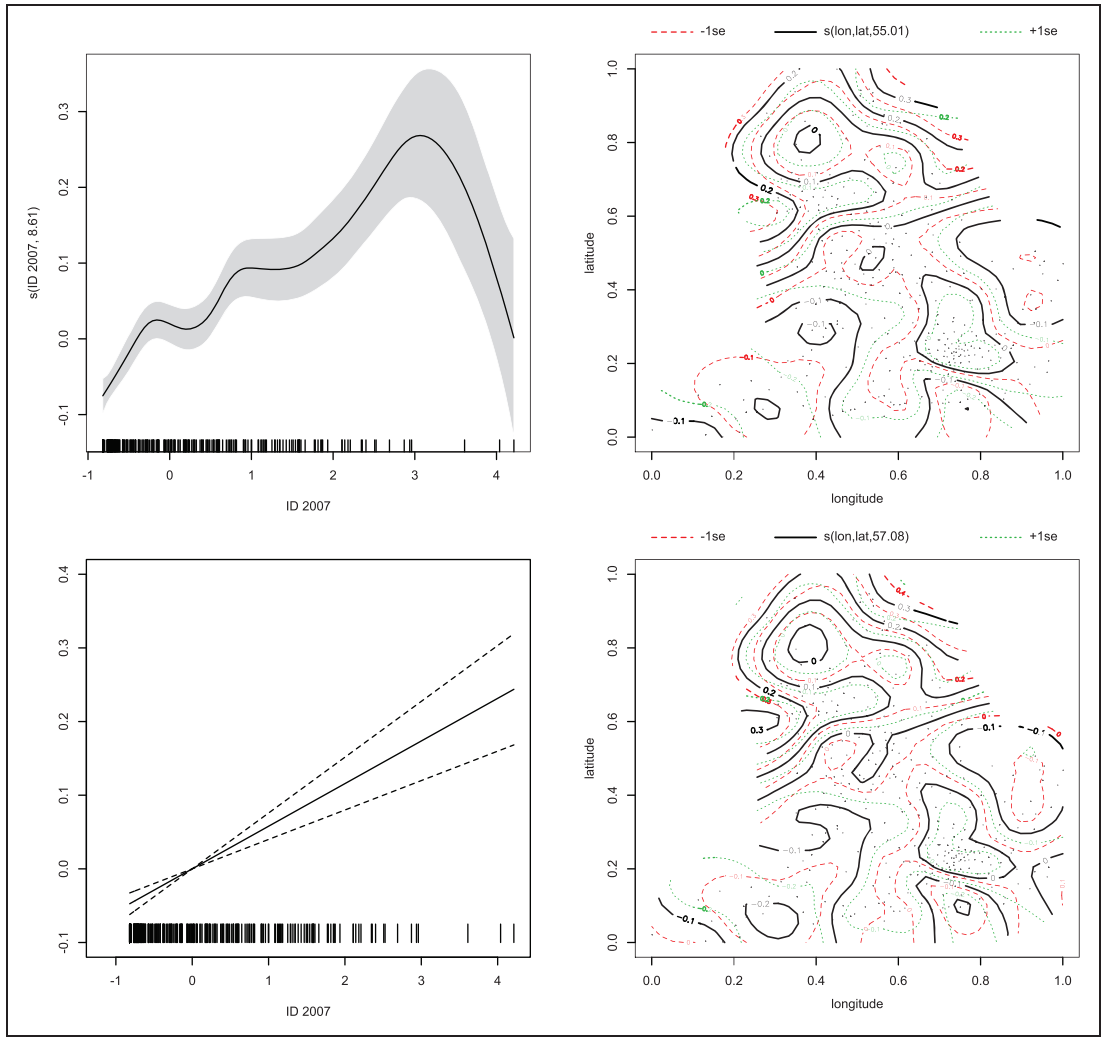

Figure 3. Penalized spline for ID 2007 (top left) and bivariate penalized spline (top right) from model (c). Parametric term for ID 2007 (bottom left) and bivariate penalized spline (bottom right) from model (d).

methods, so that they could have both sampling and nonsampling errors. For all these reasons, an outlier-robust approach seems to be a reasonable alternative to the traditional models.

Recently, semiparametric models have been extensively used on ecological regression in disease mapping.[2–7] Such models aim to cope with various types of drawbacks that could lead to a misspecification, for example:

(i) measurement errors on covariates[3,8–11];
(ii) the possible nonlinear relation between relative risk and covariates[12];
(iii) the robustness problem which can arise when disease risks and covariates exhibit a strong spatial pattern and spatial structured terms are included.[13–15]

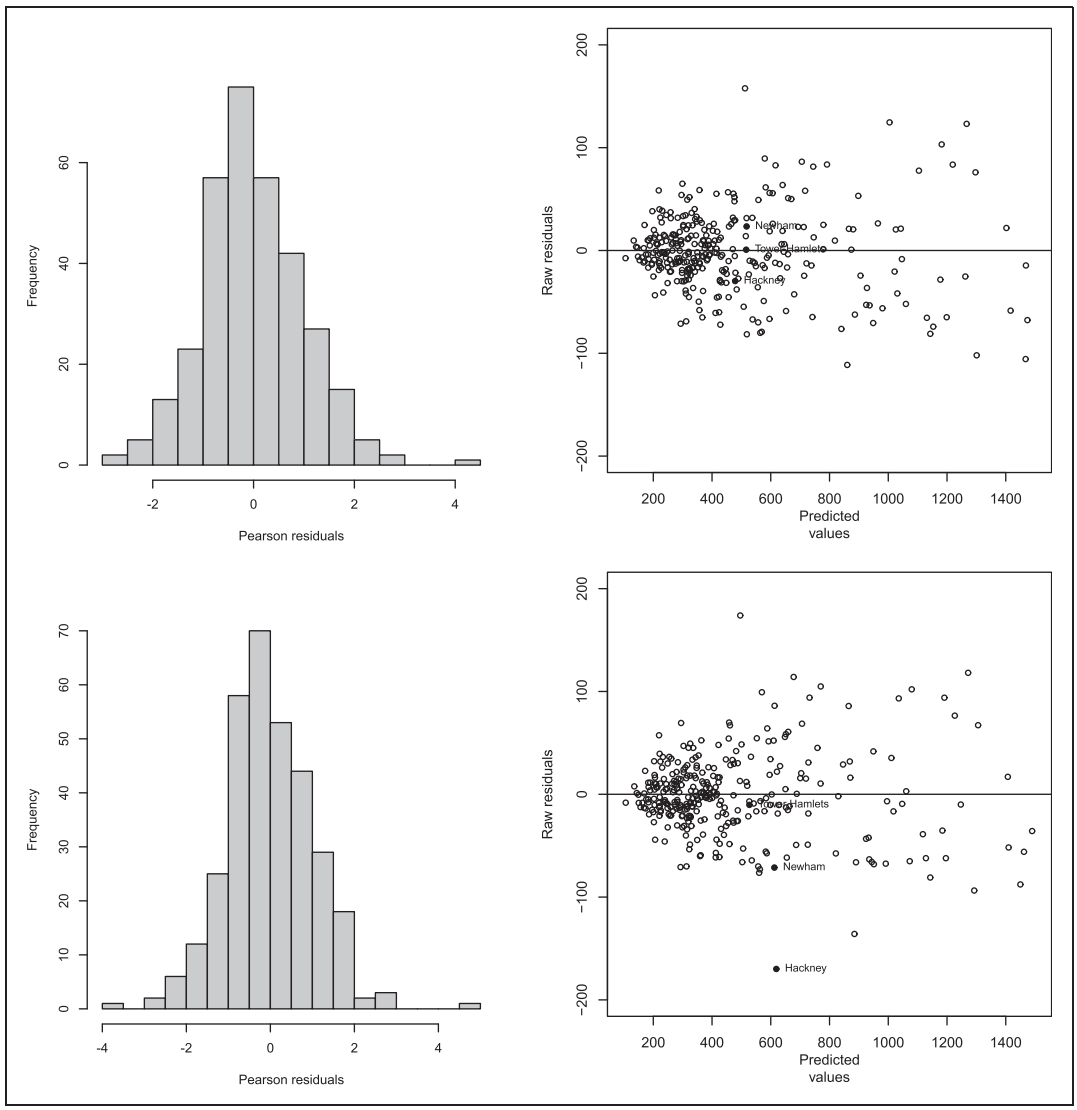

Figure 4. Pearson residuals (top left) and raw residuals versus predicted values (top right) from model (c). Pearson residuals (bottom left) and raw residuals versus predicted values (bottom right) from model (d).

In Chambers et al.,[3] the use of M-quantile regression is suggested to cope with problem (i) but (ii) is neglected and (iii) is only sketched.

This paper aims to extend the model introduced in Chambers et al.,[3] in order to handle (ii) and (iii) as well. Specifically, since disease risk and the covariates exhibit a strong spatial structure, a structured spatial component is included in the M-quantile predictor, following Pratesi et al.,[16] by means of a bivariate spline. The use of splines for the spatial part on disease mapping has been suggested, among others, in the literature.[5–7,17–20] In addition, NBGAM suggests the presence of a nonlinear relation between disease risk and one of the covariates (ID 2007). We would like to

investigate whether a robust version of models (c) and (d) may fit the data better and whether a nonlinear component in ID 2007 is really needed or not.

This paper is organized as follows. Section 2 introduces the basic notation and briefly reviews robust methods for NB data and the NB M-quantile model of Chambers et al.[3] Section 3 illustrates how the latter model is modified to take issues (ii) and (iii) into account. Results from the application to lung cancer incidence ecological regression are reported in Section 4. Finally, Section 5 concludes the paper with some comments and further topics of research.

2 Notation and background information

We denote by Y_i and t_i the number of observed and expected cases for area $i = 1,\ldots,n$, respectively. In disease mapping, NB models have been used because Y_i's usually exhibit overdispersion with respect to a Poisson model. It is common to use the mean parametrized NB,[21] where $E[Y_i] = \mu_i$ and $\mathrm{Var}[Y_i] = \mu_i + \mu_i^2/\theta$, and θ is a parameter that measures overdispersion. In particular, smaller values of θ suggest larger overdispersion. In an NBGLM setting, the aim of the ecological regression is to model μ_i. In this context, $\boldsymbol{x}$ is a $p \times 1$ vector of explanatory variables (which is assumed to include the constant term). The expected value of Y_i given $\mathbf{x}_i$ is modelled as $\mu(\mathbf{x}_i; \boldsymbol{\beta}) = \exp \eta_i = t_i \exp(\mathbf{x}_i^T \boldsymbol{\beta})$, for $i = 1, \ldots, n$, where $\boldsymbol{\beta}$ is a vector of p regression parameters and t_i represents an offset term. This leads to model relative risks instead of counts. Since the NB distribution is a member of the exponential family for fixed θ, this model is a special case of Generalized Linear Models (GLMs), with the $\log(\cdot)$ link function. In line with standard practice,[22–24] GLM methodology can be used to estimate $\boldsymbol{\beta}$, by replacing θ with a suitable estimate $\hat{\theta}$ (e.g. obtained using the method of moments) and by iterating estimation of $\boldsymbol{\beta}$ given $\hat{\theta}$.

When there is a concern about the presence of influential observations, Cantoni and Ronchetti[25] propose an approach to robust inference for GLMs based on quasi-likelihood. In particular, their robust version of the estimating equations for the parameter $\boldsymbol{\beta}$ of a GLM is of the form

$$n^{-1} \sum_{i=1}^{n} v(y_i, \mu_i) w(\mathbf{x}_i) \mu_i' - v(\boldsymbol{\beta}) = \mathbf{0} \tag{1}$$

where $E[Y_i] = \mu_i = \mu(\mathbf{x}_i; \boldsymbol{\beta}) = g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})$, $V[Y_i] = V(\mu_i)$, μ_i' is the derivative of μ_i with respect to $\boldsymbol{\beta}$, $v(\boldsymbol{\beta}) = n^{-1} \sum_{i=1}^{n} E[v(y_i, \mu_i)] w(\mathbf{x}_i) \mu_i'$ ensures the Fisher consistency of the estimator, and $v(y, \mu)$ is a bounded function of model residuals that controls the influence of the errors in y-space. Finally, weights $w(\mathbf{x}_i)$ are used to downweight leverage points. When $w(\mathbf{x}_i) = 1$, for $i = 1, \ldots, n$ and $v(y,\mu)$ is defined by Pearson residuals and the Huber influence function, Cantoni and Ronchetti[25] call the estimator defined by the solution to equation (1) as a Huber quasi-likelihood estimator. Note that the solution to equation (1) can be obtained numerically by a Fisher scoring procedure.

Chambers et al.[3] apply this approach to robust fitting of the mean parameterized NB model using the following estimating equations

$$\Psi(\boldsymbol{\beta}, c) := n^{-1} \sum_{i=1}^{n} \boldsymbol{\psi}(y_i, \mu_i) = \mathbf{0} \tag{2}$$

where $\boldsymbol{\psi}(y_i, \mu_i) = \psi(r_i) w(\mathbf{x}_i) V^{-1/2}(\mu_i) \mu_i' - v(\boldsymbol{\beta})$; $r_i = V^{-1/2}(\mu_i)(y_i - \mu_i)$ are Pearson residuals; $\psi(\cdot)$ is the Huber Proposal 2 influence function, such that

$$\psi(r) = r\, I(-c < r < c) + c\, \mathrm{sgn}(r)\, I(|r| \geq c) \tag{3}$$

where $I(\cdot)$ is the indicator function and c is the tuning constant; $\mu_i = t_i \exp(\mathbf{x}_i^T \boldsymbol{\beta})$; $\mu_i' = \mu_i \mathbf{x}_i^T$; $V(\mu_i) = \mu_i + \mu_i^2/\theta$, with $\theta > 0$ the shape parameter. The correction term $\nu(\boldsymbol{\beta})$ can be computed explicitly for the NB model (see Chambers et al.[3] for details) and a robust procedure is also proposed to estimate the scale parameter θ. In particular, a robust scale estimator[26] is considered defined by

$$n^{-1} \sum_{i=1}^{n} \left\{ \psi^2(r_i) - E\left[\psi^2\left(\frac{y_i - \mu_i}{V^{1/2}(\mu_i)} \right) \right] \right\} = \mathbf{0} \tag{4}$$

where $E\left[\psi^2\left(\frac{y_i - \mu_i}{V^{1/2}(\mu_i)} \right) \right]$ is a constant that ensures that the solution to equation (4) is Fisher consistent. Equations (2) and (4) can be solved by iterating between a solution to equation (2) given θ and a solution to equation (4) given $\boldsymbol{\beta}$.

Linear M-quantile regression provides a 'quantile-like' generalization of linear regression based on influence functions.[27] The M-quantile of order q of a continuous random variable Y with continuous distribution function $F(Y)$ is the value Q_q such that

$$\int \psi_q\left(\frac{Y - Q_q}{\sigma_q} \right) dF(Y) = 0 \tag{5}$$

where $\psi_q(r) = 2\psi(r)\{qI(r > 0) + (1 - q)I(r \leq 0)\}$, and ψ is a user-defined influence function. Here σ_q is a suitable measure of scale of the random variable $Y - Q_q$. Note that when $\psi(r) = \text{sgn}(r)$ we obtain the standard quantile of order q, while when $\psi(r) = r$ we obtain the expectile of order q, which represents a quantile-like generalization of the mean.[28] In this paper, we will always consider as influence function the Huber proposal 2 in equation (3), that depends on the tuning constant c. A linear M-quantile regression model[27] assumes that the q-th M-quantile of the conditional distribution of Y given $\boldsymbol{x}$, denoted by $Q_q(\boldsymbol{x}; c)$ is such that

$$Q_q(\boldsymbol{x}; c) = \boldsymbol{x}^T \boldsymbol{\beta}_{q,c} \tag{6}$$

where $\boldsymbol{\beta}_{q,c}$ is an unknown vector of q-th M-quantile regression slope coefficients which are independent of $\boldsymbol{x}$. For ease of notation we will drop subscript c and use $\boldsymbol{\beta}_q$. For specified q and continuous ψ, an estimate $\widehat{\boldsymbol{\beta}}_q$ of $\boldsymbol{\beta}_q$ is obtained by solving the set of estimating equations

$$n^{-1} \sum_{i=1}^{n} \psi_q\left\{ (y_i - \mathbf{x}_i^T \boldsymbol{\beta}_q) \hat{\sigma}_q^{-1} \right\} \mathbf{x}_i = \mathbf{0} \tag{7}$$

where $\hat{\sigma}_q$ is a suitable robust estimator of scale, e.g. $\hat{\sigma}_q = \text{median } |r|/0.6745$. Provided the tuning constant c is bounded away from zero, we can solve equation (7) using standard iteratively reweighted least squares. Furthermore, if ψ is continuous and monotone nondecreasing (e.g. a Huber-type function) then this algorithm is guaranteed to converge to a unique solution.[29]

M-quantile regression models allow to trade robustness for efficiency by properly tuning such constant c: robustness is increased as c decreases, while efficiency is increased as c increases with expectile regression in the limit as $c \to \infty$. In this way, we allow for a different set of regression parameters for each value of q as for quantile regression, with the extra-flexibility of the ψ function that may ensure robustness and a unique solution to the estimation procedure. Now, the quantile function of a discrete random variable is not generally a monotone increasing function of q, so a

unique solution to equation (7) for distinct values of q does not exist if Y is a count and $\psi(r)=\text{sgn}(r)$. However, this is not the case if ψ is a continuous monotone nondecreasing function, for which a unique solution always exists provided the expectation exists. This allows the concept of regression M-quantiles to be extended to count data in a straightforward way. In the case of NB data, Chambers et al.[3] consider the following model for the q-th M-quantile of the conditional distribution of Y given $\boldsymbol{x}$

$$Q_q(\boldsymbol{x}, t; c) = t \exp(\boldsymbol{x}^T \boldsymbol{\beta}_q) \tag{8}$$

In order to estimate $\boldsymbol{\beta}_q$ Chambers et al.[3] consider the extension of equation (2) to the M-quantile case, by replacing μ_i by $Q_q(\mathbf{x}_i,t_i;c)$, leading to the estimating equations

$$\Psi(\boldsymbol{\beta}_q, c) := n^{-1} \sum_{i=1}^{n} \psi_q(y_i, Q_q(\mathbf{x}_i, t_i; c)) = \mathbf{0} \tag{9}$$

3 Semiparametric M-quantile regression models for Negative Binomial data

Semiparametric regression models are a powerful extension of linear regression models in which one or more explanatory variables enter the model nonparametrically, i.e. without a pre-specified functional form. Among the many methods proposed in the literature to estimate such models, that based on penalized (*P*-) splines has gained much attention in the literature in the past 15 years, thanks to its simplicity and flexibility to be extended to handle very diverse situations (see Ruppert et al.[30] for an introduction and Ruppert et al.[31] for a more up to date review). In this section, we first illustrate how to import the flexibility of such models for M-quantile regression and when data are NB, describe the estimation procedure and then provide details on how to use such models for disease mapping.

3.1 Model specification

Pratesi et al.[16] introduce semiparametric M-quantile regression models to enhance the flexibility of linear M-quantile regression models and to grant robustness to classical semiparametric regression models based on *P*-splines. In M-quantile semiparametric regression, the q-th M-quantile of the conditional distribution of Y given $\boldsymbol{x}$ is assumed to take the form of an additive model, e.g.

$$Q_q(\boldsymbol{x}, c) = \beta_{0,q} + \beta_{1,q} x_1 + \tilde{m}_q(x_2) + \tilde{m}_q(x_3, x_4) \tag{10}$$

in which explanatory variable x_1 enters the model linearly, while the association with x_2 (x_3, x_4) is assumed to be a univariate (bivariate) unknown smooth function $\tilde{m}_q(\cdot)$. Such smooth functions are then estimated using *P*-splines.

In this paper, we wish to import the flexibility of such a model to the case in which our data are NB. In particular, we wish to extend the linear predictor of model (8) to encompass the presence of smooth components as those reported in equation (10). Hence, as an example of additive model for the linear predictor of the NB M-quantile model, we can consider the following form for the q-th M-quantile of the conditional distribution of Y given $\boldsymbol{x}$

$$Q_q(\boldsymbol{x}, t; c) = t \exp\left\{\beta_{0,q} + \beta_{1,q} x_1 + \tilde{m}_q(x_2) + \tilde{m}_q(x_3, x_4)\right\} \tag{11}$$

The aforementioned exemplification will turn out to be particularly useful for handling the data considered in this paper. In fact, model (c) considered in the Introduction is of the form log(relative risk) = intercept + SMOKE 2007 + s(ID 2007) + s(lon,lat). Note that model (d), on the other hand, simplifies equation (11) by fitting a parametric component for ID 2007, log(relative risk) = intercept + SMOKE 2007 + ID 2007 + s(lon,lat), assuming $Q_q(\boldsymbol{x}, t; c) = t\exp\{\beta_{0,q} + \beta_{1,q}x_1 + \beta_{2,q}x_2 + \tilde{m}_q(x_3, x_4)\}$.

Both smooth components in equation (11) will be handled using *P*-splines. Let us first focus on the univariate smooth component $\tilde{m}_q(x_2)$ that, as in classical nonparametric regression based on *P*-splines, is assumed to be approximated sufficiently well by a polynomial spline of degree l. In particular, $\tilde{m}_q(x_2)$ will be approximated by

$$m_q[x_2; \boldsymbol{b}_q] = \sum_{k=1}^{K} b_{k,q} B_k^{(l)}(x_2) \tag{12}$$

where the $B_k^{(l)}(x_2)$ are B-splines basis functions,[32] $b_{k,q}$ are the corresponding coefficients for $k = 1, \ldots, K$ and K is the dimension of the basis that depends on the number of fixed knots. If K is sufficiently large, the class of functions in equation (12) is very large and can approximate most smooth functions. In particular, it is common practice in the *P*-splines context, to place at most 35 knots (for small datasets, one knot every 4–5 observations) at uniformly spread quantiles of the unique values of x_2. Pratesi et al.[16] consider truncated polynomial basis functions in equation (12) instead of B-splines to approximate the function $\tilde{m}_q(\cdot)$. Truncated polynomial bases are easier to interpret, but usually exhibit a large degree of correlation that hinders convergence of numerical methods. Other bases can be used like radial or tensor basis functions (see Wood[1] for details on this).

Given the large number of knots, their influence is limited by putting a constraint on the size of the spline coefficients. Using B-splines, a difference penalty is applied to the vector of parameters $\boldsymbol{b}_q = (b_{1,q}, \ldots, b_{K,q})^T$. In particular, let $\boldsymbol{D}^{(d)}$ denote a difference matrix of order d; then the penalty matrix is given by $\boldsymbol{P}_1 = \boldsymbol{D}^{(d)T}\boldsymbol{D}^{(d)}$ and provides a penalty made of squared d-th order differences in the sequence of coefficients. For example, if $d = 1$, then $\boldsymbol{b}_q^T\boldsymbol{P}_1\boldsymbol{b}_q = \sum_{k=1}^{K-1}(b_{k+1,q} - b_{k,q})^2$.

The bivariate smooth component $\tilde{m}_q(x_3, x_4)$, on the other hand, is approximated using a set of radial basis functions (see e.g. Ruppert et al.[30]: p. 253). In particular

$$m_q[x_3, x_4; \alpha_{1,q}, \alpha_{2,q}, \boldsymbol{a}_q] = a_{1,q}x_3 + \alpha_{2,q}x_4 + \sum_{j=1}^{J} a_{j,q}Z_j(x_3, x_4) \tag{13}$$

where $a_{j,q}$, for $j = 1, \ldots, J$ are the coefficients of the radial basis functions

$$Z_j(x_3, x_4) = ||(x_3, x_4) - (\kappa_{3j}, \kappa_{4j})||^2 \log||(x_3, x_4) - (\kappa_{3j}, \kappa_{4j})||$$

that depend on the distance in the (x_3, x_4)-space between a point and the knot $(\kappa_{3j}, \kappa_{4j})$. Such basis functions are radially symmetric about each point and are, therefore, rotationally invariant. This is a particularly useful property when smoothing geographical data because the final smooth is independent of axis orientation. The choice of the number J and location of knots in two dimensions is more challenging than in the univariate case, because knots need to be a subset of the observations nicely scattered to cover the (x_3, x_4)-domain. Ruppert et al.[30] suggest $J = \max\{20; \min(n/4; 150)\}$ as a default in two dimensions. The location of the knots can be

determined using space filling designs or clustering methods. For the present data, we will use the clara function of the clustering package of R.

As for the univariate case, the influence of the knots is limited by putting a penalty to the vector of spline coefficients. For radial smoothers, the penalty matrix can be defined to be

$$\mathrm{P}_2 = ||(\kappa_{3j}, \kappa_{4j}) - (\kappa_{3j'}, \kappa_{4j'})||^2 \log||(\kappa_{3j}, \kappa_{4j}) - (\kappa_{3j'}, \kappa_{4j'})||_{1\le j,j'\le J}.$$

It is the $J \times J$ matrix of values given by the radial basis functions applied to the knots and therefore the penalty $\boldsymbol{a}_q^T \boldsymbol{P}_2 \boldsymbol{a}_q$ has the radial symmetry property. In addition, such a choice corresponds to the thin plate spline family of smoothers and to the penalization of the second derivative. In case of large correlation among the Z_j's, a transformation can be performed using Demmler–Reinsch orthogonalization to improve the performance of numerical methods (see e.g. Ruppert et al.,[30]: Appendix B).

3.2 Estimation procedure

Given data $\{y_i, t_i, x_{1i}, x_{2i}, x_{3i}, x_{4i}\}$, for $i = 1, \ldots, n$, let

$$\mathbf{u}_i = (1, x_{1i}, x_{3i}, x_{4i}, B_1^{(l)}(x_{2i}), \ldots, B_K^{(l)}(x_{2i}), Z_1(x_{3i}, x_{4i}), \ldots, Z_J(x_{3i}, x_{4i}))^T$$

be the collection of all covariates and spline representations. Then, for an M-quantile q and a value for the tuning constant c, estimation of the vector of all coefficients $\boldsymbol{\gamma}_q = (\beta_{0,q}, \beta_{1,q}, \alpha_{1,q}, \alpha_{2,q}, \boldsymbol{b}_q^T, \boldsymbol{a}_q^T)^T$ can be accommodated by mimicking penalization of an objective function and solving the following set of estimating equations

$$\Psi(\boldsymbol{\gamma}_q, c) := n^{-1} \sum_{i=1}^{n} \boldsymbol{\psi}_q(y_i, Q_q(\mathbf{u}_i, t_i; c)) + \boldsymbol{\Lambda}_q \boldsymbol{P} \boldsymbol{\gamma}_q = \mathbf{0}_{(4+K+J)} \tag{14}$$

where

$$\boldsymbol{\psi}_q(y_i, Q_q(\mathbf{u}_i, t_i; c)) = \left[\psi_q(r_{iq}) w(\mathbf{u}_i) \frac{Q_q'(\mathbf{u}_i, t_i; c)}{V^{1/2}(Q_q(\mathbf{u}_i, t_i; c))} - \nu(\boldsymbol{\gamma}_q) \right] \tag{15}$$

with

- residuals $r_{iq} = V^{-1/2}[Q_q(\mathbf{u}_i, t_i; c)](y_i - Q_q(\mathbf{u}_i, t_i; c))$,
- $Q_q(\mathbf{u}_i, t_i; c) = t_i \exp\{\mathbf{u}_i^T \boldsymbol{\gamma}_q\}$,
- $V(Q_q(\mathbf{u}_i, t_i; c)) = Q_q(\mathbf{u}_i, t_i; c) + Q_q(\mathbf{u}_i, t_i; c)^2/\theta_q, \theta_q > 0$ the shape parameter,
- $Q_q'(\mathbf{u}_i, t_i; c) = Q_q(\mathbf{u}_i, t_i; c)\mathbf{u}_i$,
- $\boldsymbol{\Lambda}_q = \mathrm{blockdiag}\{\mathbf{0}_{4\times 4}, \lambda_{1,q} \boldsymbol{I}_{K\times K}, \lambda_{2,q} \boldsymbol{I}_{J\times J}\}$, and
- $\boldsymbol{P} = \mathrm{blockdiag}\{\mathbf{0}_{4\times 4}, \boldsymbol{P}_1, \boldsymbol{P}_2\}$.

Matrix $\boldsymbol{P}$ summarizes all the penalties of the procedure, i.e. no penalization for the linear/parametric components of the model and different penalties for B-splines and thin plate splines. Matrix $\boldsymbol{\Lambda}_q$ collects the corresponding smoothing parameters, with larger values of the smoothing parameters $\lambda_{1,q}$ and $\lambda_{2,q}$ corresponding to smoother approximations. Selection of good smoothing

parameters is crucial for adequate approximation of the final model and will be discussed in Section 3.3.

The correction for ensuring Fisher consistency can be computed explicitly and is given by

$$\begin{aligned} v(\boldsymbol{\gamma}_q) = n^{-1} \sum_{i=1}^{n} w_q(r_{iq}) w(\mathbf{u}_i) \Big\{ & -c\ P(Y_i \le j_1) + c\, P(Y_i \ge j_2 + 1) \\ & + \frac{Q_q(\mathbf{u}_i, t_i; c)}{V^{1/2}(Q_q(\mathbf{u}_i, t_i; c))} P(Y_i = j_1) \left(1 + \frac{j_1}{\theta_q}\right) \\ & - \frac{Q_q(\mathbf{u}_i, t_i; c)}{V^{1/2}(Q_q(\mathbf{u}_i, t_i; c))} P(Y_i = j_2) \left(1 + \frac{j_2}{\theta_q}\right) \Big\} V^{-1/2}(Q_q(\mathbf{u}_i, t_i; c)) Q_q(\mathbf{u}_i, t_i; c) \mathbf{u}_i \end{aligned}$$

where $j_1 = \lfloor Q_q(\mathbf{u}_i, t_i; c) - c V^{1/2}(Q_q(\mathbf{u}_i, t_i; c)) \rfloor, j_2 = \lfloor Q_q(\mathbf{u}_i, t_i; c) + c V^{1/2}(Q_q(\mathbf{u}_i, t_i; c)) \rfloor$, and $w_q(r_{iq}) = 2\,[q\, I(r_{iq} > 0) + (1 - q) I(r_{iq} \le 0)]$.

Equation (14) can be solved using Fisher scoring by first computing $-E[\partial \Psi(\boldsymbol{\gamma}_q, c)/\partial \boldsymbol{\gamma}_q]$. This quantity is given by

$$-E\left[\frac{\partial \Psi(\boldsymbol{\gamma}_q, c)}{\partial \boldsymbol{\gamma}_q}\right] = -n^{-1} \sum_{i=1}^{n} E\left[\frac{\partial}{\partial \boldsymbol{\gamma}_q} \boldsymbol{\psi}_q(y_i, Q_q(\mathbf{u}_i, t_i; c))\right] - \boldsymbol{\Lambda}_q \boldsymbol{P} \tag{16}$$

By adapting the computation developed for NB M-quantile models in Chambers et al.[3] we get

$$\begin{aligned} & n^{-1} \sum_{i=1}^{n} w_i E\left[\psi_q(r_{iq}) \frac{\partial \log f(Y_i | \mathbf{u}_i; \boldsymbol{\gamma}_q)}{\partial Q_q(\mathbf{u}_i, t_i; c)} \frac{\partial Q_q(\mathbf{u}_i, t_i; c)}{\partial \boldsymbol{\gamma}_q} \mathbf{u}_i\right] + \boldsymbol{\Lambda}_q \boldsymbol{P} \\ & = n^{-1} \sum_{i=1}^{n} w_i E\left[\psi_q(r_{iq}) \frac{y_i - Q_q(\mathbf{u}_i, t_i; c)}{V(Q_q(\mathbf{u}_i, t_i; c))}\right] Q_q(\mathbf{u}_i, t_i; c) \mathbf{u}_i \mathbf{u}_i^T + \boldsymbol{\Lambda}_q \boldsymbol{P} \end{aligned}$$

By using a Fisher scoring algorithm we can obtain estimates of $\boldsymbol{\gamma}_q$ for a fixed value of the smoothing parameters matrix $\boldsymbol{\Lambda}_q$ and shape parameter θ_q using

$$\boldsymbol{\gamma}_q^{(t)} = \boldsymbol{\gamma}_q^{(t-1)} + [\mathbf{U}^T \mathbf{W} \mathbf{U} + \boldsymbol{\Lambda}_q \boldsymbol{P}]^{-1} [\mathbf{U}^T \boldsymbol{\omega} + \boldsymbol{\Lambda}_q \boldsymbol{P} \boldsymbol{\gamma}_q^{(t-1)}] \tag{17}$$

where $\mathbf{U}$ is the design matrix with $\mathbf{u}_i$ on its i-th row,

$$\mathbf{W} = \operatorname{diag}\left\{ w(\mathbf{u}_i) E\left[\psi_q(r_{iq}) \frac{y_i - Q_q(\mathbf{u}_i, t_i; c)}{V(Q_q(\mathbf{u}_i, t_i; c))}\right] Q_q(\mathbf{u}_i, t_i; c) \right\}_{i=1,\ldots,n}$$

and $\boldsymbol{\omega}_{(n \times 1)} = \{w(\mathbf{u}_i)(\psi_q(r_{iq}) - E[\psi_q(r_{iq})])\}_{i=1,\ldots,n}$

The shape parameter θ_q is estimated for a fixed value of the smoothing parameters matrix $\boldsymbol{\Lambda}_q$ and of $\boldsymbol{\gamma}_q$ following Chambers et al.[3] as the solution to

$$n^{-1} \sum_{i=1}^{n} \left\{ \psi_q^2(r_{iq}) - E\left[\psi_q^2\left(\frac{Y_i - Q_q(\mathbf{u}_i, t_i; c)}{V^{1/2}(Q_q(\mathbf{u}_i, t_i; c))}\right)\right] \right\} = \mathbf{0} \tag{18}$$

where $E\left[\psi_q^2\left(\frac{Y_i - Q_q(\mathbf{u}_i, t_i; c)}{V^{1/2}(Q_q(\mathbf{u}_i, t_i; c))}\right)\right]$ is a constant that ensures Fisher consistency for estimation of θ_q and ψ_q is given in equation (15).

3.3 Smoothing parameters selection

Up to now we have illustrated estimation of $\boldsymbol{\gamma}_q$ and θ_q for a fixed value of the smoothing parameter matrix $\boldsymbol{\Lambda}_q$. Let $\widehat{\boldsymbol{\gamma}}_q$ and $\hat{\theta}_q$ be such estimates and let

$$\hat{y}_i = \hat{Q}_q(\mathbf{u}_i, t_i; c) = t_i \exp\{\mathbf{u}_i^T \hat{\boldsymbol{\gamma}}_q\} \tag{19}$$

Here, we develop a Generalized Cross Validation (GCV) criterion to be used to determine the value of the smoothing parameters $\lambda_{1,q}$ and $\lambda_{2,q}$ collected in $\boldsymbol{\Lambda}_q$. GCV depends on the effective number of degrees of freedom used by the model and these, in turn, can be usually computed from the smoother matrix. In our model, however, a fitted value $\hat{y}_i$ is a nonlinear function of the vector of y_i; therefore, there is no matrix $\mathbf{S}_{\boldsymbol{\Lambda}q}$ such that $\hat{y}_i = \mathbf{S}_{\boldsymbol{\Lambda}_q} y_i$ and hence the usual definition of a smoother matrix does not apply. However, we will define the smoother matrix $\mathbf{S}_{\boldsymbol{\Lambda}q}$ at quantile q to be that matrix such that

$$\hat{Q}_q(\mathbf{u}_i, t_i; c) - Q_q(\mathbf{u}_i, t_i; c) \approx \mathbf{S}_{\boldsymbol{\Lambda}_q}(y_i - E[Y_i])$$

Using a Taylor approximation, we can write

$$\hat{Q}_q(\mathbf{u}_i, t_i; c) - Q_q(\mathbf{u}_i, t_i; c) \approx \mathbf{WU}[\mathbf{U}^T\mathbf{WU} + \boldsymbol{\Lambda}_q \boldsymbol{P}]^{-1}\mathbf{U}^T(y_i - E[Y_i])$$

and have a definition for the smoother matrix to be

$$\mathbf{S}_{\boldsymbol{\Lambda}_q} = \mathbf{WU}[\mathbf{U}^T\mathbf{WU} + \boldsymbol{\Lambda}_q \boldsymbol{P}]^{-1}\mathbf{U}^T \tag{20}$$

Thus, an estimate of the effective number of degrees of freedom used by the model is given by its trace. The GCV criterion can then be defined as

$$GCV(\boldsymbol{\Lambda}_q) = \frac{n^{-1}\sum_{i=1}^{n} 2\mathrm{Dev}(y_i, \hat{y}_i; \boldsymbol{\Lambda}_q)}{[1 - n^{-1}\delta \mathrm{Dev}(\mathbf{S}_{\boldsymbol{\Lambda}_q})]^2} \tag{21}$$

where

$$\mathrm{Dev}(y_i, \hat{y}_i; \boldsymbol{\Lambda}_q) = \begin{cases} \hat{\theta}_q \log\left(\dfrac{\hat{y}_i}{\hat{\theta}_q} + 1\right) & \text{if } y_i = 0 \\ (\hat{\theta}_q + 1)\log\left(\dfrac{\hat{y}_i}{\hat{\theta}_q} + 1\right) + y_i \log\dfrac{y_i}{\hat{y}_i} - (y_i + \hat{\theta}_q)\log\left(\dfrac{y_i}{\hat{\theta}_q} + 1\right) & \text{if } y_i = 1, 2, \ldots \end{cases}$$

is the i-th contribution to the deviance of the model for a fixed value of the smoothing parameter matrix $\boldsymbol{\Lambda}_q$, and δ is a constant that penalizes additional effective degrees of freedom measured by the trace of the smoother matrix.

The complete algorithm for the estimation of γ_q, smoothing parameters in $\mathbf{\Lambda}_q$ and the shape parameter θ_q can be described as follows:

(1) For each value of q and a fixed c, choose the initial values $\gamma_q^{(0)}$, $\mathbf{\Lambda}_q^{(0)}$, and $\theta_q^{(0)}$. Set $b = 0$.
(2a) calculate $\gamma_q^{(b+1)}$ from the Fisher scoring in (17);
(2b) calculate $\mathbf{\Lambda}_q$ from formula (21);
(2c) calculate $\theta_q^{(b+1)}$ from criterion (18);
(2d) set b = b + 1.
(3) Iterate steps (2a)–(2d) until convergence is achieved.

Routines in R to this end are available from the authors, and we refer to the 'ensemble' model for a range of values of q as a SPNBMQ model in what follows.

3.4 Disease mapping via SPNBMQ

Usually, models characterize the variability associated with the conditional distribution of an overdispersed count variable Y given covariates $\boldsymbol{x}$ in terms of latent clustering and heterogeneity effects. On the other hand, M-quantile models can be used to characterize overdispersion in a different way by attaching to each observed count a so-called 'M-quantile coefficient'. The M-quantile coefficient associated with the observed value y_i of a continuously distributed random variable Y and an associated covariate value $\mathbf{x}_i$ is the value q_i such that $\hat{Q}_{q_i}(\mathbf{x}_i; c) = y_i$,[29,33] i.e. that value of q for which the fitted value reproduces the observed one. Typically, this equation is solved by fitting regression M-quantiles on a finite grid $G = \{0 < q_1 < q_2 < \ldots < q_{L-1} < q_L < 1\}$ of L values of q and then using linear interpolation.

When dealing with NB data, however, the proposed model is such that $Q_q(\mathbf{u}_i, t_i; c) = t_i \exp\{\mathbf{u}_i^T \gamma_q\}$ and an M-quantile coefficient cannot be found with the aforementioned approach for those observations for which $y_i = 0$, since $Q_q(\mathbf{u}_i, t_i; c)$ has a strictly positive domain. To overcome this problem we use the definition proposed in Chambers et al.[3]

$$\hat{Q}_q(\mathbf{u}_i, t_i; c) = \begin{cases} k(\mathbf{u}_i) & y_i = 0 \\ y_i & y_i = 1, 2, \ldots \end{cases} \tag{22}$$

where $k(\cdot)$ denotes an appropriate strictly positive boundary function for the data set. In particular, since we can argue that observation $y_i = 0$ should have a smaller q-value then observation $y_j = 0$ when $\hat{Q}_{0.5}(\mathbf{u}_i, t_i; c) > \hat{Q}_{0.5}(\mathbf{u}_j, t_j; c)$, we can set $k(\mathbf{u}_i) = \min\{1 - , \epsilon[\hat{Q}_{0.5}(\mathbf{u}_i, t_i; c)]^{-1}\}$ in equation (22), where $\epsilon > 0$ is a small positive constant.

M-quantile coefficients allow for a better representation of the conditional distribution of Y given the covariates and can be used to capture residual between-area variation. In particular, we can use the deviation of the area-specific M-quantile coefficient $\hat{\gamma}_{q_i}$ from the 'median' M-quantile coefficient $\hat{\gamma}_{0.5}$. In particular, the SPNBMQ predictor of the count in area i is then given by

$$\hat{Q}_{q_i}(\mathbf{u}_i, t_i; c) = t_i \exp\left(\mathbf{u}_i^T \hat{\gamma}_{q_i}\right) = t_i \exp\left(\mathbf{u}_i^T \hat{\gamma}_{0.5} + \mathbf{u}_i^T \left(\hat{\gamma}_{q_i} - \hat{\gamma}_{0.5}\right)\right) \tag{23}$$

where the last term on the right-hand side can be interpreted as a pseudo-random effect for area i, allowing also for estimation of area effects.

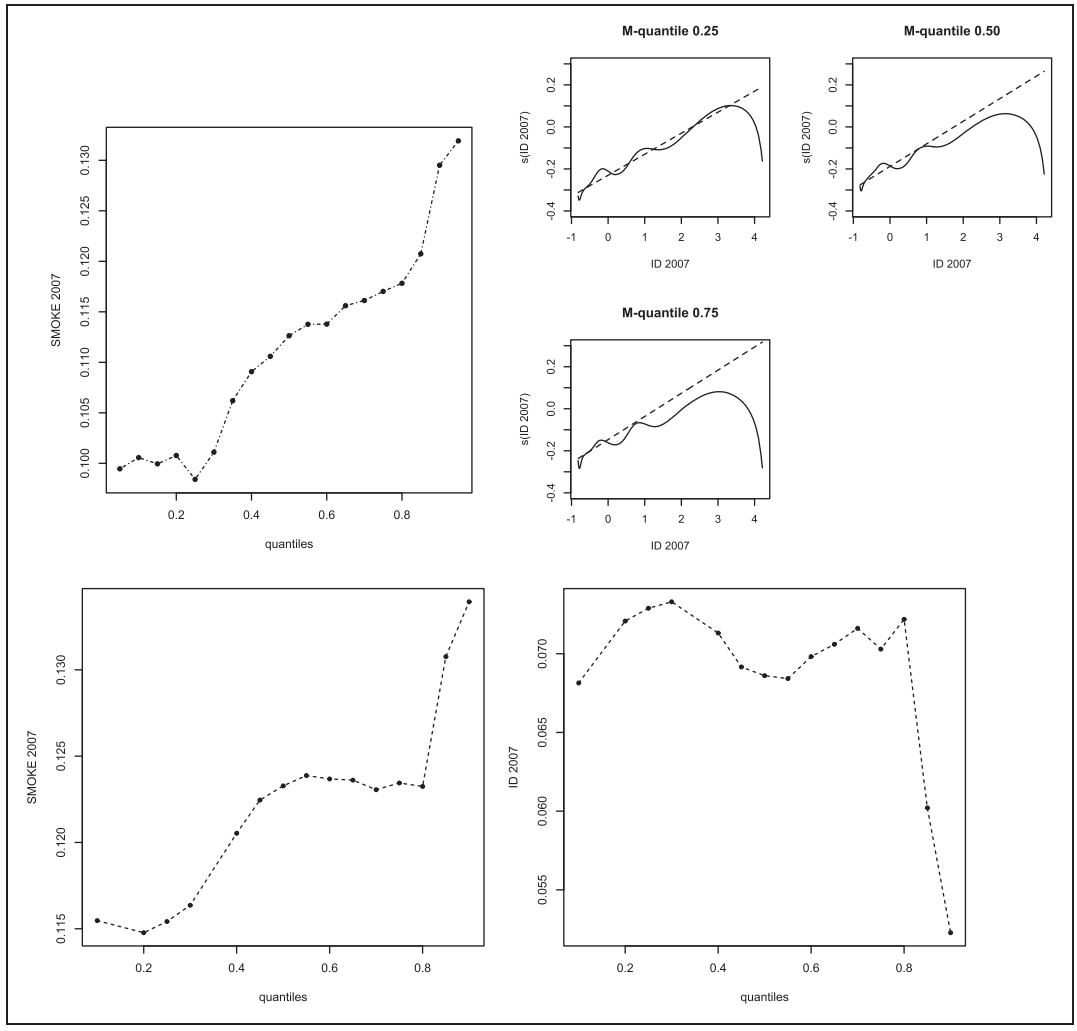

Figure 5. M-quantile coefficient for different quantiles for SMOKE 2007 (top left) and (solid line) the estimated *P*-splines on ID 2007 for different quantiles $q = 0.25$, 0.50, and 0.75 from SPNBMQ(c) model and (dashed line) the corresponding coefficient for ID 2007 from SPNBMQ(d) (top right). Bottom panel M-quantile coefficient for different quantiles for SMOKE 2007 (left) and ID 2007 (right) from SPNBMQ(d).

4 Results

In this section, we illustrate the main results of the SPNBMQ approach using the data on the lung cancer incidence over 2005–2010 for 326 LADs in England. The diagnostic analysis in Section 1 allows us to detect the presence of large residuals and it motivates the recourse to robust methods such as an M-quantile model with a bounded ψ function. In particular, we have fitted to the data the following models: (1) SPNBMQ(c), log(relative risk) = intercept + SMOKE 2007 + s(ID 2007) + s(lon,lat) and (2) the SPNBMQ(d) log(relative risk) = intercept + SMOKE 2007 + ID 2007 + s(lon,lat).

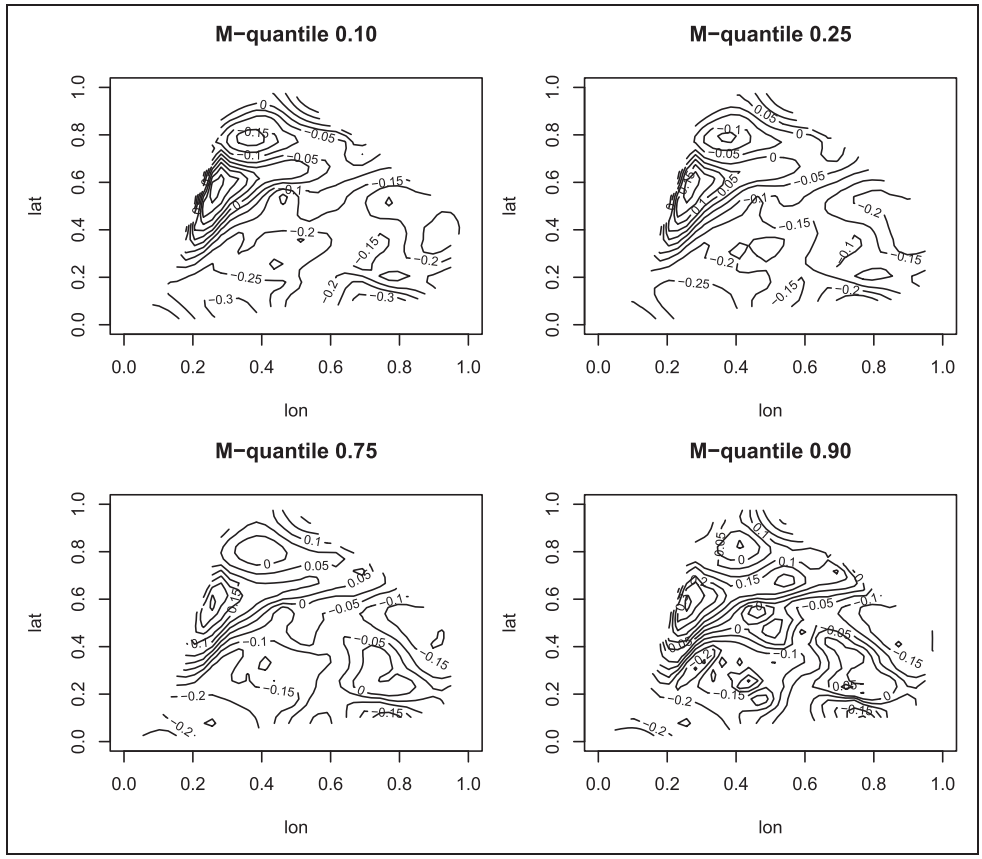

Figure 6. Contour plot of estimated bivariate spline for different quantiles $q = 0.10$, 0.25, 0.75, and 0.90 from SPNBMQ(d) model.

The values of Pearson residuals in Figure 4 suggest to use a Huber proposal 2 with the tuning constant c equal to 2. SPNBMQ(c) has been fitted by using cubic B-splines ($l = 3$) with $K = 30$ knots for ID 2007 with a difference matrix of order $d = 3$. For the bivariate smooth component of both SPNBMQ(c) and SPNBMQ(d) models, a set of radial basis functions (as specified in Section 3.1) with $J = 90$ knots, has been applied. The Demmler–Reinsch orthogonalization has been performed to ensure the numerical method convergence. The value of δ in the GCV criterion is fixed to 1.2 for all quantiles. The estimated value of θ_q is large (>100) for each quantile, by this showing that overdispersion is not present in the data, similarly to the NBGAM models.

Comparison between the two models could be carried on by using the GCV criterion in equation (21). The GCV criterion takes value 2.48 for model SPNBMQ(c) and value 2.05 for model SPNBMQ(d) by this suggesting that, when using a roust model, the nonlinear relationship with ID 2007 does not provide an improvement in the fit. The use of GCV for model comparison needs caution here because θ is not a maximum likelihood estimator and the deviance depends on its value. Nonetheless, it is a useful way to compare SPNBMQ models although further research is needed to develop other criteria possibly based on quasi-likelihoods.[34]

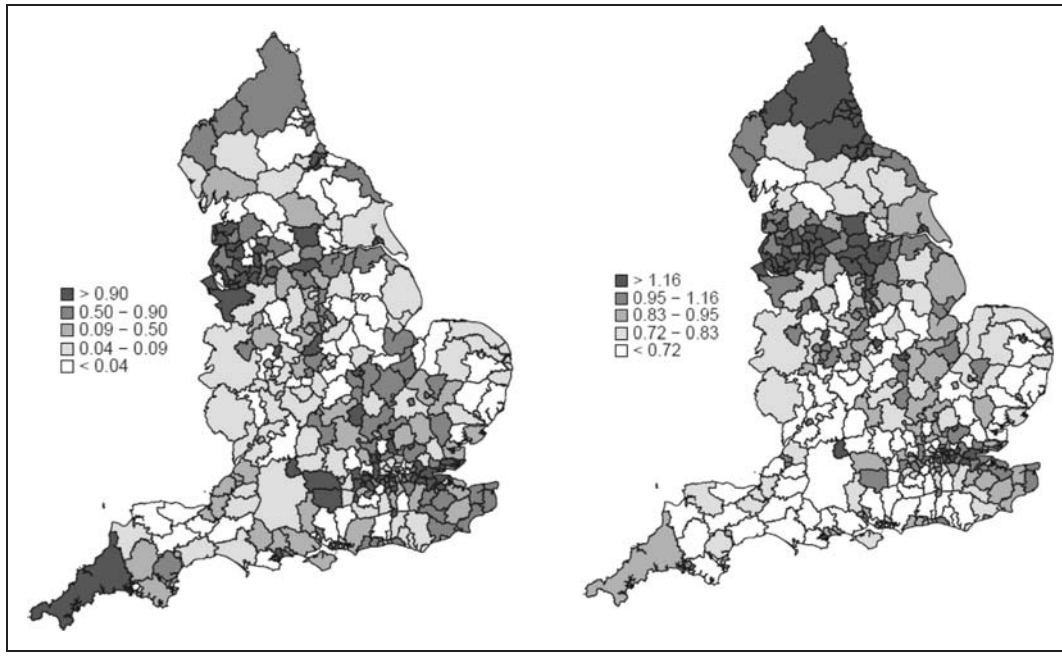

Figure 7. Maps of M-quantile coefficients q_i (right) and estimated relative risks (left) from SPNBMQ(d) model.

Figure 5 (left panel) exhibits the change in the coefficient $\beta_{1,q}$ for SMOKE 2007 on the SPNBMQ(c) (top) and SPNBMQ(d) (bottom) models as the quantile index q ranges from zero to one. The path seems to be rather linear and shows a slight increase in the value of $\beta_{1,q}$ with q. The value at the median is similar to that coming (0.12) from the NBGAM model (c) considered in the Introduction. Figure 5 (top right panel) shows, on the other hand, the estimated *P*-spline for ID 2007 at the quartiles and, in dashed line, the corresponding linear approximation from the SPNBMQ(d) model. It is clear that there is a different approximation for large values of ID 2007. In particular, by looking at the right panel of Figure 1, we can see that the *P*-splines try to approximate the pattern given by the last few LADs that show large values of ID 2007 but small relative risks. This suggests that estimates obtained by SPNBMQ(d) model for the areas with larger values of ID 2007 will be all moved upward compared to those obtained from the SPNBMQ(c) model. In fact, the estimated coefficient for ID 2007 from model SPNBMQ(d) at quantile 0.5 is sensibly larger (0.069) than that from model NBGAM(c) (0.058), thereby showing the efficacy of the bounded influence function on adjusting the role of outlying areas on estimating the effect of ID 2007 on SMR. Convergence issues for the SPNBMQ estimation algorithm for values of q near zero and near one mean that these estimated coefficients behave in a rather nonlinear fashion in the tails of the q distribution, as is particularly evident for ID 2007 in Figure 5 for SPNBMQ(d) model. These estimated tail coefficients should therefore be treated with caution.

The reliability of the results depends not only on the availability of the auxiliary information, but also on the choice of the smoothing parameters which contribute to the bivariate spline effect at each quantile. The latter evaluated at different quantiles shows that the contribution seems to be able to mimic the spatial trend of the study variable. Figure 6 describes the contour of the estimated bivariate *P*-splines at different quantiles ($q = 0.10$, 0.25, 0.75, and 0.90) from model

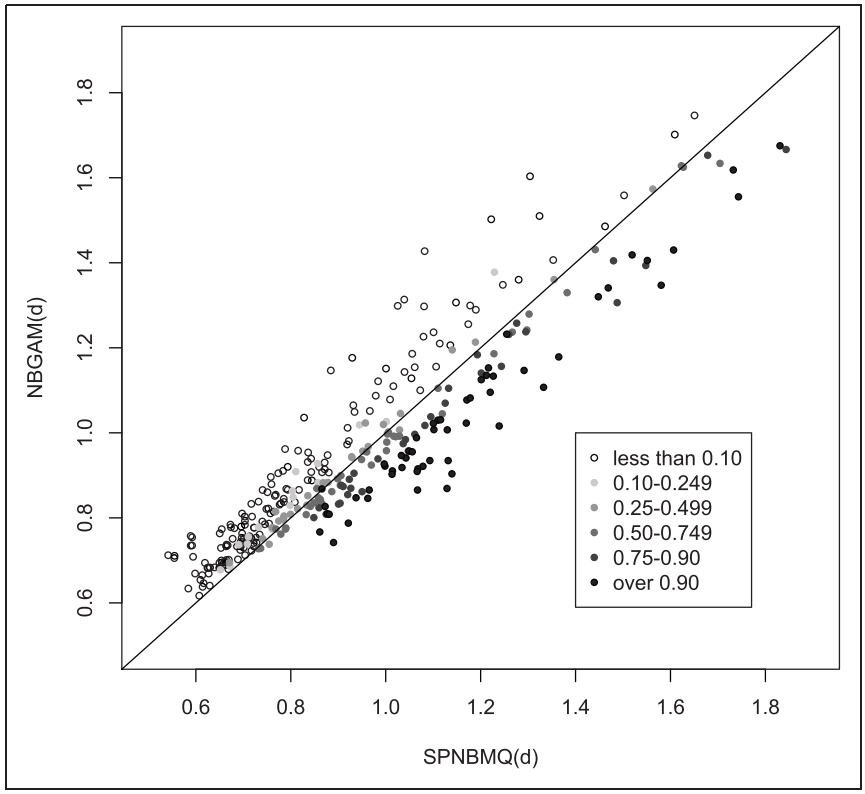

Figure 8. Estimated relative risk from NBGAM model (d) and model SPNBMQ(d). Estimates are color-coded according to estimated q-values.

SPNBMQ(d). The estimated spline effect at $q = 0.90$ takes the highest values in higher risk areas as, for example, for those areas near London, while for the highly industrialized zones of northern England, as Liverpool and Manchester, we have relatively lower values for $q = 0.10$. We observe negative values for all quantiles for the South-East region where low relative risks are estimated.

Generalized Linear Mixed Models include random area effects to account for between-area variation. The M-quantile approach avoids distributional assumptions as well as problems associated with the specification of random effects, allowing between-area differences to be characterized by the variation of area-specific M-quantile coefficients. Figure 7 (left panel) reports the spatial distribution of the M-quantile coefficients, the q_i's, which reflects variability not explained by the covariates. A similar reasoning in a completely different context can be found in Kokic et al.,[29] where M-quantile regression is employed to measure production performance and units are ranked according to their M-quantile coefficient after accounting for covariates. The q_i values for Hackney, Newham and Tower Hamlets from model SPNBMQ(c) are respectively 0.25, 0.20, and 0.74 and from model SPNBMQ(d) 0.03, 0.01, and 0.03. Once the effects for ID 2007 are linear, the q_i's decrease drastically for the three areas. This means that the values of the three areas are in the tail of the distribution and then they can be considered as outlier values under the SPNBMQ(d) model. This property of the M-quantile approach is evident also from Figure 8 that compares

estimated relative risks from SPNBMQ(d) model and those obtained using NBGAM model (d) (for a discussion of *P*-splines models for disease mapping see Goicoa et al.[17]). From this figure it is clear that the two sets of estimates are strongly correlated, but that those based on the proposed method add to those based on the NBGAM model (d) the information coming from M-quantile coefficients: LADs characterized by a relatively smaller (larger) q_i tend to have overestimated (underestimated) relative risks by NBGAM that relies on the model only.

Figure 7 (right panel) shows the relative risk estimated from the SPNBMQ(d) model. This map points out the expected geographic differences, due to variation in SMOKE 2007 and the two smoothing components. The incidence of lung cancer is generally lower than average in the south and midlands of England, with the exception of some health authorities in London and the Birmingham conurbation. There is a band of higher than average relative risk across the formerly highly industrialized parts of northern England, from Liverpool and Manchester in the west to Hull in the east. Higher values have been observed also in Teesside and Tyneside.

5 Conclusions

A semiparametric NB M-quantile model for ecological regression on disease mapping is proposed and is used to obtain relative risk estimates. The novelty of the paper consists on the inclusion of smoothing components on the quantile predictor of an NB M-quantile regression. In particular, these components can represent spatial effects and the nonlinear relation between covariates and the response variable. This approach reduces the need for parametric assumptions, so that model misspecification problems are reduced. These smoothing components have been estimated using penalized splines.

The results from the application to data on Lung cancer incidence over 2005–2010 for 326 LADs in England show that the proposed methodology provides a reasonable and useful alternative to existing methods when assumptions of the parametric model are not valid. In particular, it is a method that is robust to outlying observations and to mis-specification of the linear predictor structure.

Despite the fact that the proposed methodology provides encouraging results, further research is necessary. To start with, the estimation of the area quantile coefficients is challenging and alternative approaches should be investigated. Moreover, a drawback for all quantile-type fitted regression functions is the phenomenon of quantile crossing. This occurs when two or more estimated quantiles or M-quantile functions cross or overlap. It may be due to model misspecification, collinearity or huge outlying values. He[35] proposed a simple way to a posteriori restrict quantile regression while maintaining sufficient modelling flexibility. This method has been used by Pratesi et al.[16] for overcoming the problem of quantile crossing for *P*-splines M-quantile regression models. Hence, this approach could be easily extended to the SPNB M-quantile model. Moreover, further research is necessary in order to understand how the proposed M-quantile method, NBGAM and Empirical Bayes, Hierarchical Bayes (assuming no clustering effects and assuming clustering effects) compare in terms of their performance.

A further extension could be the development of a bootstrap-based Mean Square Error (MSE) estimator. This latter, may offer a stable approximation of the actual MSE of the estimated relative risk. For example, an MSE estimator based on the semiparametric bootstrap (proposed also in Chambers et al.[3] and Tzavidis et al.[36]) could be considered. Finally, the time dimension, as done for the spatial dimension, could be added into the proposed Semiparametric NB M-quantile model in order to provide robust estimates of the relative risk in space and time.

References

1. Wood SN. *Generalized additive models: An introduction with R*. Chapman & Hall/CRC: Boca Raton, 2006.
2. Cai B, Lawson AB, Hossain MM, et al. Bayesian semiparametric model with spatially–temporally varying coefficients selection. *Stat Med* 2013; **32**: 3670–3685.
3. Chambers R, Dreassi E and Salvati N. Disease mapping via negative binomial regression M-quantiles. arXiv:1310.3403v1 [stat.ME] 2013.
4. Lee D and Neocleous T. Bayesian quantile regression for count data with application to environmental epidemiology. *J R Stat Soc Series C* 2010; **59**: 905–920.
5. Lee D. Using spline models to estimate the varying health risks from air pollution across Scotland. *Stat Med* 2012; **31**: 3366–3378.
6. MacNab YC. Spline smoothing in Bayesian disease mapping. *Environmetrics* 2007; **18**: 727–744.
7. MacNab YC and Gustafson P. Regression B-spline smoothing in Bayesian disease mapping: with an application to patient safety surveillance. *Stat Med* 2007; **26**: 4455–4474.
8. Bernardinelli L, Pascutto C, Best NG, et al. Disease mapping with errors in covariates. *Stat Med* 1997; **16**: 741–752.
9. MacNab YC. Bayesian multivariate disease mapping and ecological regression with errors in covariates: Bayesian estimation of DALYs and 'preventable' DALYs. *Stat Med* 2009; **28**: 1369–1385.
10. MacNab YC. On Bayesian shared component disease mapping and ecological regression with errors in covariates. *Stat Med* 2010; **29**: 1239–1249.
11. Xia H and Carlin BP. Spatio-temporal models with errors in covariates: mapping Ohio lung cancer mortality. *Stat Med* 1998; **17**: 2025–2043.
12. Wakefield J. Disease mapping and spatial regression with count data. *Biostatistics* 2007; **8**: 158–183.
13. Breslow NE and Clayton DG. Approximate inference in Generalized Linear Mixed Models. *J Am Stat Soc* 1993; **88**: 9–25.
14. Latouche A, Guihenneuc-Jouyaux C, Girard C, et al. Robustness of the BYM model in absence of spatial variation in the residuals. *Int J Health Geogr* 2007; **6**: 1–39.
15. Wakefield J. Sensitivity analyses for ecological regression. *Biometrics* 2003; **59**: 9–17.
16. Pratesi M, Ranalli MG and Salvati N. Semiparametric M-quantile regression using penalized splines. *J Nonparametr Stat* 2009; **21**: 287–304.
17. Goicoa T, Ugarte MD, Etxeberria J, et al. Comparing CAR and P-spline models in spatial disease mapping. *Environ Ecol Stat* 2012; **19**: 573–599.
18. Lee D-J and Durbán M. Smooth-CAR mixed models for spatial count data. *Comput Stat Data Anal* 2009; **53**: 2968–2979.
19. MacNab YC and Dean CB. Spatio-temporal modelling of rates for the construction of disease maps. *Stat Med* 2002; **21**: 347–358.
20. Sauleau EA, Hennerfeind A, Buemi A, et al. Age, period and cohort effects in Bayesian smoothing of spatial cancer survival with geoadditive models. *Stat Med* 2007; **26**: 212–229.
21. Cameron AC and Trivedi PK. *Regression analysis of count data*. New York: Cambridge University Press, 1998.
22. Breslow NE. Extra-Poisson variation in log-linear models. *Appl Stat* 1984; **33**: 38–44.
23. Lawless JF. Negative binomial and mixed poisson regression. *Can J Stat* 1987; **15**: 209–225.
24. McCullagh P and Nelder JA. *Generalized linear models*, 2nd ed. Chapman and Hall: London – New York, 1989.
25. Cantoni E and Ronchetti E. Robust inference for generalized linear models. *J Am Stat Assoc* 2001; **96**: 1022–1030.
26. Huber PJ. *Robust statistics*. Wiley: London, 1981.
27. Breckling J and Chambers R. M-quantiles. *Biometrika* 1988; **75**: 761–771.
28. Newey WK and Powell JL. Asymmetric least squares estimation and testing. *Econometrica* 1987; **55**: 819–847.
29. Kokic P, Chambers R, Breckling J, et al. A measure of production performance. *J Bus Econ Stat* 1997; **10**: 419–435.
30. Ruppert D, Wand MP and Carroll RJ. *Semiparametric regression*. Cambridge: Cambridge University Press, 2003.
31. Ruppert D, Wand MP and Carroll RJ. Semiparametric regression during 2003–2007. *Electron J Stat* 2009; **3**: 1193–1256.
32. Eilers PHC and Marx BD. Flexible smoothing with B-splines and penalties. *Stat Sci* 1996; **11**: 89–121.
33. Chambers R and Tzavidis N. M-quantile models for small area estimation. *Biometrika* 2006; **93**: 255–268.
34. Heritier S, Cantoni E, Copt S, et al. *Robust methods*. Chichester, UK: John Wiley & Sons, 2009.
35. He X. Quantile curves without crossing. *Am Stat* 1997; **51**: 186–192.
36. Tzavidis N, Ranalli MG, Salvati N, et al. Robust small area prediction for counts. *Stat Methods Med Res* 2014; DOI: 10.1177/0962280214520731.