UNIVERSITY OF Southampton
Southampton Statistical Sciences Research Institute

Working Paper M10/12

Methodology

Small Area Estimation Via M-Quantile Geographically Weighted Regression

N. Salvati, N. Tzavidis, M. Pratesi, R. Chambers

Abstract

The effective use of spatial information, that is the geographic locations of population units, in a regression model-based approach to small area estimation is an important practical issue. One approach for incorporating such spatial information in a small area regression model is via Geographically Weighted Regression (GWR). In GWR the relationship between the outcome variable and the covariates is characterised by local rather than global parameters, where local is defined spatially. In this paper we investigate GWR-based small area estimation under the M-quantile modelling approach. In particular, we specify an M-quantile GWR model that is a local model for the M-quantiles of the conditional distribution of the outcome variable given the covariates. This model is then used to define a bias-robust predictor of the small area characteristic of interest that also accounts for spatial association in the data. An important spin-off from applying the M-quantile GWR small area model is that it can potentially offer more efficient synthetic estimation for out of sample areas. We demonstrate the usefulness of this framework through both model-based as well as design-based simulations, with the latter based on a realistic survey data set. The paper concludes with an illustrative application that focuses on estimation of average levels of Acid Neutralizing Capacity for lakes in the north-east of the USA.

Small Area Estimation Via M-quantile Geographically Weighted Regression

N. Salvati · N. Tzavidis · M. Pratesi · R. Chambers

Abstract The effective use of spatial information, that is the geographic locations of population units, in a regression model-based approach to small area estimation is an important practical issue. One approach for incorporating such spatial information in a small area regression model is via Geographically Weighted Regression (GWR). In GWR the relationship between the outcome variable and the covariates is characterised by local rather than global parameters, where local is defined spatially. In this paper we investigate GWR-based small area estimation under the M-quantile modelling approach. In particular, we specify an M-quantile GWR model that is a local model for the M-quantiles of the conditional distribution of the outcome variable given the covariates. This model is then used to define a bias-robust predictor of the small area characteristic of interest that also accounts for spatial association in the data. An important spin-off from applying the M-quantile GWR small area model is that it can potentially offer more efficient synthetic estimation for out of sample areas. We demonstrate the usefulness of this framework through both model-based as well as design-based simulations, with the latter based on a realistic survey data set. The paper concludes with an illustrative application that focuses on estimation of average levels of Acid Neutralizing Capacity for lakes in the north-east of the USA.

Keywords Borrowing strength over space · Environmental data · Estimation for out of sample areas · Robust regression · Spatial dependency.

1 Introduction

Sample survey are extensively used to collect data for calculation of reliable direct estimates of population totals and means. However, reliable estimates for domains are also often required, and geographically defined domains, for example regions, states, counties and metropolitan areas, are of particular interest. In many cases, small (or even zero) domain-specific sample sizes result in direct estimators with high variability. This problem can be avoided by employing small area estimation (SAE) techniques. An approach that is now widely used in SAE is the so-called indirect or model-based approach, and indirect estimators for small areas are typically based on unit level random effects models. In particular, the Best Linear Unbiased Predictor (BLUP) can be defined using a unit level model that assumes independence, but not necessarily normality, of the random area effects (Rao, 2003). A detailed description of this predictor and of its empirical version (EBLUP) can be found in Rao (2003, Chap. 7), Rao (2005) and Jiang & Lahiri (2006). Chambers & Tzavidis

Dipartimento di Statistica e Matematica Applicata all'Economia, Università di Pisa
Via C. Ridolfi, 10 - Pisa
Tel.: +39-50-2216492
Fax: +39-50-2216375
E-mail: salvati@ec.unipi.it

Social Statistics and Southampton Statistical Sciences Research Institute, University of Southampton E-mail: n.tzavidis@soton.ac.uk · Dipartimento di Statistica e Matematica Applicata all'Economia, Università di Pisa E-mail: m.pratesi@ec.unipi.it · Centre for Statistical and Survey Methodology, School of Mathematics and Applied Statistics, University of Wollongong E-mail: ray@uow.edu.au

(2006) describe an alternative approach to SAE that is based on regression M-quantiles. This 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. Nevertheless, the assumption of unit level independence is also implicit in M-quantile SAE models.

In economic, environmental and epidemiological applications, observations that are spatially close may be more alike than observations that are further apart. One approach to incorporating such spatial information in statistical modelling is by extending the random effects model to allow for spatially correlated area effects using, for example, a Simultaneous Autoregressive (SAR) model (Anselin, 1992; Cressie, 1993). Applications of SAR models in small area estimation have been considered by Petrucci & Salvati (2004), Singh et al. (2005) and Pratesi & Salvati (2008). An alternative approach for incorporating spatial information in a small area regression model is to assume that the model coefficients themselves vary spatially across the geography of interest. Geographically Weighted Regression (GWR) (Brundson et al., 1996; Fotheringham et al., 1997, 2002; Yu & Wu, 2004) models this spatial variation by using local rather than global parameters in the regression model. That is, a GWR model assumes spatial non-stationarity of the conditional mean of the variable of interest.

In this paper we explore the use of GWR in small area estimation based on the M-quantile modelling approach. In particular, we propose an M-quantile GWR model, i.e. a local model for the M-quantiles of the conditional distribution of the outcome variable given the covariates. This approach is semi-parametric in that it attempts to capture spatial variability by allowing model parameters to change with the location of the units, in effect by using a distance metric to introduce spatial non-stationarity into the mean structure of the model. The model is then used to define a predictor of the small area characteristic of interest (here we focus on small area means). As a consequence, the M-quantile GWR small area model integrates the concepts of bias-robust small area estimation and borrowing strength over space within a unified modeling framework. By construction, the M-quantile GWR model is a local model and so can provide more flexibility in SAE, particularly for out of sample small area estimation, i.e. areas where there are no sampled units. Empirical results presented in this paper indicate that use of this model for SAE appears to lead to more efficient predictors for this situation. However, this extra flexibility comes at the cost of having to estimate more parameters than in a global model. Model diagnostics play a crucial role in applied SAE and some approaches to deciding whether to use a local or a global small area model are discussed in this paper.

The structure of the paper is as follows. In Section 2 we review unit level mixed models with random area effects and M-quantile models for small area estimation. In Section 3 we describe GWR and extend this procedure to define the M-quantile GWR model. In Section 4 we describe SAE under the M-quantile GWR model, and in Section 5 we discuss mean squared error estimation for small area predictors based on this model. In Section 6 we present results from model-based and design-based simulation studies aimed at assessing the performance of the different small area predictors considered in this paper. In Section 7 we use the M-quantile GWR small area model for estimating average levels of Acid Neutralizing Capacity at 8-digit Hydrologic Unit Code (HUC) level using data collected in an environmental survey of lakes in the north-eastern region of the USA. Note that Opsomer et al. (2008) and Pratesi et al. (2008) have also applied non-parametric spatial SAE methods to these data. Although these studies employ global, rather than local, non-parametric spatial small area models, we find that SAE based on an M-quantile GWR model leads to qualitatively similar results. Finally, in Section 8 we summarize our main findings and provide directions for future research.

2 An overview of unit level models for small area estimation

In what follows we assume that the target population can be divided into d small areas and that unit record data are available at small area level. We index the population units by i and the small areas by j. Each small area j contains a known number N_j of units. The set s_j contains the n_j population indices of the sampled units in small area j. Note that non-sample areas have $n_j = 0$,

in which case s_j is the empty set. The set r_j contains the $N_j - n_j$ indices of the non-sampled units in small area j. The overall population size is N and the overall sample size is n. The sample data then consist of indicators of small area affiliation, values y_i of the variable of interest, values $\mathbf{x}_i$ of a vector of p auxiliary variables that characterise between unit variability and values $\mathbf{z}_i$ of a vector of k covariates that characterise between area variability. We assume that $\mathbf{x}_i$ contains 1 as its first component. The aim is to use this data to predict various area specific quantities, including (but not only) the area j mean m_j of y.

The most popular method used for this purpose employs linear mixed models. In the general case such a model has the form

$$y_i = \mathbf{x}_i^T\boldsymbol{\beta} + \mathbf{z}_i^T\boldsymbol{\gamma} + \epsilon_i, \tag{1}$$

where $\boldsymbol{\gamma}$ is a vector of k area-specific random effects, ϵ_i is an individual random effect. The empirical best linear unbiased predictor (EBLUP) of m_j (Henderson, 1975; Rao, 2003, Chap. 7) is then

$$\hat{m}_j^{LM} = N_j^{-1}\Big[\sum_{i\in s_j} y_i + \sum_{i\in r_j}\{\mathbf{x}_i^T\hat{\boldsymbol{\beta}} + \mathbf{z}_i^T\hat{\boldsymbol{\gamma}}\}\Big], \tag{2}$$

where $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\gamma}}$ are defined by substituting an optimal estimator for the covariance matrix of the random effects in (1) in the best linear unbiased estimator of $\boldsymbol{\beta}$ and the best linear unbiased predictor (BLUP) of $\boldsymbol{\gamma}$ respectively. A widely used estimator of the mean squared error (MSE) of the EBLUP is based on the approach of Prasad & Rao (1990). This estimator accounts for the variability due to the estimation of the random effects, regression parameters and variance components.

An alternative approach to small area estimation is based on the use of M-quantile models. The M-quantile of order q of a random variable Y with distribution function $F(Y)$ is the value Q_q that satisfies

$$\int \psi_q\Big(\frac{Y - Q_q}{\sigma_q}\Big)dF(Y) = 0,$$

where $\psi_q(\varepsilon) = 2\{qI(\varepsilon > 0) + (1-q)I(\varepsilon \leqslant 0)\}\psi(\varepsilon)$ and ψ is an appropriately chosen influence function. Here σ_q is a suitable measure of the scale of the random variable $Y - Q_q$. Note that when $\psi(\varepsilon) = \varepsilon$ we obtain the expectile of order q, which represents a quantile-like generalization of the mean, and when $\psi(\varepsilon) = sgn(\varepsilon)$ we obtain the standard quantile of order q. Both quantiles and expectiles have been extended to conditional distributions to provide quantile and expectile generalizations of the usual concept of a regression model (Koenker & Bassett, 1978; Newey & Powell, 1987). More generally, Brecking & Chambers (1988) define a linear M-quantile regression model as one where the M-quantile $Q_q(\mathbf{X};\psi)$ of order q of the conditional distribution of y given $\mathbf{X}$ corresponding to an influence function ψ satisfies

$$Q_q(\mathbf{x}_i;\psi) = \mathbf{x}_i^T\boldsymbol{\beta}_\psi(q). \tag{3}$$

For specified q and continuous ψ, an estimate $\hat{\boldsymbol{\beta}}_\psi(q)$ of $\boldsymbol{\beta}_\psi(q)$ can be obtained via iterative weighted least squares. Asymptotic theory for this estimator follows directly from well-known M-estimation results and is set out in Section 2.2 of Brecking & Chambers (1988). The M-quantile coefficient q_i of population unit i was introduced by Kokic et al. (1997) and is defined as the value q_i such that $Q_{q_i}(\mathbf{x}_i;\psi) = y_i$. M-quantile regression models can be used to characterise the entire conditional distribution $f(y|\mathbf{X})$ of y given $\mathbf{X}$, with the M-quantile coefficients, q_i then characterising unit level differences in this conditional distribution.

Extending this line of thinking to SAE, Chambers & Tzavidis (2006) observed that if variability between the small areas is a significant part of the overall variability of the population data, then units from the same small area are expected to have similar M-quantile coefficients. In particular, when (3) holds, and $\boldsymbol{\beta}_\psi(q)$ is a sufficiently smooth function of q, these authors suggest a predictor of m_j of the form

$$\hat{m}_j^{MQ} = N_j^{-1}\Big[\sum_{i\in s_j} y_i + \sum_{i\in r_j}\hat{Q}_{\hat{\theta}_j}(\mathbf{x}_i;\psi)\Big], \tag{4}$$

where $\hat{Q}_{\hat{\theta}_j}(\mathbf{x}_i;\psi) = \mathbf{x}_i^T\hat{\boldsymbol{\beta}}_\psi(\hat{\theta}_j)$ and $\hat{\theta}_j$ is an estimate of the average value of the M-quantile coefficients of the units in area j. Typically this is the average of estimates of these coefficients for sample units in the area, where these unit level coefficients are estimated by solving $\hat{Q}_{q_i}(\mathbf{x}_i;\psi) = y_i$ for q_i. Here $\hat{Q}_q$ denotes the estimated value of (3) at q. When there is no sample in area, we can form a 'synthetic' M-quantile predictor by setting $\hat{\theta}_j = 0.5$.

Tzavidis et al. (2010) refer to (4) as the 'naïve' M-quantile predictor and note that this can be biased. When the non-sample predicted values in (4) are estimated expectations $\hat{y}_i$ that converge in probability to the actual expected values of the y_i, we see that

$$\sum_{i\in r_j} I(\hat{y}_i \leqslant t) = \sum_{i\in r_j} I\{y_i - (y_i - \hat{y}_i) \leqslant t\} \approx \sum_{i\in r_j} I\{y_i \leqslant t + \varepsilon_i\} \neq \sum_{i\in r_j} I\{y_i \leqslant t\}.$$

Here $\varepsilon_i = y_i - \hat{y}_i$ is the actual regression error. If these errors are independently and identically distributed symmetrically about zero, we expect that the summation on the left hand side above will closely approximate the summation on the right for values of t near the mean/median of the non-sampled area j values of y but not anywhere else. More generally, for heteroskedastic and/or asymmetric errors, this correspondence will typically occur elsewhere in the support of y, although one would expect that in most reasonable situations it will be 'close' to the mean/median of y.

To rectify this problem these authors propose a bias adjusted M-quantile predictor of m_j of the form

$$\hat{m}_j^{MQ/CD} = \int_{-\infty}^{+\infty} t d\hat{F}_j(t) = N_j^{-1}\Big[\sum_{i\in U_j}\hat{Q}_{\hat{\theta}_j}(\mathbf{x}_i;\psi) + \frac{N_j}{n_j}\sum_{i\in s_j}\{y_i - \hat{Q}_{\hat{\theta}_j}(\mathbf{x}_i;\psi)\}\Big], \tag{5}$$

where $U_j = s_j \cup r_j$. Note that the superscript CD in (5) refers to the fact that it is the value of the expected value functional defined by the area j version of the distribution function estimator proposed by Chambers & Dunstan (1986). Due to the bias correction in (5), this predictor will have higher variability and so will be most effective when the naïve estimator (4) is expected to have substantial bias, e.g. when (3) is incorrectly specified. An alternative approach for dealing with the bias-variance trade off implicit in (5) is discussed by Tzavidis et al. (2010), and involves the use of robust (huberized) residuals instead of raw residuals in this bias correction term. Finally, these authors also note that under simple random sampling within the small areas, (5) can also be derived from the design-consistent estimator of the finite population distribution function proposed by Rao et al. (1990).

Following the approach by Chambers & Tzavidis (2006), an analytic estimator of the mean squared error of (5) is described by Tzavidis et al. (2010). This is of the form

$$mse(\hat{m}_j^{MQ/CD}) = \frac{1}{N_j^2}\sum_{k:n_k>0}\sum_{i\in s_k}\lambda_{ijk}\Big\{y_i - \hat{Q}_{\hat{\theta}_k}(\mathbf{x}_i;\psi)\Big\}^2, \tag{6}$$

where $\lambda_{ijk} = \{(w_{ij}-1)^2 + (n_j-1)^{-1}(N_j-n_j)\}I(k=j) + w_{ik}^2 I(k\neq j)$ and w_{ij} is the i-th component of the vector

$$\mathbf{w}_j = \frac{N_j}{n_j}\mathbf{1}_j + \mathbf{W}(\hat{\theta}_j)\mathbf{X}\{\mathbf{X}^T\mathbf{W}(\hat{\theta}_j)\mathbf{X}\}^{-1}\Big(\sum_{i\in r_j}\mathbf{x}_i - \frac{N_j - n_j}{n_j}\sum_{i\in s_j}\mathbf{x}_i\Big).$$

Here $\mathbf{1}_j$ is the n-vector with i-th component equal to one whenever the corresponding sample unit is in area j and is zero otherwise, and $\mathbf{W}(\hat{\theta}_j)$ is a diagonal matrix of order n defined by the weights obtained from the iterative weighted least squares algorithm used to fit the M-quantile regression model. Tzavidis et al. (2010) have also proposed a nonparametric bootstrap scheme for estimating the MSE of (5).

3 M-quantile geographically weighted regression

In this Section we define a spatial extension to linear M-quantile regression based on GWR. Since M-quantile models do not depend on how areas are specified, we also drop the subscript j from our notation in this Section.

Given n observations at a set of L locations $\{u_l; l = 1, \ldots, L; L \leqslant n\}$ with n_l data values $\{(y_{il}, \mathbf{x}_{il}); i = 1, \ldots, n_l\}$ observed at location u_l, a linear GWR model is a special case of a locally linear approximation to a spatially non-linear regression model and is defined as follows

$$y_{il} = \mathbf{x}_{il}^T \boldsymbol{\beta}(u_l) + \varepsilon_{il}, \tag{7}$$

where $\boldsymbol{\beta}(u_l)$ is a vector of p regression parameters that are specific to the location u_l and the ε_{il} are independently and identically distributed random errors with zero expected value and finite variance. The value of the regression parameter 'function' $\boldsymbol{\beta}(u)$ at an arbitrary location u is estimated using weighted least squares

$$\hat{\boldsymbol{\beta}}(u) = \Big\{ \sum_{l=1}^{L} w(u_l, u) \sum_{i=1}^{n_l} \mathbf{x}_{il} \mathbf{x}_{il}^T \Big\}^{-1} \Big\{ \sum_{l=1}^{L} w(u_l, u) \sum_{i=1}^{n_l} \mathbf{x}_{il} y_{il} \Big\},$$

where $w(u_l, u)$ is a spatial weighting function whose value depends on the distance from sample location u_l to u in the sense that sample observations with locations close to u receive more weight than those further away. In this paper we use a Gaussian specification for this weighting function

$$w(u_l, u) = \exp\Big\{ - d_{u_l,u}^2 / 2b^2 \Big\}, \tag{8}$$

where $d_{u_l,u}$ denotes the Euclidean distance between u_l and u and $b > 0$ is the bandwidth. As the distance between u_l and u increases the spatial weight decreases exponentially. For example, if $w(u_l, u) = 0.5$ and $w(u_m, u) = 0.25$ then observations at location u_l have twice the weight in determining the fit at location u compared with observations at location u_m. Alternative weighting functions, corresponding to density functions other than the Gaussian, can be used. For example, the bi-square function provides a continuous, near-Gaussian weighting function up to distance b from location u and then zero weights any data from locations that are further away. See Fotheringham et al. (2002) for a discussion of other weighting functions. In general, for any type of weighting function sampled observations near location u have more influence on the estimation of the GWR model parameters at u than do sampled observations that are further away. In general, the impact of the weighting function on estimation is reflected in the weight matrix used for deriving estimates of the GWR regression parameters.

The bandwidth b is a measure of how quickly the weighting function decays with increasing distance, and so determines the 'roughness' of the fitted GWR function. A spatial weighting function with a small bandwidth will typically result in a rougher fitted surface than the same function with a large bandwidth. In this paper we use a single bandwidth for our extension of GWR to M-quantile regression. This global bandwidth is defined by minimising the cross-validation criterion proposed by Fotheringham et al. (2002):

$$CV = \sum_{l=1}^{L} \sum_{i=1}^{n_l} [y_{il} - \hat{y}_{(il)}(b)]^2,$$

where $\hat{y}_{(il)}(b)$ is the predicted value of y_{il}, using bandwidth b, with the observation y_{il} omitted from the model fitting process. The value of b that minimises CV is then selected. An alternative approach is to use optimal local bandwidths (Farber & Páez, 2007). However, this significantly increases the computational intensity of the model fitting process.

The GWR model (7) is a linear model for the conditional expectation of y given $\mathbf{X}$ at location u. That is, this model characterises the local behaviour of the conditional expectation of y given $\mathbf{X}$ as a linear function of $\mathbf{X}$. However, a more complete picture of the relationship between y and $\mathbf{X}$ at location u can be constructed by specifying a model for the conditional distribution of y given

$\mathbf{X}$ at this location. Since the M-quantiles serve to characterise this conditional distribution, such a model can be defined by extending (3) to specify a linear model for the M-quantile of order q of the conditional distribution of y given $\mathbf{X}$ at location u, writing

$$Q_q(\mathbf{x}_{il}; \psi, u) = \mathbf{x}_{il}^T \boldsymbol{\beta}_\psi(u; q), \tag{9}$$

where now $\boldsymbol{\beta}_\psi(u; q)$ varies with u as well as with q. Like (7), we can interpret (9) as a local linear approximation, in this case to the (typically) non-linear order q M-quantile regression function of y on $\mathbf{X}$, thus allowing the entire conditional distribution (not just the mean) of y given $\mathbf{X}$ to vary from location to location. The parameter $\boldsymbol{\beta}_\psi(u; q)$ in (9) at an arbitrary location u can be estimated by solving

$$\sum_{l=1}^{L} w(u_l, u) \sum_{i=1}^{n_l} \psi_q\{y_{il} - \mathbf{x}_{il}^T \boldsymbol{\beta}_\psi(u; q)\}\mathbf{x}_{il} = \mathbf{0}, \tag{10}$$

where $\psi_q(\varepsilon) = 2\psi(s^{-1}\varepsilon)\{qI(\varepsilon > 0) + (1-q)I(\varepsilon \leqslant 0)\}$, s is a suitable robust estimate of the scale of the residuals $y_{il} - \mathbf{x}_{il}^T \boldsymbol{\beta}_\psi(u; q)$, e.g. $s = median|y_{il} - \mathbf{x}_{il}^T \boldsymbol{\beta}_\psi(u; q)|/0.6745$, and we typically assume a Huber Proposal 2 influence function, $\psi(\varepsilon) = \varepsilon I(-c \leqslant \varepsilon \leqslant c) + sgn(\varepsilon) I(|\varepsilon| > c)$. Provided c is bounded away from zero, we can solve (10) by combining the iteratively re-weighted least squares algorithm used to fit the 'spatially stationary' M-quantile model (3) and the weighted least squares algorithm used to fit a GWR model. Put $w_\psi(\varepsilon) = \psi_q(\varepsilon)/\varepsilon$ and $w_{\psi il} = w_\psi(\varepsilon_{il})$. Then (10) can be written as

$$\sum_{l=1}^{L} w(u_l, u) \sum_{i=1}^{n_l} w_{\psi il}\{y_{il} - \mathbf{x}_{il}^T \boldsymbol{\beta}_\psi(u; q)\}\mathbf{x}_{il} = \mathbf{0}.$$

Note that the spatial weights $w(u_l, u)$ in (10) do not depend on q. That is, the degree of spatial smoothing is the same at every value of q. Spatial weights that vary with q are straightforward to define by allowing the bandwidth underpinning these weights to vary with q. Such a q-specific optimal bandwidth b can be obtained by minimising the following function with respect to b

$$\sum_{l=1}^{L} \sum_{i=1}^{n_l} [y_{il} - \hat{y}_{(il)}(q; b)]^2,$$

where $\hat{y}_{(il)}(q; b)$ is the estimated value of the right hand side of (9) at quantile q and location u_{il}, using bandwidth b when the observation y_{il} is omitted from the model fitting process. However, using this q-specific cross-validation criterion can significantly increase the computational time. In this paper we therefore use the optimal bandwidth at $q = 0.5$ for all other values of q. We note that this choice could potentially lead to over-smoothing for small or large values of q and hence bias. Nevertheless, it is a reasonable first approximation to the q-specific optimal bandwidth that can be computed reasonably quickly.

An R function (R Development Core Team, 2004) that implements an iterative re-weighted least squares algorithm for for fitting (9) is available from Salvati & Tzavidis (2010). The steps of this algorithm are as follows:

1. For specified q and for each location u of interest, define initial estimates $\boldsymbol{\beta}_\psi^{(0)}(u; q)$.
2. At each iteration t, calculate residuals $\varepsilon_{il}^{(t-1)} = y_{il} - \mathbf{x}_{il}^T \boldsymbol{\beta}_\psi^{(t-1)}(u; q)$ and associated weights $w_{\psi il}^{(t-1)}$ from the previous iteration.
3. Compute the new weighted least squares estimates from

$$\hat{\boldsymbol{\beta}}_\psi^{(t)}(u; q) = \left\{\mathbf{X}^T \mathbf{W}^{*(t-1)}(u; q)\mathbf{X}\right\}^{-1} \mathbf{X}^T \mathbf{W}^{*(t-1)}(u; q)\mathbf{y}, \tag{11}$$

where $\mathbf{y}$ is the vector of n sample values and $\mathbf{X}$ is the corresponding matrix of order $n \times p$ of sample x values. The matrix $\mathbf{W}^{*(t-1)}(u; q)$ is a diagonal matrix of order n with entry, corresponding to a particular sample observation, set equal to the product of this observation's spatial weight, which depends on its distance from location u, and the weight that this observation has when the sample data are used to calculate the 'spatially stationary' M-quantile estimate $\hat{\boldsymbol{\beta}}_\psi(q)$.

4. Repeat steps 1-3 until convergence. Convergence is achieved when the difference between the estimated model parameters obtained from two successive iterations is less than a small pre-specified value.

The fitted regression surface $\hat{Q}_q(\mathbf{x}_{il};\psi,u) = \mathbf{x}_{il}^T\hat{\boldsymbol{\beta}}_\psi(u;q)$ then defines the fit of the M-quantile GWR model for the regression M-quantile of order q of y_{il} given $\mathbf{x}_{il}$ at location u.

Street et al. (1988) proposed an estimator of the covariance matrix of a 'standard' M-estimator of a linear regression parameter vector. Their approach can be easily generalised to the estimation of the covariance matrix of the estimators of the M-quantile and M-quantile GWR regression coefficients.

One may argue that (9) is over-parameterised as it allows for both local intercepts and local slopes. An alternative spatial extension of the M-quantile regression model (3) that has a smaller number of parameters combines local intercepts with global slopes and is defined as

$$Q_q(\mathbf{x}_{il};\psi,u) = \mathbf{x}_{il}^T\boldsymbol{\beta}_\psi(q) + \delta_\psi(u;q), \tag{12}$$

where $\delta_\psi(u;q)$ is a real valued spatial process with zero mean function over the space defined by the locations of interest. Model (12) is fitted in two steps. At the first step we ignore the spatial structure in the data and estimate $\boldsymbol{\beta}_\psi(q)$ directly via the iterative re-weighted least squares algorithm used to fit the standard linear M-quantile regression model (3). Denote this estimate by $\hat{\boldsymbol{\beta}}_\psi(q)$. At the second step we use geographic weighting to estimate $\delta_\psi(u;q)$ via

$$\hat{\delta}_\psi(u;q) = n^{-1}\sum_{l=1}^{L} w(u_l,u)\sum_{i=1}^{n_l}\psi_q\{y_{il} - \mathbf{x}_{il}^T\hat{\boldsymbol{\beta}}_\psi(q)\}. \tag{13}$$

Choosing between (9) and (12) will depend on the particular situation and whether it is reasonable to believe that the slope coefficients in the M-quantile regression model vary significantly between locations. However, it is clear that since (12) is a special case of (9), the solution to (10) will have less bias and more variance than the solution to (13). Hereafter we refer to (9) and (12) as the MQGWR and MQGWR-LI (Local Intercepts) models respectively.

Note that estimates of the local (GWR) M-quantile regression parameters are derived by solving the estimating equation (10) using iterative re-weighted least squares, without any assumption about the underlying conditional distribution of y_{il} given $\mathbf{x}_{il}$ at each location u_l. That is, the approach is distribution-free. Of course, if this conditional distribution is known, and it can be appropriately parameterised, say, by ω, then one can apply methods such as maximum likelihood to the sample data to estimate this parameter by $\hat{\omega}$. The corresponding maximum likelihood estimate of $\boldsymbol{\beta}_\psi(u,q)$ in (9) is then defined by solving the estimating equation

$$\sum_{l=1}^{L} w(u_l,u)\sum_{i=1}^{n_l}\int \psi_q\Big\{Y - \mathbf{x}_{il}^T\boldsymbol{\beta}_\psi(u_l;q)\Big\}dF(Y|\mathbf{x}_{il},u_l;\hat{\omega}) = 0,$$

where $w(v,u)$ is the spatial weighting function of interest, e.g. (8), and $F(Y|\mathbf{x}_{il},u;\omega)$ is the conditional distribution of Y given $\mathbf{x}_{il}$ at location u. A related question concerns the conditions under which the estimating equation (10) corresponds to a maximum likelihood scoring equation. Clearly, this will only be the case when $Q_q(\mathbf{x}_{il};\psi,u) = \mathbf{x}_{il}^T\boldsymbol{\beta}_\psi(u;q)$ is a parameter of the conditional distribution $F(Y|\mathbf{x}_{il},u;\omega)$ with the derivative of the corresponding log density equal to $\psi_q\Big\{Y - \mathbf{x}_{il}^T\boldsymbol{\beta}_\psi(u;q)\Big\}\mathbf{x}_{il}$. For a normal conditional distribution, ψ equals the identity function and with $q = 0.5$ this condition is satisfied. Similarly, when ψ is the sign function and the conditional distribution is Asymmetric Laplace, Koenker (2004) shows that (10) leads to a maximum likelihood solution.

When several conditional quantiles or M-quantiles are estimated, two or more estimated conditional quantile or M-quantile functions can potentially 'cross over' at some point in the space defined by the covariates. This is called *quantile crossing* and may be due to model misspecification, collinearity or the presence of outlying values. A consequence is that the estimated conditional M-quantiles defined by these functions will be incorrectly ordered with respect to q for some values

of the covariates. The problem occurs because each conditional M-quantile function is independently estimated, i.e. without enforcing the property that at each value of $\mathbf{X}$, the M-quantiles of y are ordered by q. He (1997) proposes a simple way of building this restriction into fitted quantile regression lines by a-posteriori restricting them relative to the median regression line. This approach can be easily adapted to fitting M-quantile and M-quantile GWR models. In what follows we describe this procedure for the case of a scalar covariate x. However, the extension to multiple covariates is straightforward. Without loss of generality, we assume that ε has median 0 and $|\varepsilon|$ has median 1. The restricted M-quantile GWR fit is then obtained by:

1. Computing the residuals $\hat{\varepsilon}_{il} = y_{il} - \hat{Q}_{0.5}(x_{il}; \psi, u)$ relative to the M-quantile GWR fit of order $q = 0.5$ at location u.
2. Regressing the absolute values $r_{il} = |\hat{\varepsilon}_{il}|$ of these residuals on the covariate values x_{il} using an M-quantile GWR model with $q = 0.5$ to obtain fitted values $\hat{r}_{il}$.
3. Finding the value $\kappa_q(u) \in (-\infty, +\infty)$ for which $\sum_{l=1}^{L} w(u_l, u) \sum_{i=1}^{n_l} \psi_q(\hat{\varepsilon}_{il} - \kappa_q(u)\hat{r}_{il}) = 0$. Note that if the influence function ψ underlying ψ_q above is the Huber Proposal 2 function, then $\kappa_q(u)$ is monotone in q. This can be shown by a straightforward adaptation of the argument used to prove Proposition 1 of He (1997).
4. The order-restricted M-quantile fit of order q at location u is then $\hat{Q}_q(x_{il}; \psi, u) = \hat{Q}_{0.5}(x_{il}; \psi, u) + \kappa_q(u)\hat{r}_{il}$ where $\hat{r}_{il}$ is evaluated at x_{il}.

In the empirical results reported in this paper, the above algorithm was used when there was evidence of quantile crossing in the unrestricted M-quantile GWR fit to the sample data. Although there are a number of proposals for solving the quantile crossing problem (Cole, 1988; He, 1997), to our knowledge there are no formal tests for detecting whether this problem exists for a particular data set. Obviously one can always carry out a numerical search over the covariate data space to check whether there are values of q and values of $\mathbf{X}$ where this phenomenon occurs. However, this approach quickly becomes unfeasible as the dimension of $\mathbf{X}$ increases. In general, when dealing with small sample sizes and with data that exhibit heteroscedasticity, a safe strategy is to always fit the M-quantile model so that the M-quantile lines do not cross. However, since quantile crossing is typically observed at the boundary of the covariate data space, and usually for either large or small values of q, it is unlikely that this phenomenon will have a severe impact on small area estimation based on a fitted M-quantile regression surface, since such estimates are typically calculated at or near the small area average of $\mathbf{X}$, which will typically be in the interior of the covariate data space.

Finally, we note that outlier robust estimation is not always justified, and when used unnecessarily can potentially lead to a significant loss of efficiency. In such 'outlier-free' cases we can fit instead the expectile (Newey & Powell, 1987) version of the M-quantile GWR model. This is straightforward since all we need to do is to substitute a large value for the tuning constant c in the Huber Proposal 2 influence function that underpins this model, e.g. $c = 100$. As we noted earlier, this tuning constant can be used to trade robustness for efficiency. In particular, as the value of this tuning constant decreases to zero we move towards quantile regression while as its value increases we move towards expectile regression.

4 Using M-quantile GWR models in small area estimation

In a growing number of small area applications, the small area data are geo-coded. Geographical information may be available at the unit level, allowing one to identify the locations of individuals or households, or at a more aggregate level when one has access to the centroids of geographical areas. Developing methods that make efficient use of the spatial information in SAE is therefore important. This spatial information can be incorporated directly into the model regression structure via an M-quantile GWR model and in this Section we describe how this can be achieved. In addition to assumptions about the structure of the population, the number of small areas and the sample and population sizes made at the start of Section 2, we now assume that we have only one population value per location, allowing us to drop the index l. We also assume that the geographical coordinates of every unit in the population are known, which is the case with geo-coded data. The aim is to use these data to predict the area j mean m_j of y using the M-quantile GWR models (9) and (12).

Following Chambers & Tzavidis (2006), we first estimate the M-quantile GWR coefficients $\{q_i; i \in s\}$ of the sampled population units without reference to the small areas of interest. A grid-based interpolation procedure for doing this under (3) is described by Chambers & Tzavidis (2006) and can be used directly with (12). We adapt this approach to the GWR M-quantile model (9) by first defining a fine grid of q values in the interval $(0, 1)$. Chambers & Tzavidis (2006) use a grid that ranges between 0.01 and 0.99 with step 0.01. We employ the same grid definition and then use the sample data to fit (9) for each distinct value of q on this grid and at each sample location. The M-quantile GWR coefficient for unit i with values y_i and $\mathbf{x}_i$ at location u_i is finally calculated by using linear interpolation over this grid to find the unique value q_i such that $\hat{Q}_{q_i}(\mathbf{x}_i; \psi, u_i) \approx y_i$.

Provided there are sample observations in area j, an area j specific M-quantile GWR coefficient, $\hat{\theta}_j$ can be defined as the average value of the sample M-quantile GWR coefficients in area j, otherwise we set $\hat{\theta}_j = 0.5$. Following Tzavidis et al. (2010), the bias-adjusted M-quantile GWR predictor of the mean m_j in small area j is then

$$\hat{m}_j^{MQGWR/CD} = N_j^{-1}\Big[\sum_{i\in U_j} \hat{Q}_{\hat{\theta}_j}(\mathbf{x}_i; \psi, u_i) + \frac{N_j}{n_j}\sum_{i\in s_j}\{y_i - \hat{Q}_{\hat{\theta}_j}(\mathbf{x}_i; \psi, u_i)\}\Big], \tag{14}$$

where $\hat{Q}_{\hat{\theta}_j}(\mathbf{x}_i; \psi, u_i)$ is defined via the MQGWR model (9), the MQGWR-LI model (12), or the expectile GWR model discussed at the end of the previous section. Empirical comparisons of the 'large c' (i.e. expectile) and the more robust 'small c' Huber-type M-quantile small area models are reported later in this paper.

There are situations where we are interested in estimating small area characteristics for domains (areas) with no sample observations. The conventional approach to estimating a small area characteristic, say the mean, in this case is synthetic estimation. Under the linear mixed model (1) the synthetic mean predictor for out of sample area j is $\hat{m}_j^{LM/SYNTH} = N_j^{-1}\sum_{i\in U_j}\mathbf{x}_i^T\hat{\boldsymbol{\beta}}$. Under the M-quantile GWR model (9) the synthetic mean predictor for out of sample area j is $\hat{m}_j^{MQGWR/SYNTH} = N_j^{-1}\sum_{i\in U_j}\hat{Q}_{0.5}(\mathbf{x}_i; \psi, u_i)$. We note that with MQGWR-based synthetic estimation all variation in the area-specific predictions comes from the area-specific auxiliary information, including the locations of the population units in the area. We expect that when a truly spatially non-stationary process underlies the data, use of $\hat{m}_j^{MQGWR/SYNTH}$ will lead to improved efficiency relative to more conventional synthetic mean predictors. Empirical results that address the issue of out of sample area estimation are set out in Section 6.

5 Mean squared error estimation

A 'pseudo-linearization' MSE estimator for M-quantile small area estimators was recommended by Chambers & Tzavidis (2006) and has now been used successfully in empirical studies reported in a number of published papers on SAE, including the recent publications by Tzavidis et al. (2010) and Salvati et al. (2010). We extend this approach to defining an estimator of a first order approximation to the mean squared error of (14). This extension is based on (i) a model where the regression of y on $\mathbf{X}$ for a particular population unit depends on its location, with this regression specified by the locally linear GWR model (7), and (ii) the fact that estimators derived under the MQGWR model (9) or the MQGWR-LI model (12) can be written as linear combinations of the sample values of y. For example, from (11) we see that (14) can be expressed as a weighted sum of the sample y-values

$$\hat{m}_j^{MQGWR/CD} = N_j^{-1}\mathbf{w}_j^T\mathbf{y}, \tag{15}$$

where

$$\mathbf{w}_j = \frac{N_j}{n_j}\mathbf{1}_j + \sum_{i\in r_j}\mathbf{H}_{ij}^T\mathbf{x}_i - \frac{N_j - n_j}{n_j}\sum_{i\in s_j}\mathbf{H}_{ij}^T\mathbf{x}_i. \tag{16}$$

Here $\mathbf{1}_j$ is the n-vector with i-th component equal to one whenever the corresponding sample unit is in area j and is zero otherwise and

$$\mathbf{H}_{ij} = \Big\{\mathbf{X}^T\mathbf{W}^*(u_i;\hat{\theta}_j)\mathbf{X}\Big\}^{-1}\mathbf{X}^T\mathbf{W}^*(u_i;\hat{\theta}_j),$$

where $\mathbf{W}^*(u;q)$ is the limit of the weighting matrices $\mathbf{W}^{*(t-1)}(u;q)$ defined following (11).

If we assume that the weights defining (15) are fixed, and that the values of y follow a location specific linear model, e.g. (7), then an estimator of the prediction variance of (15) can be computed following standard methods of heteroskedasticity-robust prediction variance estimation for linear predictors of population quantities (Royall & Cumberland, 1978). Put $\mathbf{w}_j = (w_{ij})$. This estimator is of the form

$$mse(\hat{m}_j^{MQGWR/CD}) = N_j^{-2}\sum_{k:n_k>0}\sum_{i\in s_k}\lambda_{ijk}\Big\{y_i - \hat{Q}_{\hat{\theta}_k}(\mathbf{x}_i;\psi,u_i)\Big\}^2, \tag{17}$$

where $\lambda_{ijk} = \Big\{(w_{ij}-1)^2 + (n_j-1)^{-1}(N_j-n_j)\Big\}I(k=j) + w_{ik}^2 I(k\neq j)$ and $\hat{Q}_{\hat{\theta}_k}(\mathbf{x}_i;\psi,u_i)$ is assumed to define an unbiased estimator of the expected value of y_i given $\mathbf{x}_i$ at location u_i in area k. Since the weights defining (16) reproduce the small area mean of $\mathbf{X}$, it also follows that (15) is unbiased for this mean in the special case where this expectation does not vary with location within the small area of interest, and so (17) then estimates the mean squared error of (15) in this case. More generally, when the expectation of y_i given $\mathbf{x}_i$ varies from location to location within the small area, this unbiasedness holds on average provided sampling within the small area is independent of location, in which case (17) is an estimator of a first order approximation to the mean squared error of (15). We observe that (17) can also be used as an MSE estimator for the expectile version of (14), since the only difference in this case is the value of the tuning constant c in the influence function that defines the M-quantiles underpinning the small area estimators.

The theoretical basis for the pseudo-linearization approach to MSE estimation is described by Chambers et al. (2009) and here we provide a summary of its main properties. This approach is generally applicable in the sense that it can be used for estimating the MSE of any estimator that can be expressed in a pseudo-linear form i.e. as a weighted sum of sample values, and is the main analytical approach for estimating the MSE of M-quantile small area predictors. As noted earlier, the approach is based on an extension of the heteroskedasticity-robust approach to prediction variance estimation described in Royall & Cumberland (1978), and leads to MSE estimators that are simple to implement and potentially misspecification bias robust. Its main drawback is that these MSE estimators can suffer from increased variability, especially when the area-specific sample sizes are very small. For this reason, we suggest that it should be used when there is reasonable evidence of failure of the assumptions of a linear mixed model and hence the potential bias robustness of this approach outweighs its increased variability.

Note also that (17) treats the weights (16) as fixed and so ignores the contribution to the mean squared error from the estimation of the area level M-quantile coefficients by $\hat{\theta}_j$. This is a pseudo-linearization assumption since for large overall sample sizes the contribution to the overall mean squared error of (15) arising from the variability of $\hat{\theta}_j$ is of smaller order of magnitude than the fixed weights prediction variance of (15). As a consequence (17) will tend to be almost unbiased. However, the potential underestimation of the MSE of (15) implicit in (17) needs to be balanced against the bias robustness of this MSE estimator under misspecification of the second order moments of y, and may well lead to (17) being preferable to other MSE estimators based on higher order approximations that depend on the model assumptions being true. Empirical results reported in Tzavidis et al. (2010) indicate that the version of the MSE estimator (17) for the linear M-quantile predictor performs well in both model-based and design-based simulation studies.

6 Simulation studies

We present results from two types of simulation studies that are used to examine the performance of the small area estimators discussed in the preceding Sections. In Section 6.1 we report results

from model-based simulations. In this case population data are generated at each simulation using a linear mixed model with different parametric assumptions about the distribution of errors and the spatial structure of the data, and a single sample is then taken from this simulated population according to a pre-specified design. In Section 6.2 we report results from a design-based simulation: actual survey data are first used to simulate a population with spatial characteristics similar to those of the original sample, and this fixed population is then repeatedly sampled according to a pre-specified design. In our case the survey data come from the Environmental Monitoring and Assessment Program (EMAP), which is part of the Space Time Aquatic Resources Modelling and Analysis Program (STARMAP) at Colorado State University.

6.1 Model-based simulations

In these simulations, synthetic population values are generated under two versions of a linear mixed model and two distributional specifications for the random area effects and the individual residuals. Each population is of size $N = 10,500$ and contains $d = 30$ equal-sized small areas. More specifically, under the first model, population values of y are generated via $y_{ij} = 1 + 2x_{ij} + \gamma_j + \varepsilon_{ij}$ where $i = 1, \ldots, 350$ and $j = 1, \ldots, 30$. The values x_{ij} in this model are independently generated from the uniform distribution over the interval $[0, 1]$, denoted as $x_{ij} \sim U[0, 1]$, and the random effects are generated under two different distributional specifications: (a) Gaussian errors with $\gamma_j \sim N(0, 0.04)$ and $\varepsilon_{ij} \sim N(0, 0.16)$ and (b) Chi-squared errors with $\gamma_j \sim \chi^2(1) - 1$ and $\varepsilon_{ij} \sim \chi^2(3) - 3$, i.e. mean corrected Chi-squared variates with 1 and 3 degrees of freedom, respectively. For the second model, random effects are still simulated as in (a) and (b), but in addition the intercept and the slope of the linear model for y are allowed to vary with longitude and latitude. In particular, these simulations are based on the two-level model $y_{ij} = \alpha_{ij} + \beta_{ij}x_{ij} + \gamma_j + \varepsilon_{ij}$ with

$$\alpha_{ij} = 0.2 \times longitude_{ij} + 0.2 \times latitude_{ij},$$

$$\beta_{ij} = -5 + 0.1 \times longitude_{ij} + 0.1 \times latitude_{ij}$$

with the known location coordinates $(longitude_{ij}, latitude_{ij})$ for each population unit independently generated from $U[0, 50]$. Note that the reason for using different parametric assumptions for the error terms of the linear mixed model is because we are interested in how the small area predictors perform both when the Gaussian assumptions of the linear mixed model are satisfied and when these assumptions are violated.

This simulation design corresponds to four scenarios (Gaussian stationary, Gaussian non-stationary, Chi-squared stationary, Chi-squared non-stationary). For each of these scenarios $T = 200$ Monte-Carlo populations are generated using the corresponding model specifications. For each generated population and for each area j we select a simple random sample (without replacement) of size $n_j = 20$, leading to an overall sample size of $n = 600$. The sample values of y and the population values of x obtained in each simulation are then used to estimate the small area means. Although a larger number of simulations would be preferable, this is not feasible due to the computer intensive nature of the model-fitting process. Note also that there is no specific motivation behind the choice of equal area specific sample sizes. Repetition of our simulation studies with unequal area-specific sample sizes does not lead to any differences in the conclusions that we draw below. These results of these simulations are not reported here, but are available from the authors.

Four different types of small area linear models are fitted to these simulated data. These are (i) a random intercepts version of (1), (ii) the linear M-quantile regression specification (3), (iii) the MQGWR model (9),(iv) the MQGWR-LI model (12), and (v) the expectile GWR model that models conditional expectations instead of conditional quantiles. The last three models make use of the known locations of the population units.

The random intercepts model (i) is fitted using the default REML option of the *lme* function (Venables & Ripley, 2002, Section 10.3) in R. The M-quantile linear regression model (ii) is fitted using a modified version of the *rlm* function (Venables & Ripley, 2002, Section 8.3) in R and so uses iteratively re-weighted least squares (Chambers & Tzavidis, 2006). An extended version of this R code, available from Salvati & Tzavidis (2010), is used to fit the MQGWR models (iii), (iv), and (v).

In particular, model (v) is fitted by setting the value of the tuning constant in the Huber Proposal 2 influence function to $c = 20$. On the other hand the outlier robust M-quantile regression and the M-quantile GWR models use the Huber Proposal 2 influence function with $c = 1.345$. This value gives 95% efficiency in the normal case while protecting against outliers (Huber, 1981). Estimated model coefficients obtained from these fits are used to compute the EBLUP (2), the bias-adjusted M-quantile predictor (5), denoted by MQ below and the MQGWR, the Expectile GWR and the MQGWR-LI versions of the corresponding bias-adjusted M-quantile predictor (14).

The performance of the different small area estimators is evaluated with respect to two basic criteria: the bias and the root mean squared error of estimates of the small area means. The bias for small area j is computed as

$$Bias_j = \frac{1}{T}\sum_{t=1}^{T}(\hat{m}_{jt} - m_{jt}),$$

and the root mean squared error for area j is computed as

$$RMSE_j = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(\hat{m}_{jt} - m_{jt})^2}.$$

Note that the subscript t here indexes the T Monte-Carlo simulations, with m_{jt} denoting the value of the small area j mean in simulation t and $\hat{m}_{jt}$ denoting the area j estimated value in simulation t.

Key percentiles of the across areas distributions of the prediction biases and root mean squared errors of these estimators over the simulations are set out in Table 1. For Gaussian random effects and a spatially stationary regression surface, we see that the EBLUP is the best predictor in terms of RMSE, as one would expect. The MQ, MQGWR and MQGWR-LI predictors all have similar bias and RMSE in this case. In contrast, when the underlying regression function is non-stationary we see that the GWR-based predictors appear to be more efficient than the MQ and EBLUP predictors. As expected, the Expectile GWR estimator is more efficient overall than the corresponding M-quantile predictors because under these scenarios there is no reason to employ outlier-robust estimation. Under Chi-squared random effects this performance is unchanged, although the absolute differences in performance between the various predictors is much smaller. For a non-stationary Chi-squared process the RMSE of the MQGWR estimator is smaller than the RMSE of the Expectile GWR estimator because, in this case, the MQGWR estimator protects against outlying values.

In Figure 1 we show how the mean squared error estimator (using (17) and averaging over simulations) tracks the true mean squared error of the MQGWR and MQGWR-LI predictors under the Chi-squared scenarios. The detailed results under Gaussian scenarios are not reported here but are available from the authors. In general, the proposed mean squared error estimator (17) provides a good approximation to the true mean squared error. These results also show that when M-quantile GWR fits are used in (17), then this estimator underestimates the true mean squared error of the corresponding predictor. This is consistent with both the MQGWR and the MQGWR-LI models over-fitting the actual population regression function. However, this bias is not excessive, being more pronounced in the case of the MQGWR model. The performance of the MSE estimator under the expectile GWR model is similar to that obtained by MQGWR. These results are not reported here, but are also available from the authors upon request.

Finally, we note that one could combine the M-quantile model-based estimators with the MSE estimation method described in Section 5 to generate 'normal theory' confidence intervals for the small area means. Coverage results based on such intervals have been produced and are available from the authors. However, this use of the estimated MSE to construct confidence intervals, though widespread, has been criticised. Hall & Maiti (2006) and more recently Chatterjee et al. (2008) discuss the use of bootstrap methods for constructing confidence intervals for small area parameters, arguing that there is no guarantee that the asymptotic behaviour underpinning normal theory confidence intervals applies in the context of the small samples that characterise small area

estimation. Further research on using bootstrap techniques to construct confidence intervals under the M-quantile GWR model is left for the future.

[Table 1 about here.]

[Fig. 1 about here.]

6.2 Design-based simulation

The data used in this design-based simulation comes from the U.S. Environmental Protection Agency's Environmental Monitoring and Assessment Program (EMAP) Northeast lakes survey (Larsen et al., 1997). Between 1991 and 1995, researchers from the U.S. Environmental Protection Agency conducted an environmental health study of the lakes in the north-eastern states of the U.S.A. For this study, a sample of 334 lakes (or more accurately, lake locations) was selected from the population of 21,026 lakes in these states using a random systematic design. The lakes making up this population are grouped into 113 8-digit Hydrologic Unit Codes (HUCs), of which 64 contained less than 5 observations and 27 did not have any. In our simulation, we defined HUCs as the small areas of interest, with lakes grouped within HUCs. The variable of interest was Acid Neutralizing Capacity (ANC), an indicator of the acidification risk of water bodies. Since some lakes were visited several times during the study period and some of these were measured at more than one site, the total number of observed sites was 349 with a total of 551 measurements. In addition to ANC values and associated survey weights for the sampled locations, the EMAP data set also contained the elevation and geographical coordinates of the centroid of each lake in the target area. In our simulations we use elevation to define the fixed part of the mixed models and the M-quantile models for the ANC variable.

The aim of the design-based simulation is to compare the performance of different predictors of mean ANC in each HUC under repeated sampling from a fixed population with the same spatial characteristics as the EMAP sample. In particular, given the 21,026 lake locations, a synthetic population of ANC individual values is first non-parametrically simulated using a nearest-neighbour imputation algorithm that retains the spatial structure of the observed ANC values in the EMAP sample data.

The algorithm for this simulation is as follows: (1) we first randomly order the non-sampled locations in order to avoid list order bias and give each sampled location a 'donor weight' equal to the integer component of its survey weight minus 1; (2) taking each non-sample location in turn, we choose a sample location as a donor for the i-th non-sample location by selecting one of the ANC values of the EMAP sample locations with probability proportional to $w(u_i, u) = \exp\left\{ -d^2_{u_i,u}/2b^2 \right\}$. Here $d_{u_i,u}$ is the Euclidean distance from the i-th non-sample location u_i to the location u of a sampled location and b is the GWR bandwidth estimated from the EMAP data; and (3) we reduce the donor weight of the selected donor location by 1. The synthetic population of ANC values created by this procedure is then kept fixed over the Monte-Carlo simulations.

A total of 200 independent random samples of lake locations are then taken from the population of 21,026 lake locations by randomly selecting locations in the 86 HUCs that containing EMAP sampled lakes, with sample sizes in these HUCs set to the greater of five and the original EMAP sample size. Lakes in HUCs not sampled by EMAP are also not sampled in the simulation study. This results in a total sample size of 652 locations within the 86 'EMAP' HUCs. The synthetic ANC values at these 652 sampled locations then constitute the sample data.

Before presenting the results from this simulation study we show some model and spatial diagnostics. Figure 2 shows normal probability plots of level 1 and level 2 residuals obtained by fitting a two-level (level 1 is the lake and level 2 is the HUC) mixed model to the synthetic population data. The normal probability plots indicate that the Gaussian assumptions of the mixed model are not met. Using a model that relaxes these assumptions, such as an M-quantile model with a bounded influence function, therefore seems reasonable for these data. An alternative to the use of a robust model would have been to transform the ANC variable of the EMAP data. A popular transformation in small area applications is the logarithmic one. For the EMAP dataset,

however, a logarithmic or a square root transformation cannot be directly applied because of the negative values in the outcome variable, where a negative value of ANC implies water acidity. This problem can be overcome by adding a sufficiently large positive constant to the individual ANC values such that the resulting modified ANC values are all strictly positive. If we do this, and refit the model with HUC-specific random effects, we obtain normal probability plots of level 1 and level 2 residuals that are closer to what is expected under normality. Nevertheless, there are still clear departures from normality. This is confirmed by a Shapiro-Wilk normality test, which rejects the null hypothesis that the residuals follow a normal distribution (p-values: level 1 = 0.03555, level 2 = 1.715e-14). Even if in some cases we are willing to assume normality, using a transformation can create further complexities at the later stages of the small area estimation process. In particular, after performing small area estimation with the transformed data, the estimates must be back-transformed to the original scale. It is well known that this back-transformation can introduce bias in small area estimation, which must then be corrected (Chandra & Chambers, 2010). There is also evidence of a non-stationary process. In particular, using an ANOVA test proposed by Brundson et al. (1999) we reject the null hypothesis of stationarity of the model parameters. Figure 3 shows contour maps of the estimated HUC-specific intercepts and slopes from the fitted GWR model. These maps support the assumption of non-stationarity in the data. Examining the contours of the slope coefficients in Figure 3 we see that the effect of elevation on ANC varies spatially, with these slope coefficients ranging from -40 to 40. The average value of ANC also shows spatial variation. In particular, the contour map of the intercept coefficients shows them ranging from 0 (East) to 4000 (Centre-West). Based on these spatial diagnostics we expect that incorporating the spatial information in small area estimation may lead to gains in efficiency for this population.

[Fig. 2 about here.]

[Fig. 3 about here.]

The relative bias (RB) and the relative root mean squared error (RRMSE) of estimates of the mean value of ANC in each HUC are computed for the same five predictors that are also the focus of the model-based simulations. These results are set out in Table 2 and show that the M-quantile GWR predictors have in general lower bias than the EBLUP predictor. Examining the performance in terms of relative root mean squared error we note that the small area predictors that account for the spatial structure of the data have on average smaller root mean squared errors with the Expectile GWR predictor performing best overall. Hence, although the robust estimators adjust for bias, reflected in the lower bias of MQGWR, this adjustment comes at the cost of higher variance, which illustrates the bias-variance trade off in deciding the value of the tuning constant c. These results also indicate that incorporating the spatial information in small area estimation via the M-quantile GWR model has promise.

[Table 2 about here.]

For the non-sampled HUCs the use of the synthetic-type predictors that borrow strength over space, as defined in Section 4, also substantially improve prediction.

Figure 4 shows how the mean squared error estimator described in Section 5 tracks the true mean squared error of the different MQGWR predictors. We can see that the GWR form (17) of the mean squared estimator described in Tzavidis et al. (2010) performs well in terms of tracking the true mean squared error. Some downward bias of (17) when used with the MQGWR model (9) can be seen, however. This is much less of a problem when (17) is combined with the MQGWR-LI model (12).

[Fig. 4 about here.]

7 Application: Assessing the ecological condition of lakes in the northeastern U.S.A.

In this Section we show how the methodology described in this paper can be practically employed for estimating the average acid neutralizing capacity (ANC) for each of the 113 8-digit HUCs that

make up the EMAP dataset described in Section 6.2. ANC is a measure of the ability of a solution to resist changes in pH and is on a scale measured in meq/L (micro equivalents per litre). A small ANC value for a lake indicates that it is at risk of acidification.

Predicted values of average ANC for each HUC are calculated using the M-quantile GWR predictor (14) under the MQGWR model (9) and the MQGWR-LI model (12), with x equal to the elevation of each lake and with location defined by the geographical coordinates of the centroid of each lake (in the UTM coordinate system). The spatial weight matrix used in fitting these M-quantile GWR models is constructed using (8), with bandwidth selected using cross-validation.

In Figure 5 we show maps of estimated values of average ANC for each HUC using (a) the MQGWR model; (b) the expectile GWR model; (c) the MQGWR-LI model; (d) the spatially stationary M-quantile model (3), and (e) the linear mixed model (1). Overall, all small area models indicate that there are lower levels of average ANC (higher risk of water acidification) in the Eastern part of the study region. However, these small area models also produce substantially different estimates of average ANC in the South-Western part of the study region. In particular, maps (a), (b) and (c) that correspond to the two M-quantile GWR models provide similar estimates of average ANC for each HUC and are consistent with the spatial distribution of ANC average values produced by previous non-parametric analyses of the EMAP data (Opsomer et al., 2008; Pratesi et al., 2008). This indicates that small area models that allow for more flexible incorporation of the spatial information produce overall consistent results. Moreover, the results from the M-quantile GWR models are substantially different from the results illustrated by map (e) which shows the estimates produced by the EBLUP under the linear mixed model (1). This map shows lower levels of average ANC (and hence greater risk of water acidification) for the target population of HUCs. Finally, we see that the M-quantile model (3) that assumes no spatial structure in the data leads to map (d), which shows even lower levels of average ANC. This is most likely due to the failure of the spatial stationarity assumption in this model when it is applied to the EMAP data.

[Fig. 5 about here.]

Finally, we briefly discuss an alternative, and more parsimonious, approach to incorporating spatial information in small area models. This allows for spatial structure by including parametrically specified spatial terms in the mean part of the model. To illustrate, consider the model used to generate the data for the model-based simulations in Section 6.1. Here, a non-stationary spatial process was generated by allowing the fixed effects in a linear mixed model to vary by longitude and latitude. In this case we know the 'true' data generating process, and so the best performing model in estimation will obviously be the one defining this process, i.e. the linear random effects model that includes longitude and latitude as main effects, together with the interactions between the covariate x and longitude, and between x and latitude. With real data, however, we will not know the true data generation mechanism. If we suspect that a non-stationary process is present in the data, we could then try to model this process by including the geographical coordinates in the fixed part of the model as well as any interactions between these coordinates and other model covariates after evaluating whether the addition of these terms improves the fit of the model. Unfortunately, in most practical situations it is difficult to capture the spatial non-stationary process by just including interaction terms. As we saw in the case of the EMAP data, an ANOVA test suggests that there is spatial non-stationarity. The question that arises then is whether this non-stationary process can be modeled in a more parsimonious way via a linear random effects model that includes the geographical coordinates and the two interaction terms between elevation and these coordinates as fixed effects. In order to answer this question we assessed the fit of three models using the AIC criterion. Model 1 includes only elevation (AIC = 7968.45), Model 2 adds longitude and latitude in the fixed part of the model (AIC = 7932.02) and finally model 3 additionally includes two interaction effects defined by longitude by elevation and latitude by elevation (AIC = 7935.54). Examining the values of the AIC we conclude that adding the geographical coordinates improves the model fit but the inclusion of the two interaction terms does not improve the fit of the model any further. In order to empirically assess the performance of global models that include the geographical coordinates as covariates, in the design-based simulation we also produced results using the EBLUP based on a global model that included these coordinates as

fixed effects in addition to elevation. Note that the population of this simulation study was constructed by non-parametrically bootstrapping the population of the original EMAP data in a way that approximately preserves the structure of the original data. The results, which are available from the authors, suggest that the EBLUP estimator that includes the geographical coordinates in the model specification does not perform better than the GWR-based small area estimators. This indicates that there may be situations where a spatial non-stationary process is present but trying to capture this process by adding covariates that are functions of the geographical coordinates may not improve the performance of the corresponding small area estimators.

8 Summary

In this paper we propose a geographically weighted regression extension to linear M-quantile regression that allows for spatially varying coefficients in the model for the M-quantiles. These M-quantile GWR models have the potential to lead to significantly better small area estimates in important application areas where geo-coded data with spatial structure is available, such as in financial, economic, environmental and public health applications.

Similarly to the linear M-quantile regression model by Chambers & Tzavidis (2006), the M-quantile GWR model described in this paper allows modelling of between area variability without the need to explicitly specify the area-specific random components of the model. In particular, this model does not explicitly depend on any particular small area geography, and so can be easily adapted to different geographies as the need arises. The properties of the MQGWR predictors have been studied through model-based and design based simulation studies. The results from these studies suggest that the M-quantile GWR model represents a promising alternative for flexibly incorporating spatial information into SAE. In addition, the performance of the proposed MSE estimator for the M-quantile GWR small area predictors is promising, but we are aware that further research in this area is necessary. The applicability of the M-quantile GWR small area methodology is demonstrated using environmental data from a survey of lakes in the north-eastern region of the USA. The results are in line with those of the previous analyses with the same data (Opsomer et al., 2008; Pratesi et al., 2008) and illustrate the need for flexible and versatile ways of incorporating spatial information in small area models.

R code for fitting the M-quantile GWR small area models that we propose in this paper is available from Salvati & Tzavidis (2010). Note, however, that a prospective user of the M-quantile GWR model should have access to an appropriate level of spatial information. For example, the survey dataset used in the application of this paper includes detailed spatial information for sampled and non-sampled locations. The model can be adapted to situations where more limited spatial information is available, e.g. when only spatial information about the centroids of the small areas or other aggregated spatial information is available. Obviously, in such cases the gains from including this information in analysis will be smaller.

One problem that arises when specifying an M-quantile GWR model is in deciding which parameters of the model vary spatially (i.e. are local parameters) and which do not (i.e. are global parameters). In this paper we have explored two M-quantile GWR models that exemplify this issue - the MQGWR/expectile GWR model (9) where both intercept and slope parameters in the model vary spatially and the MQGWR-LI model (12) where only the intercept parameter varies spatially. Further research is necessary in order to develop appropriate diagnostics for deciding between them.

An alternative approach for incorporating the spatial structure of the data in small area models is via nonparametric models. Opsomer et al. (2008) and Ugarte et al. (2009) have extended model (1) to the case in which the small area random effects can be combined with a smooth, non-parametrically specified trend. These authors express the nonparametric small area model as a random effects model. Pratesi et al. (2008) have extended this approach to the M-quantile small area estimation approach using a nonparametric specification of the conditional M-quantiles of the response variable given the covariates. Both bivariate p-spline approximations for fitting nonparametric unit level nested error and M-quantile regression models allow for spatial variation in the data, which can then be used to define nonparametric models for small area estimation.

Further research is therefore necessary in order to understand how M-quantile GWR and unit level nested error p-spline regression models compare in terms of their SAE performance. Finally, we are currently investigating the use of the M-quantile GWR small area model for estimating income distribution functions and the incidence of poverty for small areas.

Acknowledgements The work in this paper has been in part supported by project PRIN *Efficient use of auxiliary information at the design and at the estimation stage of complex surveys: methodological aspects and applications for producing official statistics* awarded by the Italian Government to the Universities of Cassino, Florence, Perugia, Pisa and Trieste, and by ARC Linkage Grant LP0776810 of the Australian Research Council. The work is also supported by the project SAMPLE 'Small Area Methods for Poverty and Living Condition Estimates' (www.sample-project.eu), financed by the European Commission under the 7th Framework Program. The authors are grateful to the Space-Time Aquatic Resources Modelling and Analysis Program (STARMAP) for providing access to the data used in this paper. The views expressed here are solely those of the authors.

References

Anselin, L. (1992). *Spatial Econometrics. Methods and Models.* Boston: Kluwer Academic Publishers.

Breckling, J. & Chambers, R. (1988). M-quantile. *Biometrika* **75**, 761-771.

Brundson, C., Fotheringham, A.S. & Charlton, M.(1996). Geographically weighted regression: a method for exploring spatial nonstationarity. *Geographical Analysis* **28**, 281-298.

Brundson, C., Fotheringham, A.S. & Charlton, M.(1999). Some notes on parametric significance tests for geographically weighted regression. *Journal of Regional Science* **39**, 497-524.

Chandra, H. & Chambers, R. (2010). Small Area Estimation Under Transformation to Linearity. Tentatively accepted in *Survey Methodology.*

Chambers, R. & Dunstan, R. (1986). Estimating distribution function from survey data. *Biometrika* **73**, 597-604.

Chambers, R. & Tzavidis, N. (2006). M-quantile Models for Small Area Estimation. *Biometrika* **93**, 255-268.

Chambers, R. & Chandra, H. & Tzavidis, N. (2009). On Bias-Robust Mean Squared Error Estimation for Linear Predictors for Domains. *Working Papers, 09-08* Centre for Statistical and Survey Methodology, The University of Wollongong, Australia. (Available from: http://cssm.uow.edu.au/publications).

Chatterjee, S., Lahiri, P. & Huilin, L. (2008). Parametric Bootstrap Approximation to the Distribution of EBLUP and Related Prediction Intervals in Linear Mixed Models. *Annals of Statistics* **36**, 1221–1245.

Cole, R.J. (1988). Fitted smoothed centile curves to reference data. *Journal of the Royal Statistical Society A* **151**, 385–418.

Cressie, N. (1993). *Statistics for Spatial Data.* New York: John Wiley & Sons.

Farber, S. & Páez, A.(2007). A systematic investigation of cross-validation in GWR model estimation: empirical analysis and Monte Carlo simulations. *Journal of Geographical Systems* **9**, 371–396.

Fotheringham, A.S., Brundson, C. & Charlton, M.(1997). Two techniques for exploring non-stationarity in geographical data. *Journal of Geographical Systems* **4**, 59-82.

Fotheringham, A.S., Brundson, C. & Charlton, M.(2002). *Geographically Weighted Regression* West Sussex: John Wiley & Sons.

Hall, P. & Maiti, T. (2006). Nonparametric estimation of mean squared prediction error in nested-error regression models. *Annals of Statistics* **34**, 1733–1750.

He, X. (1997). Quantile curves without crossing. *The American Statistician* **51**, 186–192.

Henderson, C. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics* **31**, 423–447.

Huber, P. J. (1981). *Robust Statistics.* London: Wiley.

Koenker, R. & Bassett, G. (1978). Regression Quantiles. *Econometrica* **46**, 33–50.

Koenker, R. (2004). Quantile regression for longitudinal data. *Journal of Multivariate Analysis* **91**, 74–89.

Kokic, P., Chambers, R., Breckling, J. & Beare, S. (1997). A measure of production performance. *Journal of Business and Economic Statistics* **10**, 419–435.

Jiang, J. & Lahiri, P. (2006). Mixed model prediction and small area estimation (with discussions). *TEST* **15**, 1–96.

Larsen, D. P., Kincaid, T. M., Jacobs, S. E. & Urquhart, N. S.(2001). Designs for evaluating local and regional scale trends. *Bioscience* **51**, 1049–1058.

Newey, W.K. & Powell, J.L. (1987). Asymmetric least squares estimation and testing, Econometrica. *Econometrica* **55**, 819–847.

Opsomer, J. D., Claeskens, G., Ranalli, M.G., Kauermann, G. & Breidt, F.J. (2008). Nonparametric small area estimation using penalized spline regression. *Journal of the Royal Statistical Society, Series B* **70**, 265–286.

Petrucci, A. & Salvati, N. (2004). Small area estimation: the Spatial EBLUP at area and unit level. *Metodi per l'integrazione di dati da più fonti* (Liseo B., Montanari G.E., Torelli N.), eds. Franco Angeli, Milano, 37–58.

Prasad, N. & Rao, J. (1990). The estimation of the mean squared error of small-area estimators. *Journal of the American Statistical Association* **85**, 163–171.

Pratesi, M. & Salvati, N. (2008). Small area estimation: the EBLUP estimator based on spatially correlated random area effects. *Statistical Methods & Applications* **17**, 113–141.

Pratesi, M., Ranalli, M.G. & Salvati, N.(2008). Semiparametric M-quantile regression for estimating the proportion of acidic lakes in 8-digit HUCs of the Northeastern US. *Environmetrics* **19-7**, 687–701.

R Development Core Team (2004). R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. URL: http://www.R-project.org.

Rao, J.N.K., Kovar, J.G. & Mantel, H.J. (1990). On Estimating Distribution Functions and Quantiles from Survey Data Using Auxiliary Information. *Biometrika* **77**, 365–375.

Rao, J. N. K. (2003). *Small Area Estimation.* London: Wiley.

Rao, J. N. K. (2005). Inferential issues in small area estimation: some new developments. *Statistics in Transition*, **7**, 513–526.

Royall, R.M. & Cumberland, W.G. (1978). Variance estimation in finite population sampling. *Journal of the American Statistical Association*, **73**, 351–358.

Salvati, N. & Tzavidis, N. (2010). M-quantile GWR function. Available from *URL: http://www.dipstat.ec.unipi.it/persone/docenti/salvati/.*

Salvati, N., Chandra, H., Ranalli, M.G. & Chambers, R. (2010). Small Area Estimation Using a Nonparametric Model Based Direct Estimator. *Computational Statistics and Data Analysis* **54**, 2159–2171

Singh, B., Shukla, G. & Kundu, D. (2005). Spatio-temporal models in small area estimation. *Survey Methodology* **31**, 183–195.

Street, J.O., Carroll, R.J. & Ruppert, D. (1988). A note on computing robust regression estimates via iteratively reweighted least squares. *American Statistician* **42**, 152–154.

Tzavidis, N., Marchetti, S. & Chambers, R. (2010). Robust prediction of small area means and distributions. *Australian and New Zealand Journal of Statistics* **52**, 167–186

Ugarte, M.D., Goicoa, T. A., Militino, A.F. & Durban, M. (2009). Spline Smoothing in small area trend estimation and forecasting. *Computational Statistics and Data Analysis* **53**, 3616–3629.

Venables, W.N. & Ripley, B.D. (2002). *Modern Applied Statistics with S.* Springer, NewYork.

Yu, D.L. & Wu, C. (2004). Understanding population segregation from Landsat ETM+imagery: a geographically weighted regression approach. *GISience and Remote Sensing* **41**, 145–164.

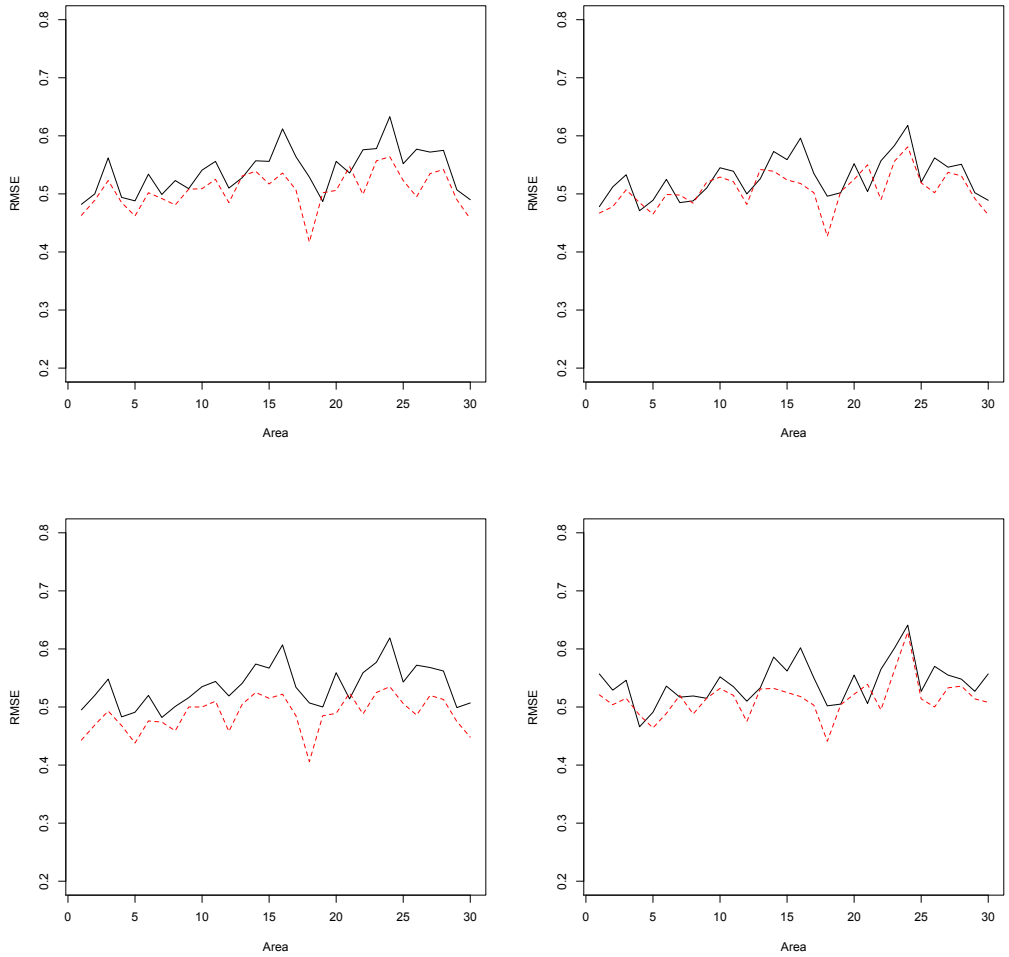


Fig. 1 Area-specific values of actual model-based RMSE (black solid line) and average estimated RMSE (red dashed line) under Chi-squared stationary (top) and non-stationary (bottom) scenarios. Top and bottom left is MQGWR version of (14) with MSE estimated using (17). Top and bottom right is the MQGWR-LI version of (14) with MSE also estimated using (17).

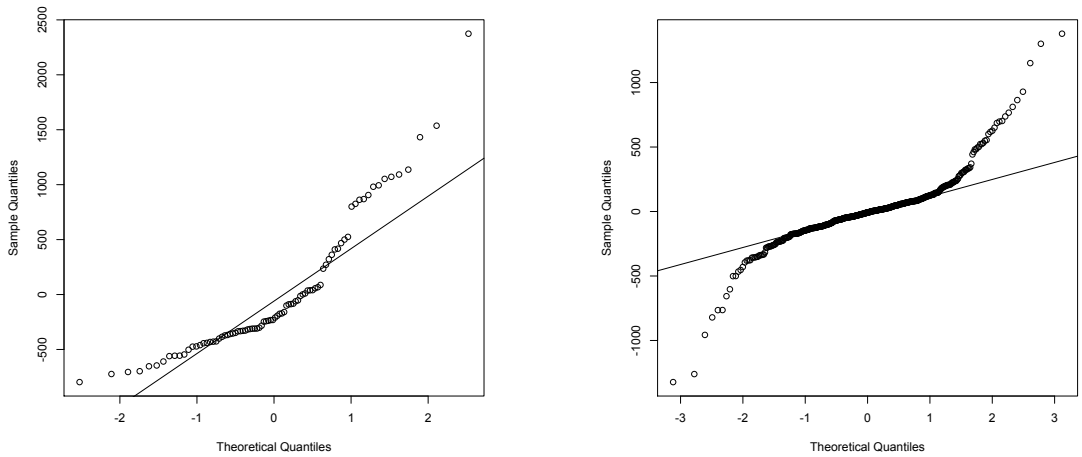
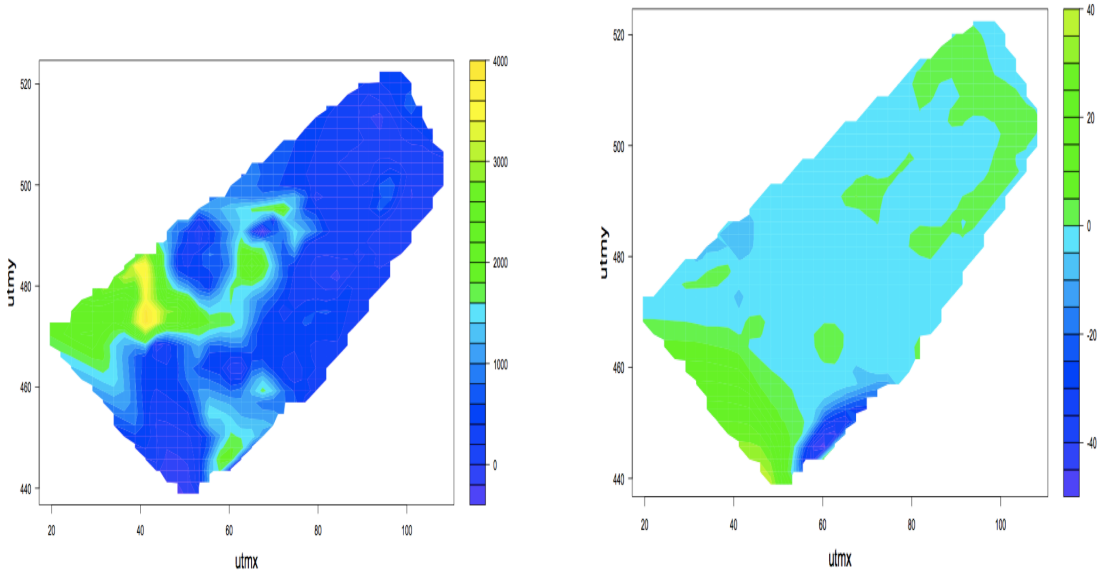


Fig. 2 Normal probability plots of level 2 (left) and level 1 residuals (right) derived by fitting a two level linear mixed model to the synthetic population data.



Intercepts Slopes

Fig. 3 Maps showing the spatial variation in the HUC specific intercept and slope estimates that are generated when the GWR model is fitted to the EMAP data.

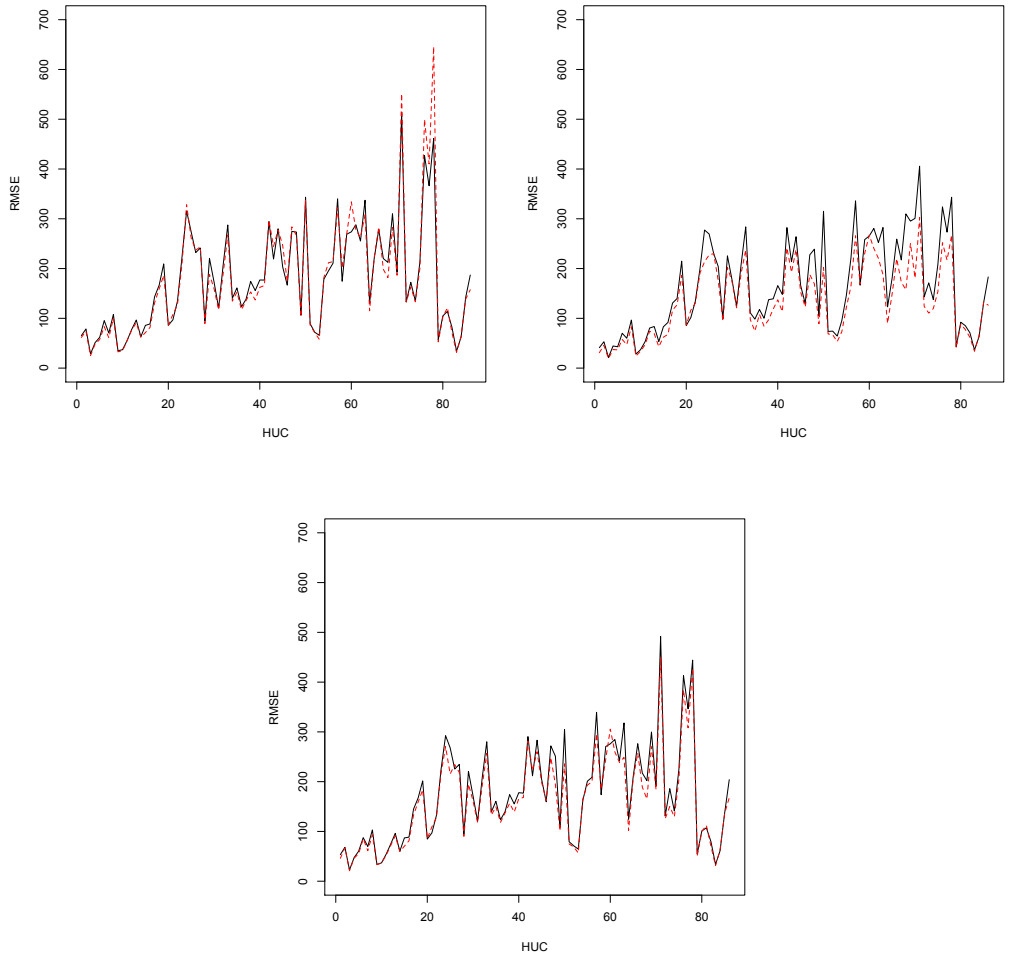


Fig. 4 HUC-specific values of actual design-based RMSE (black solid line) and average estimated RMSE (red dashed line). Top left is the M-quantile predictor (5) with MSE estimator (6) suggested by Tzavidis et al. (2010). Top right is MQGWR version of (14) with MSE estimated using (17) and bottom is the MQGWR-LI version of (14) with MSE also estimated using (17).

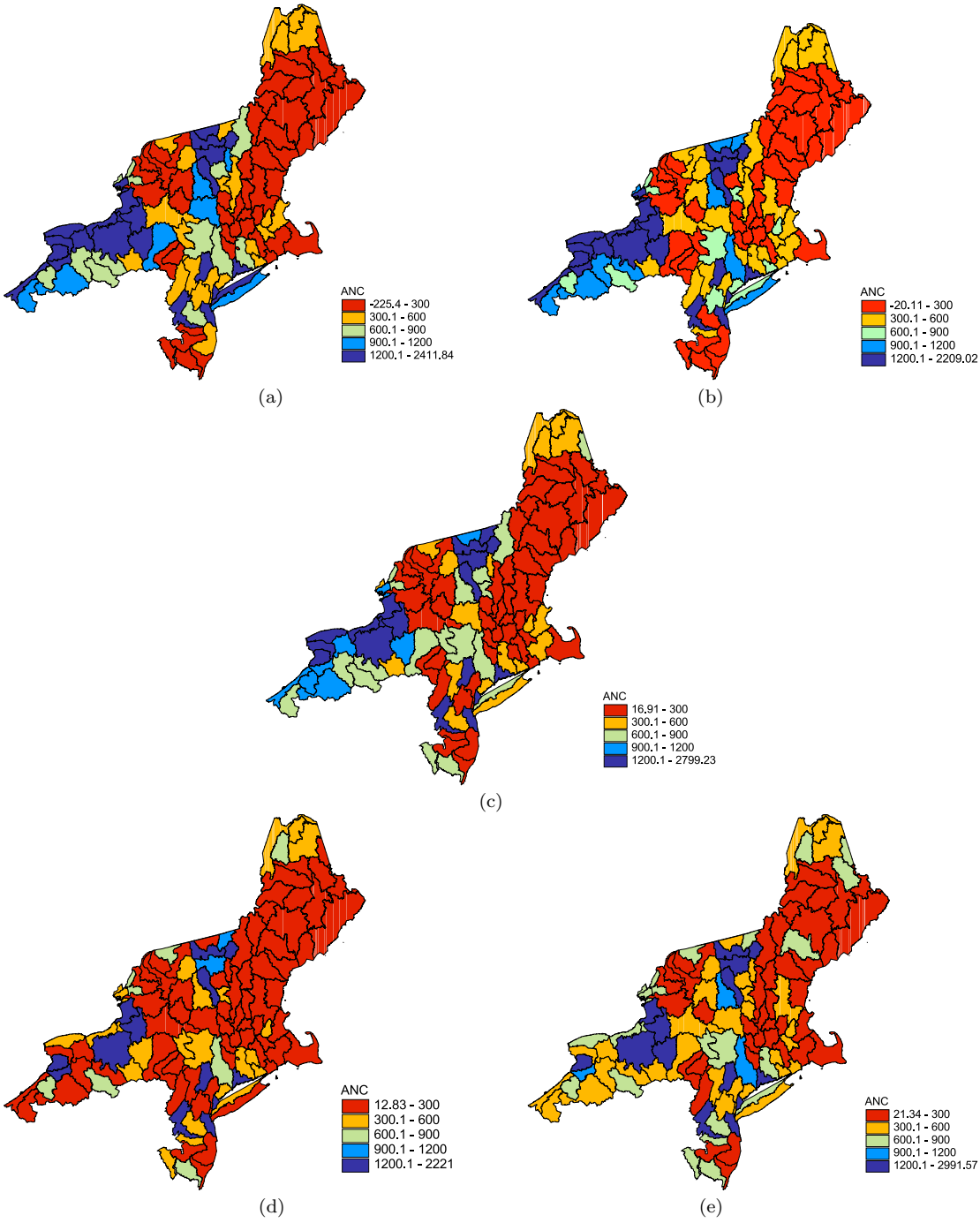


Fig. 5 Maps of estimated average ANC for all 113 HUCs. Map (a) shows estimates computed using (14) and the MQGWR model (9), map (b) presents estimates computed using (14) and the expectile GWR model, map (c) shows estimates computed using (14) and the MQGWR-LI model (12), map (d) shows estimates computed using (5) and the stationary M-quantile model (3) and finally map (e) shows estimates computed using (2) and the linear mixed model.

Table 1 Across areas distribution of Bias and RMSE over simulations.

		Summary of across areas distribution					
Predictor	Indicator	Min	Q1	Median	Mean	Q3	Max
		Stationary process, Gaussian errors					
EBLUP	Bias	-0.051	-0.034	0.001	-0.001	0.023	0.068
	RMSE	0.068	0.075	0.079	0.081	0.087	0.101
MQ	Bias	-0.015	-0.003	0.001	-0.001	0.003	0.012
	RMSE	0.074	0.083	0.088	0.087	0.091	0.100
MQGWR	Bias	-0.016	-0.007	-0.003	-0.002	0.005	0.008
	RMSE	0.067	0.084	0.088	0.087	0.091	0.100
Expectile GWR	Bias	-0.034	-0.015	-0.003	-0.003	0.006	0.046
	RMSE	0.071	0.081	0.084	0.088	0.092	0.112
MQGWR-LI	Bias	-0.010	-0.005	0.001	-0.001	0.003	0.012
	RMSE	0.073	0.085	0.087	0.086	0.090	0.097
		Non-stationary process, Gaussian errors					
EBLUP	Bias	-0.034	-0.013	-0.003	-0.002	0.011	0.031
	RMSE	0.169	0.193	0.205	0.220	0.238	0.323
MQ	Bias	-0.036	-0.011	0.000	-0.002	0.009	0.015
	RMSE	0.164	0.181	0.188	0.188	0.193	0.219
MQGWR	Bias	-0.047	-0.013	-0.005	-0.004	0.005	0.027
	RMSE	0.083	0.092	0.098	0.098	0.103	0.119
Expectile GWR	Bias	-0.059	-0.026	0.001	-0.006	0.010	0.053
	RMSE	0.076	0.088	0.093	0.097	0.104	0.130
MQGWR-LI	Bias	-0.065	-0.010	-0.005	-0.004	0.007	0.047
	RMSE	0.088	0.097	0.107	0.112	0.114	0.186
		Stationary process, Chi-squared errors					
EBLUP	Bias	-0.441	-0.097	0.075	-0.011	0.112	0.237
	RMSE	0.399	0.457	0.482	0.489	0.511	0.651
MQ	Bias	-0.063	-0.043	-0.021	-0.011	0.014	0.062
	RMSE	0.437	0.496	0.526	0.522	0.542	0.598
MQGWR	Bias	-0.075	0.002	0.035	0.028	0.060	0.113
	RMSE	0.482	0.507	0.539	0.539	0.564	0.633
Expectile GWR	Bias	-0.241	-0.071	0.008	-0.010	0.068	0.134
	RMSE	0.421	0.502	0.564	0.562	0.611	0.741
MQGWR-LI	Bias	-0.067	-0.009	0.009	0.010	0.032	0.062
	RMSE	0.471	0.500	0.525	0.528	0.552	0.618
		Non-stationary process, Chi-squared errors					
EBLUP	Bias	-0.069	-0.046	-0.021	-0.014	0.008	0.069
	RMSE	0.465	0.541	0.560	0.566	0.592	0.675
MQ	Bias	-0.086	-0.048	-0.015	-0.014	0.021	0.051
	RMSE	0.460	0.540	0.554	0.555	0.586	0.641
MQGWR	Bias	-0.083	-0.009	0.022	0.017	0.050	0.124
	RMSE	0.482	0.507	0.534	0.535	0.562	0.619
Expectile GWR	Bias	-0.295	-0.125	-0.012	-0.021	0.067	0.176
	RMSE	0.437	0.505	0.540	0.561	0.611	0.757
MQGWR-LI	Bias	-0.085	-0.018	0.004	0.007	0.041	0.080
	RMSE	0.466	0.518	0.541	0.542	0.557	0.641

Table 2 Design-based simulation results using the EMAP data. Results show across areas distribution of Relative Bias (RB) and Relative Root Mean Squared Error (RRMSE) over simulations.

		Summary of across areas distribution					
Predictor	Indicator	Min	Q1	Median	Mean	Q3	Max
				86 sampled HUCs			
EBLUP	RB(%)	-23.31	0.39	10.79	12.55	21.43	83.22
	RRMSE(%)	14.20	23.95	35.18	38.05	49.49	99.00
MQ	RB(%)	-11.09	-2.34	-0.42	-0.83	1.32	4.79
	RRMSE(%)	6.64	25.81	35.49	39.45	49.71	119.07
MQGWR	RB(%)	-8.87	-1.69	0.06	0.22	1.79	14.40
	RRMSE(%)	4.97	21.49	29.84	33.61	43.22	83.71
Expectile GWR	RB(%)	-38.52	-8.50	-0.40	-1.17	6.85	40.33
	RRMSE(%)	2.58	18.24	28.65	26.32	35.37	89.22
MQGWR-LI	RB(%)	-8.87	-2.24	-0.71	-0.78	0.85	7.20
	RRMSE(%)	5.17	23.86	34.03	35.64	46.22	81.46
				27 non-sampled HUCs			
EBLUP	RB(%)	-72.50	-57.29	-36.59	-2.47	38.14	288.11
	RRMSE(%)	5.75	40.14	53.76	60.44	62.21	288.61
MQ	RB(%)	-85.57	-73.27	-66.29	-47.46	-31.32	106.96
	RRMSE(%)	6.56	37.63	68.65	57.26	74.83	107.69
MQGWR	RB(%)	-49.98	-11.89	-3.69	-3.37	4.88	40.61
	RRMSE(%)	10.21	14.88	17.50	22.93	23.29	78.84
Expectile GWR	RB(%)	-30.51	-12.12	-3.36	-1.60	5.91	58.49
	RRMSE(%)	8.79	15.28	18.35	24.62	28.29	70.01
MQGWR-LI	RB(%)	-58.30	-38.59	-23.21	-23.13	-11.58	17.87
	RRMSE(%)	13.09	22.43	26.82	30.85	40.13	58.78