Centre for Statistical and Survey Methodology

The University of Wollongong

Working Paper

09-09

Small Area Estimation Via M-quantile Geographically Weighted Regression

Nicola Salvati, Nikos Tzavidis, Monica Pratesi and Ray Chambers

Centre for Statistical & Survey Methodology, University of Wollongong, Wollongong NSW 2522. Phone +61 2 4221 5435, Fax +61 2 4221 4845. Email: anica@uow.edu.au

Small Area Estimation Via M-quantile Geographically Weighted Regression

Nicola Salvati

Dipartimento di Statistica e Matematica Applicata all'Economia, Università di Pisa

e-mail: salvati@ec.unipi.it

Nikos Tzavidis

Centre for Census and Survey Research, University of Manchester

e-mail: Nikos.Tzavidis@manchester.ac.uk

Monica Pratesi

Dipartimento di Statistica e Matematica Applicata all'Economia, Università di Pisa

e-mail: m.pratesi@ec.unipi.it

Ray Chambers

Centre for Statistical and Survey Methodology, School of Mathematics and Applied Statistics, University of Wollongong

e-mail: ray@uow.edu.au

Abstract. One popular approach to small area estimation when data are spatially correlated is to employ Simultaneous Autoregressive (SAR) random effects models to define the Spatial Empirical Best Linear Unbiased Predictor (SEBLUP). See Singh *et al.* (2005) and Pratesi and Salvati (2008). SAR models allow for spatial correlation in the error structure. An alternative approach that incorporates the spatial information in the regression model is to use Geographically Weighted Regression (GWR). See Brunsdon *et al.* (1996) and Fotheringham *et al.* (1997). GWR extends the traditional regression model by characterising the relationship between the outcome variable and the covariates via local rather than global parameters. In this paper we investigate GWR-based small area estimation under the M-quantile modelling approach (Chambers and Tzavidis, 2006). In particular, we integrate the concepts of outlier-robust small area estimation and borrowing strength over space within a unified modelling framework by specifying 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 an outlier-robust predictor of the small area characteristic of interest that also accounts for spatial association in the data. An additional important spin-off from applying the M-quantile GWR small area model is 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 simulation, with the latter based on a realistic survey data set. The paper concludes with an application to environmental data for predicting average levels of the Acid Neutralizing Capacity at 8-digit Hydrologic Unit Code level in the Northeast states of the U.S.A.

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

1. Introduction

Sample survey data are extensively used for providing reliable direct estimates of totals and means for the survey population. However, reliable estimates for domains are also usually 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 resolved 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. Indirect estimators for small areas are often based on unit level random effects models, and the Best Linear Unbiased Predictor (BLUP) is typically defined under the unit level random effects model that assumes independence of the random area effects. A detailed description of this predictor and of its empirical version (EBLUP) can be found in Rao (2003, Chap. 7), Rao (2005) and Jiang and Lahiri (2006). Chambers and Tzavidis (2006) describe an alternative approach to SAE that is based on regression M-quantiles. This approach avoids conventional Gaussian assumptions and 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 small area estimation models.

In economic, environmental and epidemiological applications, observations that are spatially close may be more related than observations that are further apart. This spatial correlation can be accounted for 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). The application of SAR models in small area estimation enables researchers to borrow strength over space and hence potentially improve the precision of small area estimates. In this context, Singh *et al.* (2005) and Pratesi and Salvati (2008) have proposed the use of the Spatial Empirical Best Linear Unbiased Predictor (SEBLUP).

SAR models allow for spatial correlation in the error structure. An alternative approach for incorporating spatial information in the regression model is by assuming that the regression coefficients themselves vary spatially across the geography of interest. Geographically Weighted Regression (GWR) (Brunsdon *et al.*, 1996; Fotheringham *et al.*, 1997; 2002; Yu and Wu, 2004) extends the traditional regression model by allowing local rather than global parameters to be estimated. That is, GWR directly models spatial non-stationarity in the mean structure of the model. In this paper we explore the use of GWR in small area estimation based on the M-quantile modelling approach. In doing so we first 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 model is then used to define a predictor of the small area characteristic of interest (here we focus on small area means) that accounts for spatial

association in the data. The M-quantile GWR small area model integrates the concepts of outlier-robust small area estimation and borrowing strength over space within a unified modeling framework. In this context, Richardson and Welsh (1995) and Richardson (1997) have investigated outlier-robust inference for the linear mixed model and Sinha and Rao (2009) have proposed an outlier robust version of the small area EBLUP. However, we are not aware of any related extension to outlier-robust small area estimation under the SAR model or under another model that borrows strength over space. An additional important spin-off from applying the M-quantile GWR small area model appears to be more efficient predictors for out of sample areas.

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 to define the M-quantile GWR model. In section 4 we show how the M-quantile GWR model can be utilised for small area estimation. In section 5 we discuss mean squared error estimation for small area predictors defined under the M-quantile GWR model. In section 6 we present a series of model-based and design-based simulation studies for assessing the performance of the different small area predictors considered in this paper. In section 7 we use data from the U.S. Environmental Protection Agency's Environmental Monitoring and Assessment Program (EMAP) to predict average levels of the Acid Neutralizing Capacity at 8-digit Hydrologic Unit Code (HUC) level in the Northeast states of the U.S.A. Finally, in section 8 we summarize our main findings.

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, each containing a known number N_j of units, with the value $\mathbf{x}_{ij}$ of a vector x of p auxiliary variables known for each unit i in small area j and with the value y_{ij} for the variable of interest y known for each unit in the sample. We assume that $\mathbf{x}_{ij}$ contains 1 as its first component (so the model includes an intercept). The overall sample size is n, with the sample size in area j equal to n_j (this can be zero). 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

$$y_{ij} = \mathbf{x}_{ij}^T\boldsymbol{\beta} + z_{ij}\gamma_j + \varepsilon_{ij},\ i = 1, \ldots, n_j, j = 1, \ldots, d, \qquad (1)$$

where ε_{ij} is an individual random effect, γ_j denotes a random area effect and z_{ij} is an auxiliary 'contextual' variable whose value is known for all units in the population. The role of the γ_j in (1) is to

characterise differences in the conditional distribution of y given x between the small areas. The empirical best linear unbiased predictor (EBLUP) of m_j (Henderson, 1975; Rao, 2003, Chapter 7) is then

$$\hat{m}_j^{MX} = N_j^{-1}\left(\sum_{i\in s_j} y_i + \sum_{i\in r_j}\left\{\mathbf{x}_i^T\hat{\boldsymbol{\beta}} + z_i\hat{\gamma}_j\right\}\right) \quad (2)$$

where s_j denotes the n_j sampled units in area j, r_j denotes the remaining $N_j - n_j$ units in the area and $\hat{\boldsymbol{\beta}}$, $\hat{\gamma}_j$ are defined by substituting an optimal estimate of the covariance matrix of the random effects in (1) into the best linear unbiased estimator of $\boldsymbol{\beta}$ and the best linear unbiased predictor (BLUP) of γ_j respectively. A widely used approach to mean squared error (MSE) estimation of the EBLUP is based on the approach taken by Prasad and Rao (PR) (1990). This estimator accounts for the variability due to the estimation of the random effects, regression parameters and variance components.

In recent years there has been growing interest in methods that incorporate the spatial structure of the data in small area estimation. A popular approach does this by fitting a SAR model to the random area effects in (1). In matrix form the resulting model can be expressed as

$$\mathbf{y} = \mathbf{X}^T\boldsymbol{\beta} + \mathbf{Z}\left(\mathbf{I} - \rho\mathbf{W}\right)^{-1}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}, \quad (3)$$

where $\mathbf{I}$ is a $d\times d$ identity matrix, $\mathbf{Z}$ is a $n\times d$ matrix of known positive constants, the $\mathbf{W}$ matrix describes the neighbourhood structure of the small areas and ρ defines the strength of the spatial relationship between the random effects of neighbouring areas.

The application of SAR models in small area estimation enables researchers to borrow strength over space and hence potentially improve the precision of small area estimates. In this context, Petrucci and Salvati (2004), Singh *et al.* (2005) and Pratesi and Salvati (2008) have proposed the use of the Spatial Empirical Best Linear Unbiased Predictor (SEBLUP):

$$\hat{m}_j^{MX/SAR} = N_j^{-1}\left(\sum_{i\in s_j} y_i + \sum_{i\in r_j}\left\{\mathbf{x}_i^T\hat{\boldsymbol{\beta}} + z_i\hat{v}_j\right\}\right) \quad (4)$$

where $\hat{v}_j = \mathbf{b}_j^T\hat{\mathbf{G}}\mathbf{Z}^T\hat{\mathbf{V}}\left(\mathbf{y} - \mathbf{X}^T\hat{\boldsymbol{\beta}}\right)$, $\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\hat{\mathbf{V}}\mathbf{X}\right)^{-1}\mathbf{X}^T\hat{\mathbf{V}}\mathbf{y}$, $\hat{\mathbf{V}} = \hat{\sigma}_\varepsilon^2\mathbf{I}_n + \mathbf{Z}\hat{\mathbf{G}}\mathbf{Z}^T$ ($\mathbf{I}_n$ is an $n\times n$ identity matrix), $\hat{\mathbf{G}} = \hat{\sigma}_\gamma^2\left[\left(\mathbf{I} - \hat{\rho}\mathbf{W}^T\right)\left(\mathbf{I} - \hat{\rho}\mathbf{W}\right)\right]^{-1}$, $\hat{\sigma}_\gamma^2$, $\hat{\sigma}_\varepsilon^2$ and $\hat{\rho}$ are asymptotically consistent estimators of the parameters obtained by Maximum Likelihood (ML) or Restricted Maximum Likelihood (REML) estimation and $\mathbf{b}_j^T$ is $1\times d$ vector $\left(0,0,\ldots,1,\ldots,0\right)$ with value 1 in the *j*-th position. These authors derive a MSE estimator for the SEBLUP following the results of Kackar and Harville (1984), Prasad and Rao

(1990) and Datta and Lahiri (2000). Note that due to the introduction of the additional parameter ρ the last term of this MSE estimator, which measures the uncertainty arising from the estimation of the variance components, is not the same as in the case of the PR MSE estimator for the EBLUP. In practical applications, an approximately unbiased estimator of the MSE of SEBLUP can be obtained following the results of Harville and Jeske (1992) and Zimmerman and Cressie (1992) (Singh *et al.*, 2005; Pratesi and Salvati, 2008). Molina *et al.* (2008) recently proposed a computationally intensive parametric and nonparametric bootstrap-based estimator of the MSE of the Spatial Fay-Herriot model. These bootstrap procedures can be extended to the case of the unit level SAR model.

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 m_q that satisfies

$$\int \psi_q\left(\frac{y-m_q}{\sigma_q}\right)dF(y)=0$$

where $\psi_q(\varepsilon)=\{(1-q)I(\varepsilon<0)+qI(\varepsilon\geq 0)\}\psi(\varepsilon)$ and ψ is an appropriately chosen influence function. Here σ_q is a suitable measure of the scale of the random variable $Y-m_q$. Note that when $\psi(\varepsilon)=\varepsilon$ we obtain the expectile of order *q*, which represents a quantile-like generalization of the mean, while when $\psi(\varepsilon)=\text{sgn}(\varepsilon)$ 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 and Bassett, 1978; Newey and Powell, 1987). More generally, Breckling and Chambers (1988) define a linear M-quantile regression model as one where the M-quantile $Q_q(x;\psi)$ of order *q* of the conditional distribution of y given x corresponding to an influence function ψ satisfies

$$Q_q(\mathbf{x}_{ij};\psi)=\mathbf{x}_{ij}^T\boldsymbol{\beta}_\psi(q). \quad (5)$$

For specified *q* and continuous ψ, an estimate $\hat{\boldsymbol{\beta}}_\psi(q)$ of $\boldsymbol{\beta}_\psi(q)$ can be obtained via an iterative weighted least squares algorithm. Asymptotic theory for this estimator follows directly from well-known M-estimation results and is set out in section 2.2 of Breckling and Chambers (1988). The M-quantile coefficient q_i of population unit *i* was introduced by Kokic *et al.* (1997) and is 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|x)* of *y* given *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 and 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 a particular small area can be expected to have similar M-quantile coefficients. Instead of using parametric random effects, these authors therefore propose the use of area-level M-quantile coefficients, i.e. suitably averaged area-specific unit-level M-quantile coefficients, for characterising area differences.

In particular, when (5) holds, with $\boldsymbol{\beta}_\psi(q)$ a sufficiently smooth function of q, they suggest a predictor of m_j of the form

$$\hat{m}_j^{MQ} = N_j^{-1}\left[\sum_{i\in s_j} y_i + \sum_{i\in r_j} \hat{Q}_{\hat{\theta}_j}(\mathbf{x}_i;\psi)\right] \qquad (6)$$

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 (5) at *q*. When there is no sample in the area, we can form a 'synthetic' M-quantile predictor by setting $\hat{\theta}_j = 0.5$.

Tzavidis *et al.* (2008) refer to (6) as the 'naive' M-quantile predictor and note that it can be biased. To rectify this problem these authors propose a bias adjusted M-quantile predictor of m_j of the form

$$\hat{m}_j^{MQ/CD} = \int_{-\infty}^{\infty} t d\hat{F}_j(t) = N_j^{-1}\left[\sum_{i\in U_j} \hat{Q}_{\hat{\theta}_j}(\mathbf{x}_i;\psi) + \frac{N_j}{n_j}\sum_{i\in s_j}\left\{y_i - \hat{Q}_{\hat{\theta}_j}(\mathbf{x}_i;\psi)\right\}\right], \qquad (7)$$

where $U_j = s_j \cup r_j$. Note that the superscript CD in (7) refers to the fact that it is based on evaluating the area *j* expected value functional defined by integrating with respect to the area *j* version of the distribution function estimator proposed by Chambers and Dunstan (1986). Tzavidis *et al.* (2008) note that, under simple random sampling within the small areas, predictor (7) can also be derived from the design-consistent and model-consistent estimator of the finite population distribution function proposed by Rao, Kovar and Mantel (1990). Due to the bias correction in (7), this predictor will have higher variability and so will be most effective when the naïve estimator (6) is expected to have substantial bias, e.g. when the functional form of (5) is incorrectly specified. An alternative approach to dealing with the bias-variance trade off in (7) in such a situation is to limit the variability of the bias correction term in (7) by using robust (huberized) residuals instead of raw residuals. A predictor of this type is described in Tzavidis *et al.* (2008). An estimator of the mean squared error of (7) was proposed in Tzavidis *et al.* (2008). See also Chambers *et al.* (2007) for a detailed discussion of this approach.

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.

Given n observations at a set of L locations $\{u_l; l=1,...,L; L \le n\}$, with n_l data values $\{y_{il}, \mathbf{x}_{il}; i=1,...,n_l\}$ 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

$$y_{il} = \mathbf{x}_{il}^T \boldsymbol{\beta}(u_l) + \varepsilon_{il} \qquad (8)$$

where $\boldsymbol{\beta}(u_l)$ is a $(p \times 1)$ vector of 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

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

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

$$w(u_l, u) = \exp\left(-d_{u_l,u}^2 / 2b^2\right), \qquad (9)$$

where $d_{u_l,u}$ denotes the Euclidean distance between u_l and u and b 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 also be used. 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. For the purposes of this paper we use a global (i.e. single) bandwidth whose value is optimally defined by a cross validation criterion (Fotheringham *et al.*, 2002):

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

where $\hat{y}_{il}\left(b\right)$ is the fitted value of y_{il} using bandwidth b. The value of b that minimizes CV_a is then selected. An alternative approach is to use optimal local bandwidths. However, this significantly increases the computational intensity of the model fitting process.

The GWR model (8) is a linear model for the conditional expectation of y given x at location u. That is, this model characterises the local behaviour of the conditional expectation of y given x as a linear function of *x*. However, a more complete picture of the relationship between y and x at location u can be constructed by specifying a model for the conditional distribution of y given x at this location. Since the M-quantiles of a distribution serve to characterise it, such a model can be defined by extending (5) to specify a linear model for the M-quantile of order q of the conditional distribution of y given x at location u, writing

$$Q_q(\mathbf{X};\psi,u) = \mathbf{X}^T\boldsymbol{\beta}_\psi(u;q) \qquad (10)$$

where now $\boldsymbol{\beta}_\psi(u;q)$ varies with u as well as with *q*. Like (8), (10) can be interpreted as a local linear approximation, in this case to the (typically) non-linear order *q* M-quantile regression function of *y* on *x*, thus allowing the entire conditional distribution (not just the mean) of y given x to vary non-linearly from location to location. The parameter $\boldsymbol{\beta}_\psi(u;q)$ in (8) at an arbitrary location u can be estimated by solving

$$\sum_{l=1}^{L} w(u_l,u)\sum_{i=1}^{n_l}\psi_q\left(y_{il} - \mathbf{x}_{il}{}^T\boldsymbol{\beta}_\psi(u;q)\right)\mathbf{x}_{il} = 0\,. \qquad (11)$$

where $\psi_q(\varepsilon) = 2\psi(s^{-1}\varepsilon)\left\{qI(\varepsilon>0) + (1-q)I(\varepsilon\le 0)\right\}$. Here 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. $s = median\left|y_{il} - \mathbf{x}_{il}{}^T\boldsymbol{\beta}_\psi(u;q)\right| / 0.6745$ and we will typically assume a Huber Proposal 2 influence function, $\psi(\varepsilon) = \varepsilon I(-c\le\varepsilon\le c) + c\,\mathrm{sgn}(\varepsilon)I(|\varepsilon|>c)$. Provided c is bounded away from zero, we can solve (11) by combining the iteratively re-weighted least squares algorithm used to fit the 'spatially stationary' M-quantile model (5) 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 (11) can be written as

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

An R function (R Development Core Team, 2004) that implements an iterative re-weighted least squares algorithm for solving this equation is available from the authors. The steps in it 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

$$\boldsymbol{\beta}_{\psi}^{t}(u;q) = \left\{\mathbf{X}^{T}\mathbf{W}^{*(t-1)}(u;q)\mathbf{X}\right\}^{-1}\mathbf{X}^{T}\mathbf{W}^{*(t-1)}(u;q)\mathbf{y}. \qquad (12)$$

   Here $\mathbf{y}$ 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 very small value.

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

Street *et al.* (1988) proposed an estimator of the covariance matrix of a 'standard' M-estimator of the regression parameters. 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 (10) is over-parameterised as it allows for both local intercepts and local slopes. An alternative spatial extension of the M-quantile regression model (5) that has a smaller number of parameters is one that combines local intercepts with global slopes and is defined as

$$Q_{q}(\mathbf{X};\psi,u) = \mathbf{X}^{T}\boldsymbol{\beta}_{\psi}(q) + \delta_{\psi}(u;q). \qquad (13)$$

Here $\delta_{\psi}(u;q)$ is a real valued spatial process with zero mean function over the space defined by locations of interest. The model (13) 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 (5). Denote this estimate by $\hat{\boldsymbol{\beta}}_{\psi}(q)$. At the second step we use geographic weighting to estimate $\delta_{\psi}(u;q)$ via

$$\hat{\delta}_{\psi}(u;q) = n^{-1}\sum_{l=1}^{L} w(u_l,u)\sum_{i=1}^{n_l}\psi_q\left(y_{il} - \mathbf{x}_{il}^T\hat{\boldsymbol{\beta}}_{\psi}(q)\right). \quad (14)$$

Choosing between (10) and (13) 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 (13) is a special case of (10), the solution to (11) will have less bias and more variance than the solution to (14). Hereafter we refer to (10) and (13) 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 (11) using iterative reweighted least squares, without any assumption about the underlying conditional distribution of *y* given *x* at each location *u*. That is, the approach is distribution-free. Of course, if this conditional distribution is known, and can be appropriately parameterised by ω, say, 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 $\beta_{\psi}(u,q)$ in (8) is then defined by solving the estimating equation

$$\sum_{l=1}^{L} w(u_l,u)\sum_{i=1}^{n_l}\int \psi_q\left(y - x_{il}^T\beta_{\psi}(u,q)\right)dF(y|x_{il},u;\hat{\omega}) = 0$$

where $w(v,u)$ is the spatial weighting function of interest, e.g. (9), and $F(y|x,v;\omega)$ is the conditional distribution of *y* given *x* at location *v*. A related question concerns the conditions under which the estimating equation (11) corresponds to a maximum likelihood scoring equation. Clearly, this will only be the case when $x^T\beta_{\psi}(u,q)$ is a parameter of the conditional distribution $F(y|x,u;\omega)$ with the derivative of the corresponding log density equal to $\psi_q\left(y - x^T\beta_{\psi}(u,q)\right)x$. For a normal conditional distribution, ψ equal to the identity function and $q = 0.5$ this condition is satisfied. Similarly, when ψ is the sign function and the conditional distribution is Asymmetric Laplace, Koenker (2004) shows that (11) 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 phenomenon is called *quantile crossing* and may be due to model misspecification, collinearity or the presence of outlying values. A consequence then 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 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 these lines relative to the median regression line. This approach can be easily

adapted to fitting M-quantile and M-quantile GWR models as follows. Note that we restrict our definition ourselves to a single covariate x_1 below. However, the extension to multiple covariates is straightforward. We assume without loss of generality that ε has median 0 and $|\varepsilon|$ has median 1. The restricted M-quantile GWR fit for the covariate value x_u at location u is then obtained by:

1. Computing the residuals $\varepsilon_{il} = y_{il} - \hat{Q}_{0.5}(x_{1il};\psi,u)$ relative to the M-quantile GWR fit of order q = 0.5 at location u;
2. Regressing the absolute values $r_{il} = |\varepsilon_{il}|$ of these residuals on the covariate values x_{1il} 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 $\sum_{l=1}^{L} w(u_l,u) \sum_{i=1}^{n_l} \psi_q\left(\varepsilon_{il} - \kappa_q \hat{r}_{il}\right) = 0$. 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 $\hat{Q}_q(x_1;\psi,u) = \hat{Q}_{0.5}(x_1;\psi,u) + \kappa_q(u)\hat{r}(x_u)$ where $\hat{r}(x_u)$ is the value of $\hat{r}_{il}$ at $x_{il} = x_u$.

In the empirical results reported later 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.

4. USING M-QUANTILE GWR MODELS IN SMALL AREA ESTIMATION

As mentioned in Section 1, SAR models allow for spatial correlation in the error structure. Alternatively, this spatial information can be incorporated directly into the regression structure via an M-quantile GWR model. In this section we describe how this can be achieved. In addition to the assumptions 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 for example with geo-referenced data. The aim is to use these data to predict the area j mean m_j of y using the M-quantile GWR models (10) and (13).

Following Chambers and 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 (5) is described in Chambers and Tzavidis (2006) and can

be used directly with (13). We adapt this approach to the GWR M-quantile model (10) by first defining a fine grid of q values in the interval (0,1). Chambers and Tzavidis (2006) use a grid that ranges between (0.01 to 0.99) with step 0.01. We employ the same grid definition and then use the sample data to fit (10) 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 $\hat{Q}_{q_i}(\mathbf{x}_i;\psi,u_i)=y_i$.

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.* (2008), the bias-adjusted M-quantile GWR predictor of the mean m_j in small area j is then

$$\hat{m}_j^{MQGWR/CD}=N_j^{-1}\left[\sum_{i\in U_j}\hat{Q}_{\hat{\theta}_j}(\mathbf{x}_i;\psi,u_i)+\frac{N_j}{n_j}\sum_{i\in s_j}\left\{y_i-\hat{Q}_{\hat{\theta}_j}(\mathbf{x}_i;\psi,u_i)\right\}\right] \quad (15)$$

where $\hat{Q}_{\hat{\theta}_j}(\mathbf{x}_i;\psi,u_i)$ is defined either via the MQGWR model (10) or via the MQGWR-LI model (13).

Variants of the M-quantile GWR model (10) can be defined by changing the value of the tuning constant c in the Huber Proposal 2 influence function. For example, an expectile version of the M-quantile GWR model can be fitted by substituting a large positive value for the tuning constant c in this influence function. Empirical comparisons of the 'large c' (i.e. expectile) and the more robust 'small c' Huber-type M-quantile small area models are reported in Chambers and Tzavidis (2006).

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 mixed model (1) the synthetic mean predictor for out of sample area j is $\hat{m}_j^{MX/SYNTH}=N_j^{-1}\sum_{i\in U_j}\mathbf{x}_i^T\hat{\boldsymbol{\beta}}$. Under the M-quantile model (5) the synthetic mean predictor for out of sample area j is $\hat{m}_j^{MQ/SYNTH}=N_j^{-1}\sum_{i\in U_j}\hat{Q}_{0.5}\left(\mathbf{x}_i;\psi\right)$. We note that with synthetic estimation all variation in the area-specific predictions comes from the area-specific auxiliary information. One way of potentially improving the conventional synthetic estimation for out of sample areas is by using a model that borrows strength over space such as SAR random effects model and M-quantile GWR model. In this last case a synthetic-type mean predictor for out of sample area j is defined by

$$\hat{m}_j^{MQGWR/SYNTH}=N_j^{-1}\sum_{i\in U_j}\hat{Q}_{0.5}\left(\mathbf{x}_i;\psi,u_i\right).$$

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 bias-robust estimator of the mean squared error of (7) was proposed in Tzavidis *et al.* (2008), and below we extend their argument to define an estimator of a first order approximation to the mean squared error of (15). A more detailed discussion of this approach to mean squared error estimation is set out in Chambers *et al.* (2007). Here we just note that it is based on (i) a model where the regression of y on x for a particular population unit depends on its location, with this regression specified by the locally linear GWR model (8), and (ii) the fact that estimators derived under the MQGWR model (10) or the MQGWR-LI model (13) can be written as linear combinations of the sample values of y. For example, from (12) we see that (15) can be expressed as a weighted sum of the sample y-values

$$\hat{m}_j^{MQGWR/CD} = N_j^{-1}\mathbf{w}_{sj}^T\mathbf{y}, \quad (16)$$

where

$$\mathbf{w}_{sj} = \frac{N_j}{n_j}\mathbf{1}_{sj} + \sum_{i\in r_j}\mathbf{H}_{ij}^T\mathbf{x}_i - \frac{N_j - n_j}{n_j}\sum_{i\in s_j}\mathbf{H}_{ij}^T\mathbf{x}_i . \qquad (17)$$

Here $\mathbf{1}_{sj}$ 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

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

If we treat the weights defining the linear representation (16) as fixed, and assume that the values of y follow a location specific linear model, e.g. (8), then an estimator of the prediction variance of (16) can be computed following standard methods of heteroskedasticity-robust variance estimation for linear predictors of population quantities (Royall and Cumberland, 1978). Put $\mathbf{w}_{sj} = (w_{ij})$. This estimator is of the form

$$v(\hat{m}_j^{MQGWR/CD}) = \frac{1}{N_j^2}\sum_{k:\, n_k>0}\sum_{i\in s_k}\lambda_{ijk}\left\{y_i - \hat{Q}_{\hat{\theta}_k}(\mathbf{x}_i,\psi,u_i)\right\}^2 \qquad (18)$$

where $\lambda_{ijk} = \left\{(w_{ij}-1)^2 + (n_j-1)^{-1}(N_j-n_j)\right\}I(k=j) + w_{ik}^2 I(k\neq j)$ 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. Since the weights defining

(17) reproduce the small area mean of x, it also follows that (16) 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 (18) then estimates the mean squared error of (16) in this special 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 (18) is an estimator of a first order approximation to the mean squared error of (16).

Note that (18) treats the weights (17) as fixed, i.e. it ignores the contribution to the mean squared error from the estimated area level M-quantile coefficients $\hat{\theta}_j$. Chambers *et al.* (2007) refer to this as a pseudo-linearization assumption since for large overall sample sizes the contribution to the overall mean squared error of (16) arising from the variability of $\hat{\theta}_j$ will be of smaller order of magnitude then the fixed weights prediction variance of (16). As a consequence (18) will tend to be biased low. However, this potential underestimation needs to be balanced against the bias robustness of (18) under misspecification of the second order moments of *y*, and may well lead to this MSE estimator being preferable to other MSE estimators based on higher order approximations that depend on the model assumptions being true. Empirical results reported in Chambers *et al.* (2007) indicate that the MSE estimator (18) for M-quantile predictor performs well both in model-based and design-based studies with small area sample sizes.

6. SIMULATION STUDIES

In this section we present results from simulation studies that were used to examine the performance of the small area estimators discussed in the preceding sections. Two types of simulations were carried out. In section 6.1 we used model-based simulations. That is, at each simulation population data were first generated 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 was then taken from this simulated population according to a pre-specified design. In section 6.2 on the other hand we used design-based simulation. Here real survey data were first used to simulate a population with spatial characteristics and this fixed population was then repeatedly sampled according to a pre-specified design. In our case the survey data came from the Environmental Monitoring and Assessment Program (EMAP) that forms part of the Space Time Aquatic Resources Modelling and Analysis Program (STARMAP) at Colorado State University.

6.1 MODEL-BASED SIMULATIONS

Synthetic population values were generated under two spatial versions of a linear mixed model and two scenarios for the distribution of the random area effects and the individual residuals. Each population was of size $N = 10{,}500$ and contained $d = 30$ equal-sized small areas. More specifically, the first method of simulation generated population values of y and x according to the two-level model $y_{ij} = 1 + 2x_{ij} + \gamma_j + \varepsilon_{ij}$ where $x_{ij} \sim U[0,1]$, $i = 1...350$ and $j = 1...30$, with random effects generated under two scenarios: (a) $\gamma_j \sim N(0, 0.04)$ and $\varepsilon_{ij} \sim N(0, 0.16)$ and (b) $\gamma_j \sim \chi^2(1) - 1$ and $\varepsilon_{ij} \sim \chi^2(3) - 3$, i.e. mean corrected chi-square variates with 1 and 3 degrees of freedom, respectively. The second method of simulation generated population values with random effects simulated under the same scenarios (a) and (b) but in addition allowed the intercept and slope of the linear model for y to vary according to longitude and latitude. In particular, these location coordinates were independently generated as $U[0,50]$ with

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

and

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

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 defines four model-based scenarios (Gaussian stationary, Gaussian non-stationary, Chi-square stationary, Chi-square non-stationary). For each of these scenarios 200 Monte-Carlo populations were generated using the corresponding model specifications. For each generated population and for each area j we selected 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 were then used to estimate the small area means.

Four different types of small area linear models were fitted to these simulated data. These were (i) a random intercepts version of (1) with uncorrelated and correlated random area effects (3), (ii) the linear M-quantile regression specification (5), (iii) the MQGWR model (10), and (iv) the MQGWR-LI model (13). Two types of random intercepts model were used in (i). The first had uncorrelated random area effects and was fitted using the default REML option of the *lme* function (Venables and Ripley, 2002, section 10.3) in R. The second random intercepts model used in (i) had correlated random area effects and was fitted using the SEBLUP function of the SAE package in R (Gomez Rubio, 2006). The M-quantile linear regression model (ii) was fitted using a modified version of the *rlm* function (Venables

and Ripley, 2002, section 8.3) in R and so uses iteratively re-weighted least squares to fit this model (Chambers and Tzavidis, 2006). The MQGWR models in (iii) and (iv) were fitted using a modification of the functions used to fit (ii). The M-quantile regression and the M-qntile GWR models have been fitted using the Huber Proposal 2 influence function with c=1.345. Estimated model coefficients obtained from these fits were then used to compute the EBLUP (2), the Spatial EBLUP (4), the bias-adjusted M-quantile predictor (7), denoted MQ below, and the MQGWR and the MQGWR-LI versions of corresponding bias-adjusted M-quantile predictor (15).

Although a larger number of simulations would have been preferable, this was not feasible due to the computer intensive nature of the model-fitting process. Note that there was 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 the simulations have not been reported here, but they are available from the authors.

Key percentiles of the across areas distributions of the prediction biases and root mean squared errors of these estimators over these 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, as one would expect. The SEBLUP, 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 MQGWR and MQGWR-LI predictors are considerably more efficient than the MQ, EBLUP and SEBLUP predictors. Under Chi-squared random effects this performance is unchanged, although here the absolute differences in performance between the various predictors is much smaller. Finally, in Table 2 we show key percentiles of the across area distributions of the area level true and estimated mean squared errors (the latter based on (18) and averaged over the simulations) of the MQGWR and MQGWR-LI predictors, as well as the corresponding area level coverage rates for nominal 95 per cent prediction intervals. In general the proposed mean squared error estimator (18) provides a good approximation to the true mean squared error. These results also show that when M-quantile GWR fits are used in (18), then this estimator underestimates the true mean squared error of the corresponding predictor, leading to some undercoverage of prediction intervals. This is consistent with both the MQGWR and the MQGWR-LI models overfitting the actual population regression function. However, this bias is not excessive, being more pronounced in the case of the MQGWR model.

Note that the construction of confidence intervals for small area parameters requires careful consideration. In our simulations we used the MSE estimation method described in section 5 to generate 'normal theory' confidence intervals based on M-quantile model-based estimators. Similarly, we used the approach of Prasad and Rao (1990) to estimate the MSE of the EBLUP and to then construct similar confidence intervals based on this estimator, while the SEBLUP version of the PR

MSE estimator (see Petrucci and Salvati, 2004; Pratesi and Salvati, 2008) was used to estimate the MSE of the SEBLUP as well as to define corresponding confidence intervals based on it.

This use of estimated MSE to construct normal theory confidence intervals, though widespread, has been criticised, however. Hall and Maiti (2006) and more recently Chatterjee *et al.* (2008) discuss the use of bootstrap methods for constructing confidence intervals for small area parameters since 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. Our aim here, however, is more limited in that we present results on point and mean squared error estimation under different versions of the M-quantile GWR model. Further research on the construction of more accurate confidence intervals under the M-quantile GWR model (perhaps using bootstrap techniques) is left for the future.

6.2 A 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.*, 2001). Between 1991 and 1995, researchers from the U.S. Environmental Protection Agency (EPA) 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 used 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 was 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 order to do this, given the 21,026 lake locations, a synthetic population of ANC individual values were non parametrically simulated using a nearest-neighbour imputation algorithm that retained the spatial structure of the observed ANC values in the EMAP sample data.

The algorithm was defined as follows: (1) we first randomly ordered the non-sampled locations in order to avoid list order bias and gave 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 chose 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 reduced the donor weight of the selected donor location by 1. The synthetic population of ANC values created by this procedure was then kept fixed over the Monte-Carlo simulations.

A total of 200 independent random samples of lake locations were 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 were also not sampled in the simulation study. This resulted in a total sample size of 652 locations selected within the 86 'EMAP' HUCs. The synthetic ANC values at these 652 sampled locations were then noted.

Figure 1 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. Hence, the use of a model that relaxes these assumptions, such as an M-quantile model with a bounded influence function, seems reasonable for these data.

The relative bias (RB) and the relative root mean squared error (RRMSE) of estimates of the mean value of ANC in each HUC were computed for the same four predictors that were also the focus of the model-based simulations. These results are set out in Table 3 and show that the M-quantile GWR predictors have significantly lower bias than the EBLUP and SEBLUP predictors with the MQGWR predictor performing best. 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 SEBLUP and MQGWR predictors performing best. These results indicate that incorporating spatial information in small area estimation via the M-quantile GWR model has promise. The slightly higher relative root mean squared error of the MQGWR predictor (compared to the SEBLUP predictor) can be explained by the bias-variance trade off associated with the use of robust methods. Approaches to tackling this were outlined at the end of section 2. For the non-sampled HUCs the use of the synthetic-type predictors that borrow strength over space, defined in section 4, substantially improve prediction. Figure 2 shows how different mean squared estimators tracked the true mean squared error of the different predictors in this simulation. Here we see that

mean squared estimator described in Tzavidis *et al.* (2008), and its GWR form (18), perform well in terms of tracking the true mean squared error of the M-quantile predictors. Some downward bias of (18) when used with the MQGWR model (10) can be seen, however. This is much less of a problem when (18) is combined with the MQGWR-LI model (13). We also see that the PR estimator of the mean squared error of the EBLUP performs poorly as far as tracking area-specific mean squared error is concerned. This is also the case for the analogous estimator of the mean squared error of SEBLUP, and may be attributed in this case to the violation of the linear mixed model assumptions.

An alternative model specification that could be used with spatial data corresponds to adding location (i.e. longitude and latitude) in the fixed part of the mixed model. To investigate the performance of this extended model specification, we repeated the model-based simulations (stationary and non stationary) and the design-based simulation experiments with latitude and longitude included in the fixed part of the model. The results that were obtained are not reported here but are available from the authors. For the model-based simulations, compared to the results reported in Table 1, the inclusion of the longitude and latitude in the fixed part of the mixed model resulted in: (a) for a stationary process under Gaussian or Chi-squared errors the results remain unchanged, (b) for a non-stationary Gaussian process the RRMSE of the EBLUP is reduced. However, the RRMSE of the MQGWR is still lower that that of the EBLUP and (c) for a non-stationary Chi-squared process the RRMSE of the EBLUP and the MQGWR estimators are similar. For the design-based simulation, the results under the new model specification show that for the 86 sampled HUCs the inclusion of longitude and latitude as covariates in the fixed part of the mixed model has no impact on the performance of the EBLUP. However, for the 27 out of sample HUCs the synthetic estimator using the new mixed model specification performs rather badly. This can be attributed to the differences in the locations of lakes in the sampled and not sampled HUCs, with the non-sampled HUCs mainly located in the southwest of the study region. Consequently, the linear mixed model that includes location and is fitted using the data from sampled HUCs is not a good predictor for locations in non-sampled HUCs. This in turn impacts upon the performance of the synthetic small area estimator defined by the linear mixed model.

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. Application of the Brunsdon *et al.* (1999) ANOVA test for spatial stationarity indicates that the EMAP data are consistent with a process characterised by spatially varying relationships.

Predicted values of average ANC for each HUC were calculated using the M-quantile GWR predictor (15) under the MQGWR model (10) and the MQGWR-LI model (13), 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 was constructed using (9), with bandwidth selected using cross-validation.

Figure 3 shows contour maps of the estimated HUC-specific intercepts and slopes from the fitted MQGWR model (10), i.e. when this model is fitted using the HUC-specific M-quantile coefficients $\hat{\theta}_j$. These maps support the assumption of non-stationarity in the data. Finally, in Figure 4 we show maps of estimated values of average ANC for each HUC using the (a) MQGWR model; (b) the MQGWR-LI model; (c) the spatially stationary M-quantile model (5); (d) the linear mixed model (1) with uncorrelated area effects; and (e) the linear mixed model (3) with correlated random area effects. The maps (a) and (b) corresponding to the two M-quantile GWR models provide similar estimates of average ANC for each HUC and are consistent with the patterns produced by other analyses of the EMAP data using non-parametric models (Opsomer *et al.*, 2008). They are also substantially different from the maps (d) and (e) that show the estimates produced by the EBLUP under the spatially uncorrelated linear mixed model (1) and the SEBLUP under the spatially correlated linear mixed model (3). These are very similar and show lower levels of average ANC (and hence greater risk of water acidification) for the target population of HUCs. Finally, we see that the map (c) produced by the M-quantile model (5) that assumes no spatial correlation shows even lower levels of average ANC, most likely due to the failure of the spatial stationarity assumption in this model when it is applied to the EMAP data.

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-referenced data are available, such as financial and economic statistics, environmental and public health modelling. Like the linear M-quantile regression model of Chambers and 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. R code for fitting the M-quantile GWR small area model is straightforward to develop and is available from the authors. It should be noted, however, that a prospective user of the M-quantile GWR model needs to have access to an appropriate level of spatial information for fitting it. In this paper we present an application to modelling environmental data where detailed spatial information is available for sampled and non-sampled locations. More generally it is not difficult to see that 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 less.

One problem that arises with specifying an M-quantile GWR model is 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 model where both intercept and slope parameters in the model vary spatially and the MQGWR-LI model where only the intercept parameter varies spatially. Further research is necessary in order to develop appropriate diagnostics for deciding between them.

Extending the arguments of Chambers *et al.* (2007) we defined an estimator of a first order approximation to the mean squared error of (15). The results obtained in model-based and in design-based simulation studies are promising but we are aware of its potential underestimation, which must be further researched. However, the bias robustness of (18) under misspecification or failure of the model assumptions is an appealing property. In addition, current research on this topic has already produced empirical results that indicate that the MSE estimator (18) has good design based and model based properties in small area estimation (Chambers *et al.* 2007).

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 non-parametric small area estimation problem as a mixed effect model regression. 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. The use of bivariate p-spline approximations for fitting nonparametric unit level nested error and M-quantile regression models allows for reflecting the spatial variation in the data and then uses these nonparametric models for small area estimation. Further research is necessary to contrast SAR, M-quantile GWR and unit level nested error p-spline regression models in terms of their performance when borrowing strength over space in small area estimation.

ACKNOWLEDGEMENTS

The work in this paper has been 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. 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: Method and models*, Kluwer Academic Publishers, Boston.

Breckling, J. and Chambers, R. (1988) M-quantiles, *Biometrika*, **75**, 4, 761-771.

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

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

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

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

Chambers, R., Chandra, H. and Tzavidis, N. (2007) 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. and Huilin, L. (2008) Parametric Bootstrap Approximation to the Distribution of EBLUP and Related Prediction Intervals in Linear Mixed Models. [*To appear in the Annals of Statistics*]

Cressie, N. (1993) *Statistics for spatial data*, John Wiley & Sons, New York.

Datta, G. S. and Lahiri , P. (2000) A Unified Measure of Uncertainty of Estimates for Best Linear Unbiased Predictors in Small Area Estimation Problem, *Statistica Sinica*, 10, 613-627.

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

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

Gomez Rubio, V. (2006) *Introduction to Small Area Estimation, SAE Package*, R Vignette.

Hall, P. and 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.

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

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

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

Kackar, R. and Harville, D. A. (1984) Approximations for standard errors of estimators for fixed and random effects in mixed models, *Journal of the American Statistical Association*, **79**, 853-862.

Harville, D. A. and Jeske, D. R. (1992) Mean squared error of estimation or prediction under a general linear model, *Journal of the American Statistical Association*, **87**, 724-731.

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

Molina, I., Salvati, N. and M. Pratesi (2008) Bootstrap for estimating the MSE of the Spatial EBLUP, *Computational Stat.*, DOI 10.1007/s00180-008-0138-4.

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

Opsomer, J. D., Claeskens, G., Ranalli, M.G., Kauermann, G. and 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. and Salvati N. (2004) Small Area Estimation: the Spatial EBLUP at area and at 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.G.N. and Rao J.N.K. (1990) The estimation of the mean squared error of small-area estimators, *J. Amer. Statist. Assoc.*, **85**, 163-171.

Pratesi, M. and Salvati, N. (2008). Small Area Estimation: the EBLUP estimator based on spatially correlated random area effects, *Statistical Methods &Applications*, **17**, 114-131.

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. and 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*, John Wiley & Sons, New York.

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

Richardson, A.M. (1997) Bounded Influence Estimation in the Mixed Linear Model, *Journal of the American Statistical Association*, **92**, 154-161.

Richardson, A.M and Welsh, A.H. (1995), Robust Estimation in the Mixed Linear Model, *Biometrics*, **51**, 1429-1439.

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

Singh, B.B., Shukla, G.K. and Kundu, D. (2005) Spatio-Temporal Models in Small Area Estimation, *Survey Methodology*, **31**, *2*, 183-195.

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

Tzavidis, N., Marchetti, S. and Chambers, R. (2008) Robust estimation of small area means and quantiles. To appear in the *Australian and New Zealand Journal of Statistics*.

Ugarte, M.D., Goicoa, T. A., Militino, A.F. and Durban, M. (2009) Spline Smoothing in Small Area Estimation, to appear in *Computational Statistics and Data Analysis*.

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

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

Zimmerman, D. L. and Cressie, N. (1992) Mean squared prediction error in the spatial linear model with estimated covariance parameters, *Ann. Inst. Stat. Math*, **44**, 27-43.

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

Predictor	Indicator	Summary of across areas distribution					
		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
SEBLUP	Bias	-0.065	-0.033	-0.005	-0.001	0.024	0.076
	RMSE	0.064	0.074	0.082	0.082	0.088	0.106
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
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
SEBLUP	Bias	-0.104	-0.018	-0.008	-0.004	0.016	0.096
	RMSE	0.155	0.193	0.208	0.221	0.248	0.321
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
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
SEBLUP	Bias	-0.455	-0.176	0.043	-0.019	0.132	0.275
	RMSE	0.383	0.448	0.475	0.490	0.523	0.613
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
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
SEBLUP	Bias	-0.266	-0.129	0.002	-0.022	0.073	0.215
	RMSE	0.488	0.524	0.551	0.554	0.575	0.656
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
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. Across areas distribution of true (i.e. Monte Carlo) root mean squared errors (True RMSE), area averages of estimated root mean squared errors (Est. RMSE) and area coverage rates (CR%) for nominal 95% prediction intervals.

Predictor	Indicator	Percentile of across areas distribution					
		10	25	median	Mean	75	90
Stationary process, Gaussian errors							
MQGWR	True RMSE	0.080	0.084	0.088	0.087	0.091	0.093
	Est. RMSE	0.076	0.078	0.081	0.081	0.083	0.085
	CR(%)	89.51	90.34	91.72	91.88	93.71	94.48
MQGWR-LI	true RMSE	0.079	0.085	0.087	0.086	0.090	0.090
	Est. RMSE	0.077	0.079	0.082	0.082	0.083	0.086
	CR(%)	90.45	91.13	93.00	92.88	94.50	95.00
Non-stationary process, Gaussian errors							
MQGWR	true RMSE	0.090	0.092	0.098	0.098	0.103	0.106
	Est. RMSE	0.074	0.076	0.078	0.079	0.081	0.084
	CR(%)	84.30	85.00	87.00	87.08	89.38	90.50
MQGWR-LI	true RMSE	0.096	0.097	0.107	0.112	0.114	0.138
	Est. RMSE	0.085	0.088	0.098	0.100	0.103	0.122
	CR(%)	88.50	90.50	91.50	91.25	92.88	93.05
Stationary process, Chi-squared errors							
MQGWR	true RMSE	0.489	0.507	0.539	0.539	0.564	0.577
	Est. RMSE	0.463	0.489	0.507	0.506	0.529	0.542
	CR(%)	85.71	89.10	90.38	90.24	92.15	92.44
MQGWR-LI	true RMSE	0.488	0.500	0.525	0.528	0.552	0.574
	Est. RMSE	0.467	0.486	0.505	0.508	0.528	0.543
	CR(%)	87.00	90.50	91.00	90.88	92.50	93.10
Non-stationary process, Chi-squared errors							
MQGWR	true RMSE	0.494	0.507	0.534	0.535	0.562	0.574
	Est. RMSE	0.448	0.470	0.488	0.488	0.512	0.524
	CR(%)	85.50	88.13	90.00	89.40	91.00	92.05
MQGWR-LI	true RMSE	0.505	0.518	0.541	0.542	0.557	0.588
	Est. RMSE	0.485	0.501	0.515	0.514	0.529	0.537
	CR(%)	88.95	90.63	91.50	91.07	92.38	93.05

Table 3. 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.

Predictor	Indicator	Summary of across areas distribution					
		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
SEBLUP	RB (%)	-16.87	-5.12	2.50	5.27	12.33	62.04
	RRMSE (%)	8.08	20.46	29.01	31.50	38.61	75.44
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
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
SEBLUP	RB (%)	-68.46	-51.05	-27.35	11.80	58.49	345.09
	RRMSE (%)	16.03	37.71	53.81	66.21	68.13	346.34
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 (%)	-48.98	-11.89	-3.69	-3.37	4.88	40.61
	RRMSE (%)	10.21	14.88	17.50	22.93	23.29	78.24
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

Figure 1. 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.

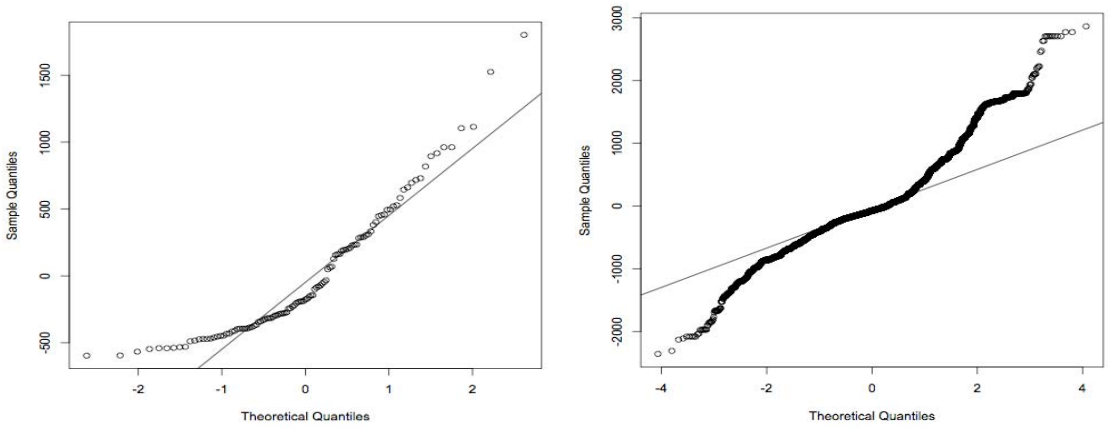

Figure 2. HUC-specific values of actual design-based RMSE (solid line) and average estimated RMSE (dashed line). Top left is the EBLUP predictor (2) with RMSE estimator suggested by Prasad and Rao (1990). Top right is the SEBLUP predictor (4) with RMSE estimator proposed by Petrucci and Salvati (2004). Centre is the M-quantile predictor (7) with RMSE estimator suggested by Tzavidis *et al.* (2008). Bottom left is MQGWR version of (15) with RMSE estimated using (18) and bottom right is the MQGWR-LI version of (15) with RMSE also estimated using (18).

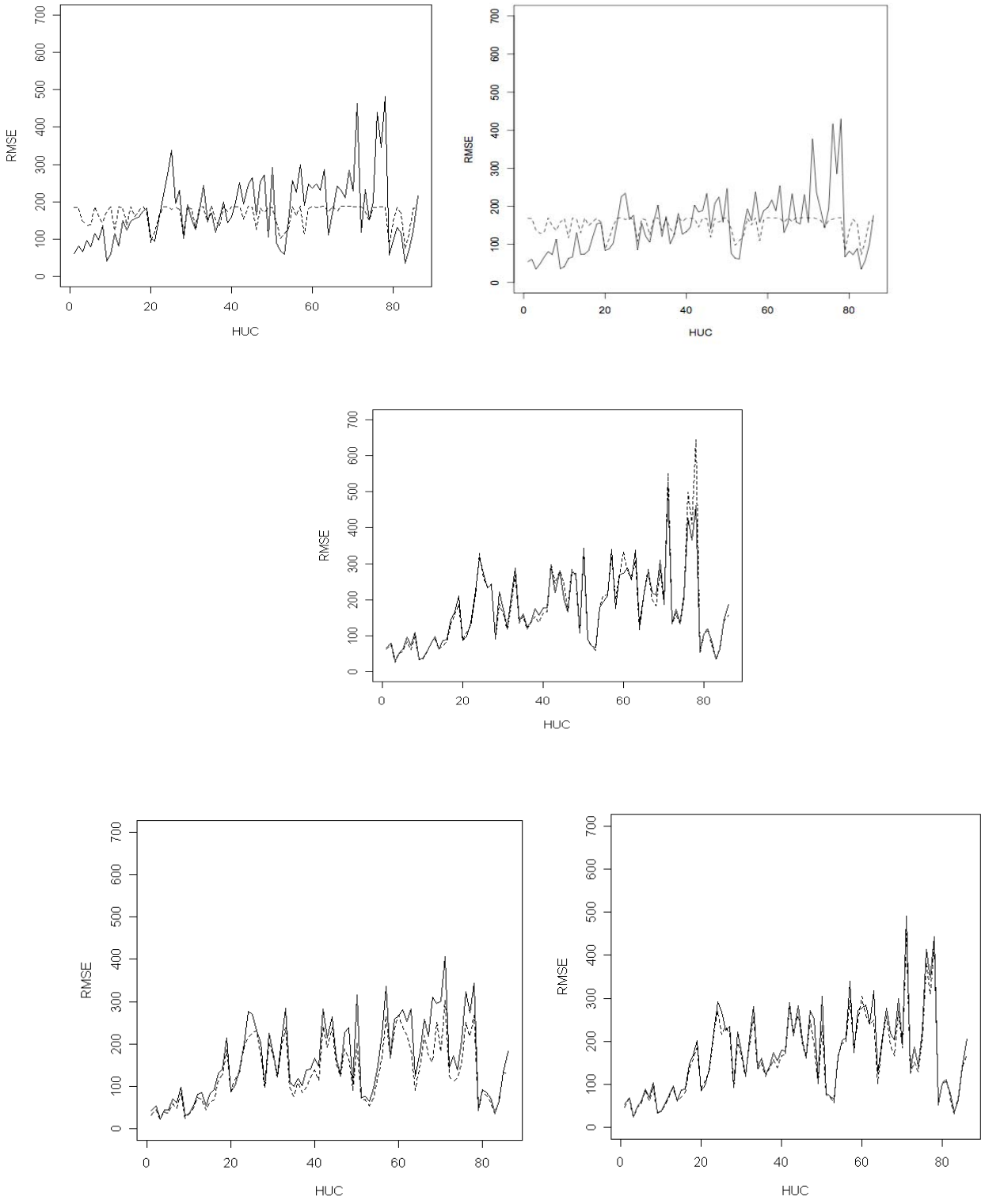

Figure 3. Maps showing the spatial variation in the HUC specific intercept and slope estimates that are generated when the MQGWR model is fitted to the EMAP data.

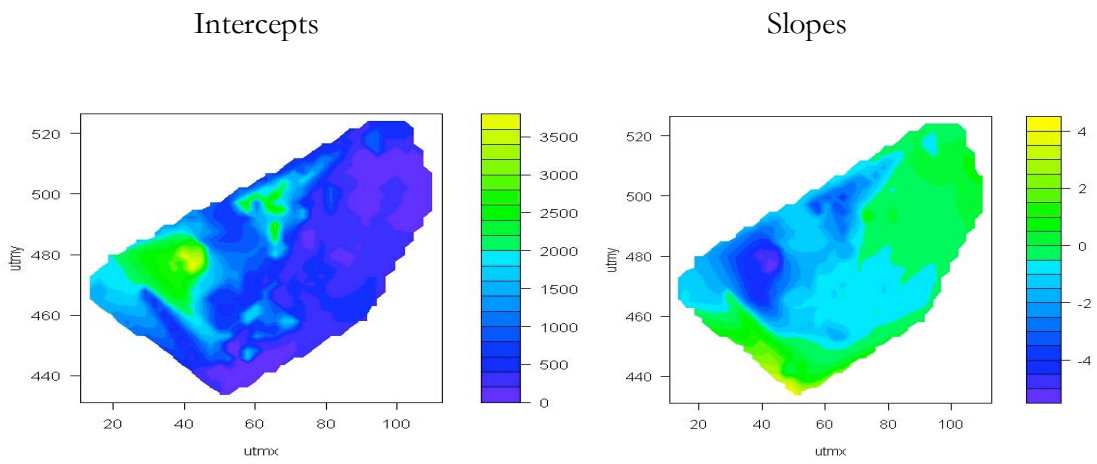

Figure 4. Maps of estimated average ANC for all 113 HUCs. The map (a) shows estimates computed using (15) and the MQGWR model (10), map (b) shows estimates computed using (15) and the MQGWR-LI model (13), map (c) shows estimates computed using (7) and the stationary M-quantile model (5) and finally maps (d) and (e) show estimates computed using (2), (4) and the linear mixed model (1) and (3) with uncorrelated and correlated random area effects.