MANCHESTER 1824

The University of Manchester

Cathie Marsh Centre for Census and Survey Research

Small Area Estimation via M-Quantile Geographically Weighted Regression

CCSR Working Paper 2007-09

Nicola Salvati, Nikos Tzavidis, Monica Pratesi & Ray Chambers
Nikos.tzavidis@manchester.ac.uk

One popular approach to small area estimation when data are spatially correlated is to employ Simultaneous Autoregressive Regression (SAR) random effects models to define an extension to the Empirical Best Linear Unbiased Predictor namely, the Spatial Empirical Best Linear Unbiased Predictor (SEBLUP) (Singh *et al.*, 2005 and Pratesi and Salvati, 2007). SAR models allow for spatial correlation in the error structure. An alternative approach for incorporating the spatial information in the regression model is via Geographically Weighted Regression (GWR) (Brunsdon *et al.*, 1996; Fotheringham *et al.*, 1997). GWR extends the traditional regression model by allowing local rather than global parameters to be estimated. In this paper we investigate the use of GWR in small area estimation based on the M-quantile modelling approach (Chambers and Tzavidis, 2006). In doing so we first propose 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

Small Area Estimation Via M-quantile Geographically Weighted Regression

Nicola Salvati[(1)], Nikos Tzavidis[(2)], Monica Pratesi[(1)] and Ray Chambers[(3)]

ABSTRACT

One popular approach to small area estimation when data are spatially correlated is to employ Simultaneous Autoregressive Regression (SAR) random effects models to define an extension to the Empirical Best Linear Unbiased Predictor namely, the Spatial Empirical Best Linear Unbiased Predictor (SEBLUP) (Singh *et al.*, 2005 and Pratesi and Salvati, 2007). SAR models allow for spatial correlation in the error structure. An alternative approach for incorporating the spatial information in the regression model is via Geographically Weighted Regression (GWR) (Brunsdon *et al.*, 1996; Fotheringham *et al.*, 1997). GWR extends the traditional regression model by allowing local rather than global parameters to be estimated. In this paper we investigate the use of GWR in small area estimation based on the M-quantile modelling approach (Chambers and Tzavidis, 2006). In doing so we first propose 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 predictor of the small area characteristic of interest that accounts for spatial association in the data. An important spin-off from this approach is more efficient

(1) Dipartimento di Statistica e Matematica Applicata all'Economia, Università di Pisa, Via Ridolfi 10, Pisa 56124, Italy e-mail: salvati@ec.unipi.it pratesi@ec.unipi.it

(2) Centre for Census and Survey Research, University of Manchester, Manchester M13 9PL, UK e-mail: Nikos.Tzavidis@manchester.ac.uk

(3) Centre for Statistical and Survey Methodology, School of Mathematics and Applied Statistics, University of Wollongong, Wollongong, NSW 2522, Australia e-mail: ray@uow.edu.au

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

Unit level random effects models are widely used in small area estimation. See Rao (2003) Typically, such models assume independence of random area effects and individual effects. This assumption of unit level independence is also implicit when M-quantile models (Chambers and Tzavidis, 2006) are used in small area estimation. In economic, environmental and epidemiological applications, however, observations that are spatially close may be more related than observations that are further apart. This spatial correlation can be modelled by extending random effects models to allow for spatially correlated area effects, e.g. via a Simultaneous Autoregressive Regression (SAR) random effects model (Anselin, 1992; Cressie, 1993), and Singh *et al*. (2005) and Pratesi and Salvati (2007) have investigated the use of the Spatial Empirical Best Linear Unbiased Predictor (SEBLUP) for small area estimation in this situation.

SAR models allow for spatial correlation in the error structure. An alternative approach to incorporating the spatial information in the regression model is by assuming that the regression coefficients 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 spatially non-stationarity in the mean structure of the outcome variable. 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 the small area mean) that accounts for spatial association in the data. An important spin-off from this approach is more efficient synthetic estimators for out of sample areas.

The structure of the paper is as follows: In section 2 we briefly 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 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. 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 these 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

$$y_{ij} = x_{ij}^T \beta + z_{ij}^T \gamma_j + \varepsilon_{ij}, i = 1, \ldots, n_j, j = 1, \ldots, d, \quad (1)$$

where ε_{ij} is an individual random effect, γ_j is a vector of area level random effects and z_{ij} is a vector of auxiliary 'contextual' variables whose values are 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) is then

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

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 $\hat{\beta}$, $\hat{\gamma}_j$ are defined by substituting an optimal estimate of the covariance matrix of the random effects in (1) into the best linear unbiased estimator of β and the best linear unbiased predictor (BLUP) of γ_j respectively.

An alternative approach to small area estimation is based on the use of M-quantile models (Breckling and Chambers, 1988). A linear M-quantile regression model is one where the q^{th} M-quantile $Q_q(x;\psi)$ of the conditional distribution of y given x satisfies

$$Q_q(x_{ij};\psi) = x_{ij}^T \beta_\psi(q). \tag{3}$$

Here ψ denotes the influence function associated with the M-quantile. For specified q and continuous ψ, an estimate $\hat{\beta}_\psi(q)$ of $\beta_\psi(q)$ is obtained via an iterative weighted least squares algorithm.

The M-quantile coefficient q_i of population unit i was introduced by Chambers and Tzavidis (2006) and is the value q_i such that $Q_{q_i}(x_i;\psi) = y_i$. These authors observed that if variability between small areas is a significant part of the overall variability of the population data, then we expect units from a particular small area to have similar M-quantile coefficients. When (3) holds, with $\beta_\psi(q)$ a sufficiently smooth function of q, they suggested a predictor of m_j of the form

$$\hat{m}_j^{MQ} = N_j^{-1}\left[\sum_{i\in s_j} y_i + \sum_{i\in r_j}\left\{x_i^T \hat{\beta}_\psi(\hat{\theta}_j)\right\}\right] \tag{4}$$

where $\hat{\theta}_j$ 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 $\hat{Q}_{q_i}(x_i;\psi) = y_i$ for q_i. Here $\hat{Q}_q$ denotes the estimated value of (3) at q. When there is no sample in the area $\hat{\theta}_j = 0.5$.

Tzavidis and Chambers (2007) refer to (4) as the 'naive' M-quantile predictor and note that this can be biased. To rectify this problem these authors propose a bias adjusted M-quantile predictor of m_j of the form

$$\hat{m}_j^{MQ/CD} = \int_{-\infty}^{\infty} t d\hat{F}_j(t) = N_j^{-1}\left[\sum_{i\in s_j} y_i + \sum_{i\in r_j} x_i^T \hat{\beta}_\psi(\hat{\theta}_j) + \frac{N_j - n_j}{n_j}\sum_{i\in s_j}\left(y_i - \hat{y}_i\right)\right], \qquad (5)$$

where $\hat{y}_i = x_i^T \hat{\beta}_\psi(\hat{\theta}_j)$. Note that the superscript CD in (5) refers to the fact that it is derived from the expected value functional of the area *j* version of the distribution function estimator proposed by Chambers and Dunstan (1986). Tzavidis and Chambers (2007) note that, under simple random sampling, predictor (5) is also derived from the expected value functional of the area *j* version of the Rao-Kovar-Mantel (1990) distribution function estimator, which is a design-consistent and model-consistent estimator of the finite population distribution function.

3. M-QUANTILE GEOGRAPHICALLY WEIGHTED REGRESSION

In this section we define a spatial extension to 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 $\{y_{il}, x_{il}; i = 1,...,n_l\}$ observed at location u_l, a GWR model is defined as follows

$$y_{il} = x_{il}\beta(u_l) + \varepsilon_{il}. \qquad (6)$$

The value of the regression 'function' $\beta(u)$ at an arbitrary location u is estimated using weighted least squares

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

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 have more weight than those further away. One popular approach to defining such a weighting function puts

$$w(u_l,u)=\begin{cases}\exp\left[1-(d_{u_l,u}/b)^2\right]^2 & if\ d_{u_l,u}\le b\\ 0 & otherwise\end{cases}, \tag{7}$$

where $d_{u_l,u}$ denotes the Euclidean distance between u_l and u and b is the bandwidth, which can be optimally defined using a least squares criterion (Fotheringham *et al*., 2002). In what follows we will use (7) to define the weighting function. However, it should be noted that alternative weighting functions, for example the bi-square function, can also be used.

The GWR model (6) is a model for the conditional expectation of y given x at location u. This is easily generalised to a model for the M-quantile of order q of the conditional distribution of y given x at u by allowing (3) to depend on u. That is, we write

$$Q_q(x;\psi,u)=x^T\beta_\psi(u;q) \tag{8}$$

where now $\beta_\psi(u;q)$ varies with u as well as with q. That is, (8) allows the entire conditional distribution (not just the mean) of y given x to vary from location to location. The parameter $\beta_\psi(u;q)$ in (8) can be estimated by solving

$$\sum_{l=1}^{L}w(u_l,u)\sum_{i=1}^{n_l}\psi_q\left\{y_{il}-x_{il}{}^T\beta_\psi(u;q)\right\}x_{il}=0. \tag{9}$$

where $\psi_q(t)=2\psi(s^{-1}t)\{qI(t>0)+(1-q)I(t\le 0)\}$. Here s is a suitable robust estimate of the scale of the sample y values, e.g. the MAD estimate $s=median\left|y_{il}-x_{il}{}^T\hat{\beta}_\psi(u;q)\right|/0.6745$ and we will typically assume a Huber Proposal 2 influence function, $\psi(t)=tI(-c\le t\le c)+c\,\mathrm{sgn}(t)I(|t|>c)$. Provided c is bounded away from zero, an iteratively re-weighted least squares algorithm that combines the

iteratively re-weighted least squares algorithm used to fit 'spatially stationary' M-quantile model (3) and the weighted least squares algorithm used to fit a GWR model can then be used to solve (9), leading to estimates of the form

$$\hat{\beta}_{\psi}(u;q) = \left\{ X_s^T W_s^*(u;q) X_s \right\}^{-1} X_s^T W_s^*(u;q) y_s . \qquad (10)$$

Here y_s is the vector of n sample y values and X_s is the corresponding matrix of order $n \times p$ of sample x values. The matrix $W_s^*(u;q)$ is a diagonal matrix of order n with entry corresponding to a particular sample observation equal to the product of this observation's spatial weight, which depends on its distance from location u, with the weight that this observation has when the sample data are used to calculate the 'spatially stationary' M-quantile estimate $\hat{\beta}_{\psi}(q)$.

One may argue that (8) is over-parametrised 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 is one that combines local intercepts with global slopes and is defined as

$$Q_q(x;\psi,u) = x^T \beta_{\psi}(q) + \delta_{\psi}(u;q) . \qquad (11)$$

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 (11) is fitted in two steps. At the first step we ignore the spatial structure in the data and estimate $\beta_{\psi}(q)$ directly via the iterative re-weighted least squares algorithm used to fit the standard linear M-quantile regression model (3). Denote this estimate by $\hat{\beta}_{\psi}(q)$. At the second step we use geographic weighting to estimate $\delta_{\psi}(u;q)$ via

$$\hat{\delta}_{\psi}(u;q) = n^{-1} \sum_{l=1}^{L} w(u_l,u) \sum_{i=1}^{n_l} \psi_q \left\{ y_{il} - x_{il}^T \hat{\beta}_{\psi}(q) \right\} . \qquad (12)$$

Choosing between (8) and (11) will depend on the particular situation and whether it is reasonable to believe that the slope coefficient in the M-quantile regression model varies significantly between locations. However, it is clear that since (11) is a special case of (8), the solution to (9) will have less bias and more variance than the solution to (12). Hereafter we refer to (8) and (11) as the MQGWR and MQGWR-LI (Local Intercepts) models respectively.

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

In this section we describe how the spatial extensions of the M-quantile model can be used for small area estimation. In addition to the assumptions made at the start of section 2, we now assume that we have only one population value per location. That is, we can 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 (8) and (11).

Following Chambers and Tzavidis (2006), we first estimate the M-quantile GWR coefficients $\{q_{is}; i \in s\}$ 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 in Chambers and Tzavidis (2006) and can be directly used with (11). We adapt this approach to the GWR M-quantile model (8) by first defining a fine grid of *q* values over the interval (0,1) and then using the sample data to fit (8) 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 x_i at location u_i is finally calculated by interpolating over this grid to find the value q_i such that $Q_{q_i}(x_i; \psi, u_i) = y_i$. In either case, 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*. Following Tzavidis and Chambers (2007), the bias-adjusted M-quantile GWR predictor of the mean m_j in small area *j* is

$$\hat{m}_j^{MQGWR/CD} = N_j^{-1}\left[\sum_{i\in s_j} y_i + \sum_{i\in r_j} \hat{Q}_{\hat{\theta}_j}(x_i;\psi,u_i) + \frac{N_j - n_j}{n_j}\sum_{i\in s_j}\left\{y_i - \hat{Q}_{\hat{\theta}_j}(x_i;\psi,u_i)\right\}\right] \quad (13)$$

where $\hat{Q}_{\hat{\theta}_j}(x_i;\psi,u_i)$ is defined either via the MQGWR model (8) or via the MQGWR-LI model (11).

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 $\hat{m}_j^{MX/SYNTH} = N_j^{-1}\sum_{\iota\in U_j} x_i^T\hat{\beta}$. Under the M-quantile model (5) the synthetic mean predictor for out of sample area *j* is $\hat{m}_j^{MQ/SYNTH} = N_j^{-1}\sum_{i\in U_j} x_i^T\hat{\beta}_\psi(0.5)$. 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 an M-quantile GWR model. In this case a synthetic-type mean predictor for out of sample area *j* is defined by

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

We expect that when a truly spatially non stationary process is present, $\hat{m}_j^{MQGWR/SYNTH}$ will improve the efficiency of the other 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 robust estimator of the mean squared error of (3) was proposed in Tzavidis and Chambers (2007). Here we extend this argument to define an estimator of a first order approximation to the mean squared error of (13) under the MQGWR model (8). Our argument is easily extended to the MQGWR-LI model (11). A more detailed discussion of this approach to mean squared error estimation is set out in Chambers *et al*. (2007). To start we note from (10) that (13) can be expressed as a weighted sum of the sample *y*-values

$$\hat{m}_j^{MQGWR/CD} = N_j^{-1} w_{sj}^T y_s , \tag{14}$$

where

$$w_{sj} = \frac{N_j}{n_j} 1_{sj} + \sum_{i \in r_j} H_{ij}^T x_i - \frac{N_j - n_j}{n_j} \sum_{i \in s_j} H_{ij}^T x_i . \tag{15}$$

Here 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

$$H_{ij} = \left\{ X_s^T W_s^*(u_i; \hat{\theta}_j) X_s \right\}^{-1} X_s^T W_s^*(u_i; \hat{\theta}_j) .$$

Given the linear representation (14), an estimator of a first order approximation to the mean squared error of this predictor can be computed following standard methods of robust mean squared error estimation for linear predictors of population quantities (Royall and Cumberland, 1978). Put $w_{sj} = (w_{ij})$. This estimator is of the form

$$v(\hat{m}_j^{MQGWR/CD}) = \sum_{k: n_k > 0} \sum_{i \in s_k} \lambda_{ijk} \left\{ y_i - \hat{Q}_{\hat{\theta}_k}(x_i, \psi, u_i) \right\}^2 \tag{16}$$

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

6. SIMULATION STUDIES

In this section we present results from simulation studies that were used to examine the performance of the small area estimators that were discussed in the preceding

sections. In section 6.1 we employ model-based simulations in which small area population and sample data were simulated based on different parametric assumptions about the distribution of errors and the spatial structure of the data. In section 6.2 we present a design-based simulation that is based on real survey data 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

Two methods were used to simulate population data. In both, N = 10500 population values of x and y in J = 30 small areas were first simulated. For each area j we then independently selected a simple random sample (without replacement) of size $n_j = 20$, leading to an overall sample size of n = 600. This process was repeated 200 times. The sample values of y and the population values of x obtained in each simulation were used to estimate the small area means.

The first method of simulation generated population values of y and x in small area j according to the two-level model $y_{ij} = 1 + 2x_{ij} + \gamma_j + \varepsilon_{ij}$ where $x_{ij} \sim U[0,1]$, 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) $\gamma_j \sim X^2(1) - 1$ and $\varepsilon_{ij} \sim X^2(3) - 3$. 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 α and slope β 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

$$\alpha = 0.2 \times longitude + 0.2 \times latitude$$

and

$$\beta = -5 + 0.1 \times longitude + 0.1 \times latitude\,.$$

Four different types of small area linear models were fitted to these simulated data. These were (i) a random intercepts version of (1), (ii) the linear M-quantile regression specification (3), (iii) the MQGWR model (8), and (iv) the MQGWR-LI model (11). The random intercepts model used in (i) was fitted using the *lme* function (Venables and Ripley, 2002, section 10.3) in R (R Development Core Team, 2004). 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 (R Development Core Team, 2004). The MQGWR models in (iii) and (iv) were fitted using a straightforward modification of the functions used to fit (ii). Estimated model coefficients obtained from these fits were then used to compute the EBLUP (2), the bias-adjusted M-quantile predictor (5), denoted MQ below, and the MQGWR and the MQGWR-LI versions of corresponding bias-adjusted M-quantile predictor (13).

Biases and root mean squared errors over these simulations, averaged over the 30 areas, are set out in Table 1. For Gaussian random effects and a spatially stationary regression surface, we can see that the EBLUP is the best predictor, 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 MQGWR and MQGWR-LI predictors are considerably more efficient than the MQCD predictor and the EBLUP. Under Chi-squared random effects this relative 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 (16) 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 (16) provides a good approximation to the true mean squared error. These results also show that when M-quantile GWR fits are used in (16), then this estimator provides some underestimation of the true mean squared error of the corresponding predictor that also results in 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.

6.2 A DESIGN-BASED SIMULATION

The actual survey 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 was selected from the population of 21,026 lakes in these states using a random systematic design. The lakes making up this population were grouped according to 113 8-digit Hydrologic Unit Codes (HUCs), of which 64 contained less than 5 observations and 27 did not have any observations. 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.

The aim of the design-based simulation was to compare the performance of different predictors of mean ANC in each HUC. In order to do this, we first created a population of ANC values with similar spatial characteristics to that of the lakes sampled by EMAP. A total of 200 independent random samples were then taken from each HUC that had been sampled by EMAP, with sample sizes set to the greater of five and the original EMAP sample size in the HUC. No samples were taken from HUCs that had not been sampled by EMAP, leading to a total sample size of 652 ANC values from 86 HUCs.

In order to generate a population dataset that had similar spatial structure to that of the EMAP sample data, we allocated ANC values to the non-sampled lakes 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 $w(u_i, u) = \exp\left\{-0.5(d_{u_i,u} / b)^2\right\}$. 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 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 the focus of the model-based simulations. These results are set out in Table 3 and show that the M-quantile GWR predictors are much more efficient

that the EBLUP and M-quantile based predictors that ignore the spatial structure in the data. In particular, we see that for the non-sampled HUCs the use of the synthetic-type predictors that borrow strength over space, defined in section 4, substantially improve prediction. Figure1 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 and Chambers (2007), and its GWR form (16), perform well in terms of tracking the true mean squared error of the M-quantile predictors. Some downward bias of (16) when used with the MQGWR model is reported, however. This is much less of a problem when (16) is combined with the MQGWR-LI model. Finally, we see that the Prasad-Rao estimator of the mean squared error of the EBLUP performs poorly as far as tracking area-specific mean squared error is concerned. This phenomenon has also been reported in other design-based studies (e.g. Chambers *et al*., 2007).

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. Figure 2(a) shows the region of interest and the locations of the sampled lakes. 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 liter). A small ANC value for a lake indicates that it is at risk of acidification. Figure 2(b) shows the distribution of ANC in the EMAP data. This is skewed and may contain influential data points. Furthermore, 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 (13) under the MQGWR model (8) and the MQGWR-LI model (11), 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 (7), with bandwidth selected using cross-validation.

Figure 3 shows contour maps of the estimated HUC-specific intercepts and slopes from the fitted MQGWR model (8), 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 MQGWR model, the MQGWR-LI model, the spatially stationary M-quantile model (3) and the linear mixed model (1). 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*., 2005). There are also substantially different from the estimates produced by the spatially stationary models (1) and (3), which show lower levels of average ANC (and hence greater risk of water acidification) for these HUCs.

8. SUMMARY

In this paper we propose a geographically weighted regression extension to 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 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.

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.

AKNOWLEDGEMENTS

The work in this paper has been supported by project PRIN "Metodologie di stima e problemi non-campionari nelle indagini in campo agricoloambientale" 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 robust mean squared error estimation for linear predictors for domains. *[Paper submitted for publication. A copy is available upon request]*.

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

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.

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

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

Opsomer, J. D., Claeskens, G., Ranalli, M.G., Kauermann, G. and Breidt, F.J. (2005). Nonparametric small area estimation using penalized spline regression. In I. S. U. Department of Statistics (Ed.), Preprint Series.

Pratesi, M. and Salvati N. (2007). Small Area Estimation: the EBLUP estimator based on spatially correlated random area effects. Forthcoming in *Statistical Methods & Applications*.

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.

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. (2006): Spatio-Temporal Models in Small Area Estimation, *Survey Methodology*, **31**, *2*, 183-195.

Tzavidis, N. and Chambers, R. (2007). Robust prediction of small area means and distributions, *CCSR Working Paper 2007-08*, University of Manchester [*available from http://www.ccsr.ac.uk/publications/working/#2007-08*]

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.

Table 1 Median values of Bias and RMSE over areas and simulations.

Predictor	Stationary process		Non-stationary process	
	Bias	RMSE	Bias	RMSE
Gaussian random effects				
EBLUP	0.001	0.079	-0.003	0.205
MQ	0.001	0.088	0.001	0.188
MQGWR	-0.003	0.088	-0.005	0.098
MQGWR-LI	0.001	0.087	-0.005	0.107
Chi-squared random effects				
EBLUP	0.075	0.482	-0.017	0.558
MQ	-0.021	0.526	-0.015	0.554
MQGWR	0.035	0.539	0.022	0.534
MQGWR-LI	0.009	0.525	0.004	0.541

Table 2 Across areas distribution of true (i.e. Monte Carlo) root mean squared errors, area averages of estimated root mean squared errors 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

	true RMSE	0.505	0.518	0.541	0.542	0.557	0.588
MQGWR-LI	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 medians of Relative Bias (RB) and Relative Root Mean Squared Error (RMSE) over areas and simulations.

Predictor	RB (%)	RRMSE (%)
86 sampled HUCs		
EBLUP	8.51	43.41
MQ	-1.15	40.29
MQGWR	-0.25	26.12
MQGWR-LI	-0.69	28.52
27 non-sampled HUCs		
EBLUP	-36.59	53.76
MQ	-66.29	68.65
MQGWR	-3.69	17.50
MQGWR-LI	-3.69	17.51

Figure 1 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 M-quantile predictor (5) with RMSE estimator suggested by Tzavidis and Chambers (2007). Bottom left is MQGWR version of (13) with RMSE estimated using (16) and bottom right is the MQGWR-LI version of (13) with RMSE also estimated using (16).

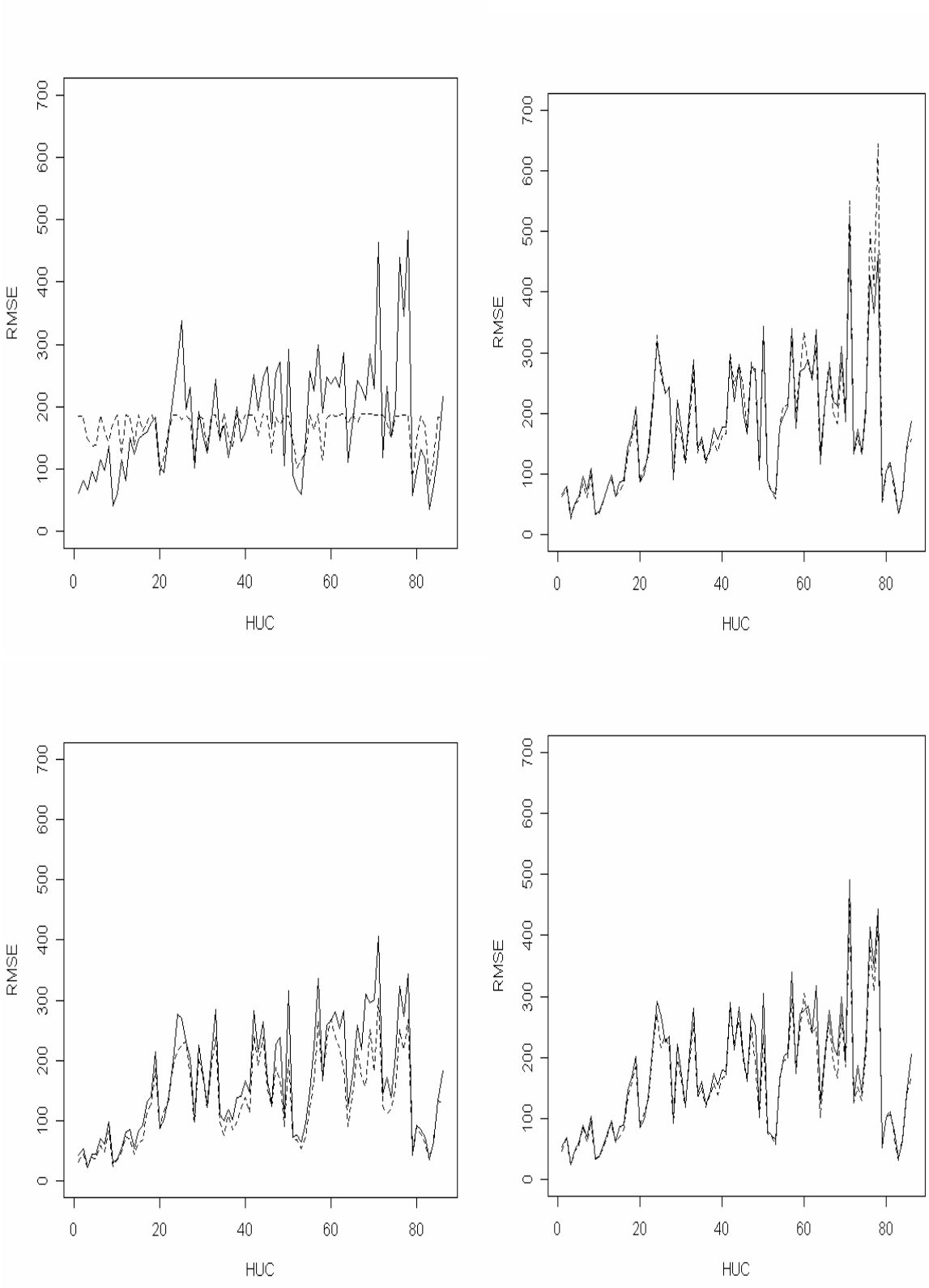


Figure 2 (a) Locations of the sampled lakes in Northeastern U.S.A. (b) Histogram of ANC values in the EMAP data.

(a) (b)

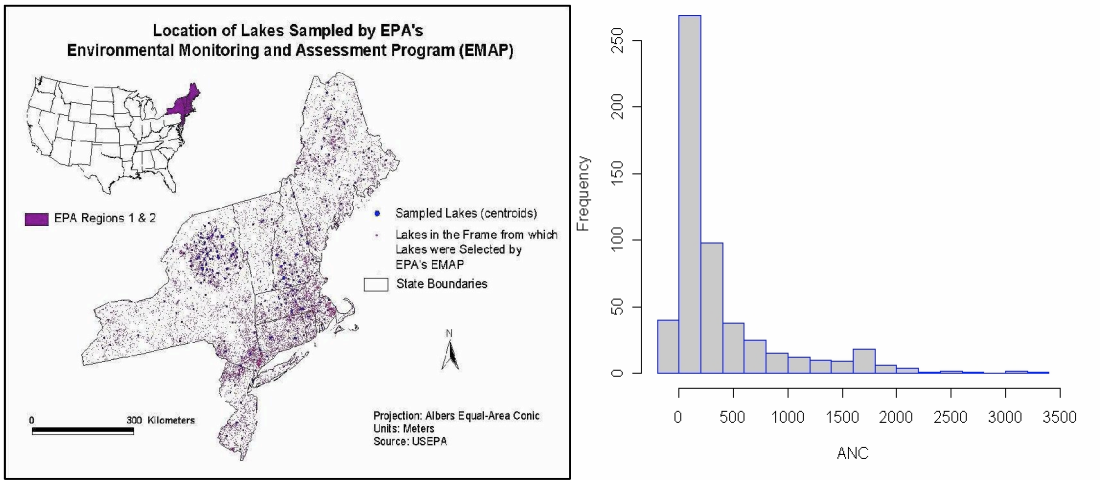


Figure 3 Maps showing the spatial variation in the HUC specific intercept and slope estimates that are generated when the MQGWR model is fitted to the EMAP data.

Intercepts **Slopes**

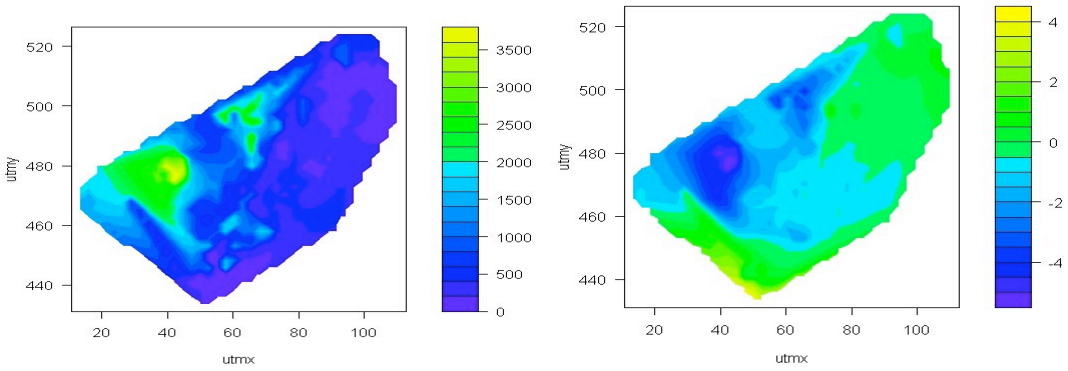
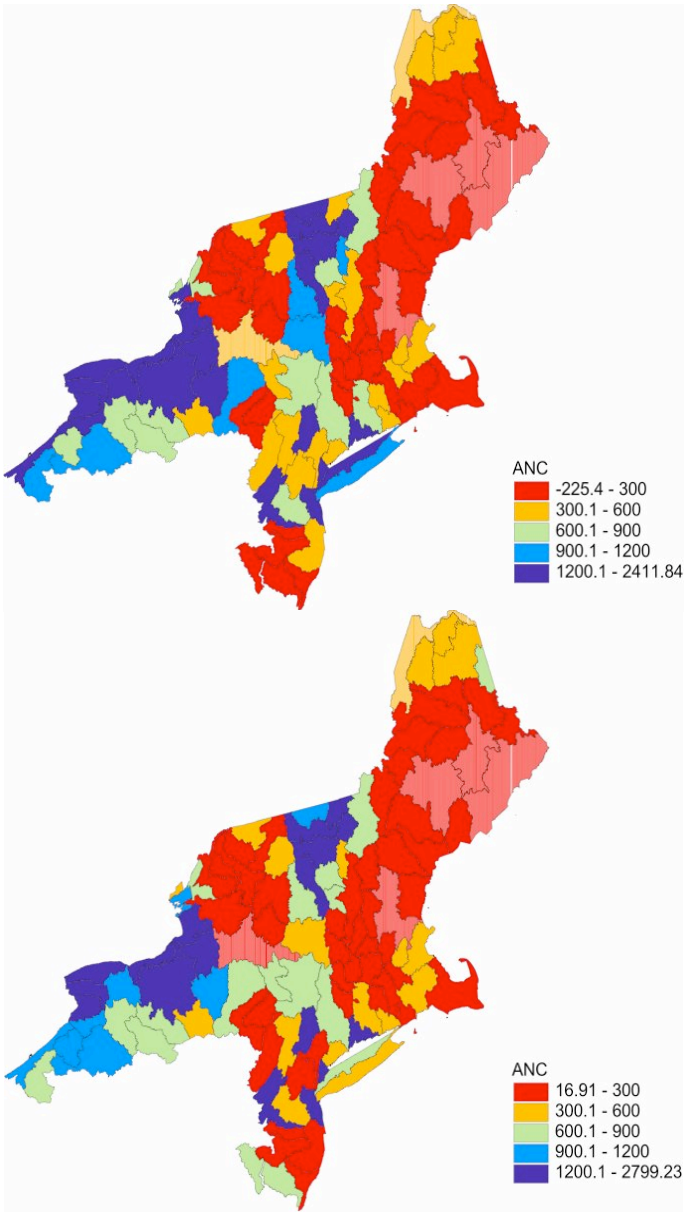
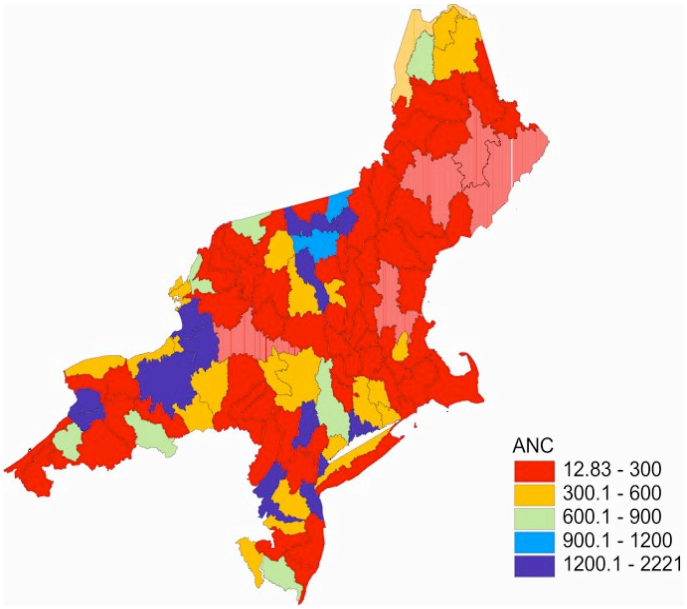
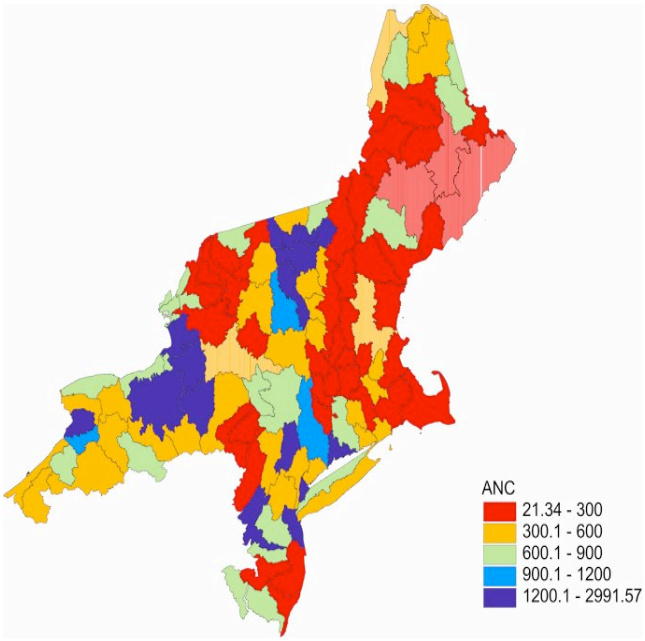


Figure 4 Maps of estimated average ANC for all 113 HUCs. The first map shows estimates computed using (13) and the MQGWR model (8), the second map shows estimates computed using (13) and the MQGWR-LI model (11), the third map shows estimates computed using (5) and the stationary M-quantile model (3) and finally the fourth map shows estimates computed using (2) and the linear mixed model (1).




ANC
12.83 - 300
300.1 - 600
600.1 - 900
900.1 - 1200
1200.1 - 2221


ANC
21.34 - 300
300.1 - 600
600.1 - 900
900.1 - 1200
1200.1 - 2991.57