Università degli Studi di Pisa
Dipartimento di Statistica e Matematica
Applicata all'Economia

Report n. 266

M-quantile Geographically Weighted Regression for Nonparametric Small Area Estimation

Nicola Salvati

Pisa, Maggio 2005

- Stampato in Proprio –

M-quantile Geographically Weighted Regression for Nonparametric Small Area Estimation

Nicola Salvati

Dipartimento di Statistica e Matematica Applicata all'Economia,
Università degli Studi di Pisa, via Ridolfi, 10 – 56124 Pisa
salvati@ec.unipi.it

Abstract: In Small Area Estimation (SAE) often the information included in the geographical location itself of the survey data is not taken into account. However, in many practical fields the data are generally related with the geographical locations where they are observed. Geographically Weighted Regression (GWR) technique is a newly developed statistical methodology that introduces the spatial non-stationarity in the regression model.

The GWR model explains the average behaviour of **Y** given a set of explanatory variables **X**, but it may not be appropriate for modelling the extreme behaviour of **Y** conditional on **X**, as the M-quantile regression models do.

In the paper the M-quantile regression model is extended to include spatially varying regression coefficients as an approach to specific modelling of data which are associated with extreme points in the sample. Following Chambers and Tzavidis (2005), that have developed a new approach to small area estimation based on quantile-like parameters of the conditional distribution of the variable of study given the covariates, we propose to use the M-quantile Geographically Weighted regression for small area estimation.

Keywords: Nonparametric Small area estimation, Geographically Weighted Regression, , M-quantile regression.

1. Introduction

For small domains geographically defined, when traditional area-specific direct estimator does not provide adequate precision it is possible to employ indirect estimators that "borrow strength" from related areas. The indirect estimators can incorporate specific random area effects that account for between areas variation beyond what is explained by auxiliary variables included in the model. Traditionally the random area effects are considered independent, but in practice, basically in most of the applications on environmental data, it should be more reasonable to assume that the random area effects between the neighbouring areas (for instance the neighbourhood could be defined by a contiguity criterion) are correlated and the correlation decays to zero as distance increases (Pratesi and Salvati, 2004; 2005). However, this kind of modelling also depends on strong distributional assumptions, requires a formal specification of the random part of the model and does not easily allow for outliers robust inference.

A new approach to small area estimation based on quantile-like parameters of the conditional distribution of the study variable given the covariates has been recently proposed by Chambers and Tzavidis (Chambers and Tzavidis, 2005). This technique does not depend on strong distributional assumption like the small area models that use both covariates and random effects, and it is robust against outlying area values. Moreover the approach allows for the estimation of the distribution function in each small area and overcomes an important problem in small area estimation that is the impact on the estimates of changing small area geographies.

Nevertheless, any relationship that is not stationary over space will not be represented particularly well by a global model, that produces parameter estimates which represent an average type of behaviour that is not likely to be followed at local level. As a result, the global value of the parameter estimate can be very misleading locally. To overcome this limit, in this paper we propose to specify a local version of the M-quantile regression model, that allows the regression coefficients to vary over the whole map.

The outline of the paper is the following: in Section 1 we recall the basis of the Geographically Weighted Regression (GWR). Section 2 describes the M-quantile regression when it is generalized to include geographical weights (Section 3). Then, the small area prediction process is presented (Section 4) with the focus on the problem of the estimation of

Mean Squared Error of the small area estimators (Section 5). In section 6 some final remarks are reported.

2. Geographically Weighted Regression

In many practical fields such as economics, environmental science and epidemiology, the data are generally related with the geographical locations where they are observed. This type of data are called spatial data. The spatial relationship among data at different locations is usually based on developing neighborhoods and the autocorrelation of locations within neighborhoods. The spatial dependence among data at different locations can be introduced by specifying a linear model with spatially correlated error (Anselin, 1992; Cressie, 1993):

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \tag{1}$$

where **X** is the $n \times p$ matrix of the area specific auxiliary covariates $\mathbf{x}_i = (x_{i1}, x_{i2}, ..., x_{ip})$, $\boldsymbol{\beta}$ is the regression parameters vector $p \times 1$, **e** is the $n \times 1$ vector of the second order variation. Basically there are two approaches to describe the spatial second order variation: Simultaneously Autoregressive models (SAR) and Conditional Autoregressive models (CAR). In spatial regression we assume that the relationship we are modelling holds everywhere in the study area - that is, the regression parameters are "whole-map" statistics. In many situations this is not necessarily the case, as mapping the residuals may reveal. Geographically Weighted Regression (GWR) technique is a newly developed statistical methodology in dealing with spatial non-stationarity among regressed relationship. The technique, originally proposed by Brunsdon *et al.* (1996), has recently received intensive attention (Fotheringham *et al.* 1997 and 2002; Yu and Wu, 2004).

The GWR extends the traditional regression framework by allowing local rather than global parameters to be estimated. The model can be written as:

$$y_i = \sum_{k=1}^{p} \beta_k(u_i, v_i)x_{ik} + e_i \quad i = 1...n \tag{2}$$

where (u_i, v_i) denotes the coordinates of the ith point in space, $\beta_k(u_i, v_i)$ is a realisation of the continuous function $\beta_k(u, v)$ at point i, $x_{i1}, x_{i2}, \ldots, x_{ip}$ are the explanatory variables at location (u_i, v_i) in the studied geographical region, and e_i are error terms. For a given data set, the coefficients are locally estimated by the weighted least squares approach. The weights $w_j(u_i, v_i)$, $j = 1 \ldots n$, at each location (u_i, v_i) are taken as a function of the distance from (u_i, v_i) to other locations where the observations are collected. The parameters at location (u_i, v_i) are estimated by minimizing

$$\sum_{j=1}^{n} w_j(u_i, v_i)\{y_j - \beta_1(u_i, v_i)x_{j1} - \beta_2(u_i, v_i)x_{j2} - \ldots - \beta_p(u_i, v_i)x_{jp}\}^2 . \tag{3}$$

Let

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_1(u_1, v_1) & \beta_2(u_1, v_1) & \cdots & \beta_p(u_1, v_1) \\ \beta_1(u_2, v_2) & \beta_2(u_2, v_2) & \cdots & \beta_p(u_2, v_2) \\ \vdots & \vdots & \vdots & \vdots \\ \beta_1(u_m, v_m) & \beta_2(u_m, v_m) & \cdots & \beta_p(u_m, v_m) \end{bmatrix}, \tag{4}$$

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}, \ \mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \tag{5}$$

and

$$\mathbf{W}(u_i, v_i) = \begin{bmatrix} w_{i1} & 0 & \cdots & 0 \\ 0 & w_{i2} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & w_{in} \end{bmatrix} \tag{6}$$

the estimated parameters at (u_i, v_i) are

$$\hat{\boldsymbol{\beta}}(u_i, v_i) = \left(\mathbf{X}^T \mathbf{W}(u_i, v_i)\mathbf{X}\right)^{-1} \mathbf{X}^T \mathbf{W}(u_i, v_i)\mathbf{Y}. \tag{7}$$

We can note that the weights matrix $\mathbf{W}(u_i, v_i)$ in GWR varies according to the location of point *i*, and the parameters can be estimated for any point in space, even at location where data have not been observed.

The choice of the weights assumes importance in generating the estimated parameters. In spatial analysis, observations close to a location (u_i, v_i) generally exert more influence on the parameter estimates at location (u_i, v_i) than those farther away.

One obvious choice is

$$w_j(u_i, v_i) = \exp\left[-1/2(d_{ij}/b)^2\right] \tag{8}$$

where d_{ij} is the distance between the points *i* and *j* and *b* is called bandwidth. An alternative kernel utilises the bi-square function,

$$w_j(u_i, v_i) = \begin{cases} \exp\left[1 - (d_{ij}/b)^2\right]^2 & \text{if } d_{ij} \le b \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

which is particularly useful because it provides a continuous, near-Gaussian weighting function up to distance *b* from the regression point and then zero weights for any data point beyond *b* (Fotheringham *et al.*, 2002). The bandwidth *b* can be determined by a “least square” criterion:

$$b_0 = \min \Delta(b) = \min \sum_{i=1}^{n} \left[y_i - \hat{y}_{(i)}(b)\right]^2 \tag{10}$$

where $\hat{y}_{(i)}(b)$ is the fitted value of y_i with the observation at location (u_i, v_i) omitted from the fitting process. Other methods of producing spatially varying kernels exist (Fotheringham *et al.*, 2002). The different possible specifications of $w_j(u_i, v_i)$ do not modify the statistical properties of the estimator. Moreover, as a result, the modalities of inclusion of the GWR in M-quantile regression are not modified and so in the economy of this paper the attention is limited to expressions (8) and (9).

3. M-quantile Geographically Weighted Regression

The GWR model explains the average behaviour of **Y** given a set of explanatory variables **X**, but it may not be appropriate for modelling the extreme behaviour of **Y** conditional on **X**. The M-quantile regression (Breckling and Chambers, 1988) is an approach to specific modelling of data which are associated with extreme points in the sample. The M-quantile regression integrates the concepts of quantile regression and expectile regression within a common framework defined by a "quantile-like" generalisation of regression based on influence functions. The M-quantile of order q for the conditional density of **Y** given **X** is defined as the solution $Q_q(\mathbf{X},\psi)$ of the estimating equation $\int \psi_q(\mathbf{Y}-Q)f(\mathbf{Y}\mid\mathbf{X})dy = 0$, where ψ is a specified influence function associated with the M-quantile. A linear M-quantile regression model is:

$$Q_q(\mathbf{X},\psi) = \mathbf{X}^T\boldsymbol{\beta}_\psi(q) \tag{11}$$

and a different set of regression parameters for each value of q can be obtained.

In many economics and environmental study, the data are related with the geographical locations where they are observed. We propose to take into account the spatial information and in particular the spatial non stationarity modelling with spatially varying regression coefficients $\boldsymbol{\beta}_\psi(u_i,v_i;q)$. A linear M-quantile Geographically Weighted Regression (M-quantile GWR) model is then:

$$Q_q(\mathbf{X},\psi) = \mathbf{X}^T\boldsymbol{\beta}_\psi(u_i,v_i;q) \tag{12}$$

where (u_i,v_i) denotes the coordinates of the *i*th point in space and $\boldsymbol{\beta}_\psi(u_i,v_i;q)$ are locally estimated at each location (u_i,v_i). For specific q and ψ, estimates of spatially varying regression parameters at location (u_i,v_i) can be obtained solving the estimating equations:

$$\sum_{j=1}^{n} w_j(u_i,v_i)\psi_q(r_{jq\psi})\mathbf{x}_j = 0 \qquad (13)$$

where $w_j(u_i,v_i)$ are the weights defined as above (6), $r_{jq\psi} = y_j - x_j^T \boldsymbol{\beta}_\psi(u_i,v_i;q)$, $\psi_q(r_{jq\psi}) = 2\psi(s^{-1}r_{jq\psi})\{qI(r_{jq\psi} > 0) + (1-q)I(r_{jq\psi} \leq 0)\}$ and s is a robust measure of spread that is employed in preference to the standard deviation of the residuals. For example, a common approach is to take $s = MAR/0.675$, where MAR is the median absolute residual. Solving the estimating equation is a weighted least-square problem that requires an iterative solution called iteratively reweighted least squares (IRLS). There are a lot of influence functions that can be used and they allow for more flexibility in modelling M-quantile GWR. For example the Huber Proposal 2 influence function, $\psi(u) = uI(-c \leq u \leq c) + c\,\mathrm{sgn}(u)$, is selected (Huber, 1981) because of its optimal and known properties. The Huber function is monotone, bounded, and twice differentiable at 0. Moreover, changing the constant c can be used to trade robustness for efficiency in the M-quantile regression fit. As c decreases ($c \downarrow 0$), the robustness increases whereas efficiency decreases; as c increases ($c \uparrow \infty$), the robustness decreases and the efficiency increases. We can note that if the weights $w_j(u_i,v_i)$ are all equal to 1, the GWR is equal to the traditional regression, and as a consequence, M-quantile GWR corresponds to the traditional M-quantile regression.

4. Estimation of small area parameters

Following Chambers and Tzavidis (2005), that have developed a new approach to small area estimation based on quantile-like parameters of the conditional distribution of the variable of study given the covariates, we propose to use the M-quantile Geographically Weighted regression for small area estimation. Our approach, beyond not depending on strong distributional assumption - like the small area models that use both covariates and random effects - and beyond to being robust against outlying area values, allows to deal with spatial non-stationarity among regressed relationship. Moreover it allows for the estimation of distribution function for each small area and still overcomes the difficulties due to changing small area geographies on the estimates like the M-quantile method.

In the following s denotes a probability sample of size n, where $n=\sum_{k=1}^{m} n_k$ and m is the number of small areas. The estimation procedure requires the knowledge of $n \times p$ matrix of covariates $\mathbf{X}$ and the weights matrix $\mathbf{W}(u_i, v_i)$ for each sampled unit i.

The phases of the estimation procedure are:

1. estimation of the regression parameters of the linear M-quantile GWR model for different values of the quantile of order q specifying the influence function ψ. The regression parameters can be obtained by solving the estimating equations (13).
2. computation of sample M-quantile GWR coefficients, denoted by $\{q_{is}; i \in s\}$. This is done defining a fine grid on the (0,1) interval and using the sample data to fit the M-quantile Geographically Weighted regression line at each value q on this grid. For each unit i, as much regression lines as the number of the values q are obtained. To obtain q_{is} a linear interpolation over this grid is used.
3. computation of the average value of the sample M-quantile GWR coefficient of the unit in area k, $\hat{\bar{q}}_k = \sum_{i \in k} q_{is}$. This is appropriate if $\bar{q}_k$ is defined as the mean value of the population q_i value in area k.
4. In order to estimate the small area parameter (small area total y_k, small area mean $\bar{y}_k$) it needs to estimate $\hat{\boldsymbol{\beta}}_{\psi}(u_i, v_i; \hat{\bar{q}}_k)$, the regression parameters, for each area M-quantile coefficient, $\hat{\bar{q}}_k$. The predictor of small area mean $\bar{y}_k$ assumes the form:

$$\hat{\bar{y}}_k = \frac{1}{N_k}\left(\sum_{i \in s_k} y_j + \sum_{i \in r_k} \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_{\psi}(u_i, v_i; \hat{\bar{q}}_k)\right) \tag{14}$$

where s_k and r_k respectively denote the sampled and non sampled units in area i and N_k is the population size in area k. The unobserved value y_i for population unit $i \in r_k$ is predicted using $\mathbf{x}_i^T \hat{\boldsymbol{\beta}}_{\psi}(u_i, v_i; \hat{\bar{q}}_k)$. Then, for each unit belonging to area k we have different estimates of $\hat{\boldsymbol{\beta}}_{\psi}(u_i, v_i; \hat{\bar{q}}_k)$ depending on the spatial location of the unit i. If the spatial location (u_i, v_i) for population unit $i \in r_k$ is not known, the unobserved value y_i for population unit is predicted using $\mathbf{x}_i^T\hat{\boldsymbol{\beta}}_\psi(\bar{u}_k,\bar{v}_k;\hat{\bar{q}}_k)$ where $(\bar{u}_k,\bar{v}_k)$ are the coordinates of the centroid of the kth small area.

5. Estimation of MSE

The mean squared error of the study parameter can be obtained applying the standard method developed for unbiased weighted linear estimators by Royall and Cumberland (1978). The method has been already applied by Chambers and Tzavidis under the traditional M-quantile model (2005). Under this approach we assume that $\hat{\bar{q}}_k$ is constant. The estimator of the mean squared error is:

$$\hat{M}_k = \hat{V}_k + \hat{B}_k^2 \tag{15}$$

where the estimator of the prediction variance of (14) is either

$$\hat{V}_k = \sum_{i\in s} p_{ik}\left(y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}}(u_i,v_i;0.5)\right)^2 \tag{16a}$$

or

$$\hat{V}_k = \sum_j\sum_{i\in s_j} p_{ik}\left(y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}}(u_i,v_i;\hat{\bar{q}}_k)\right)^2 \tag{16b}$$

with $p_{ik} = N_k^{-2}\left(u_{ik}^2 + I(i\in k)(N_k - n_k)/(n_k-1)\right)$, $\mathbf{u}_k = [u_{ik}] = \mathbf{W}_s(\hat{\bar{q}}_k)\mathbf{X}_s\left(\mathbf{X}_s^T\mathbf{W}_s(\hat{\bar{q}}_k)\mathbf{X}_s\right)^{-1}\mathbf{t}_{rk}$, $\mathbf{W}_s(\hat{\bar{q}}_k)$ is a diagonal matrix that contains the final set of weights produced by IRLS used to compute $\hat{\boldsymbol{\beta}}(u_i,v_i;\hat{\bar{q}}_k)$, and $\mathbf{t}_{rk}$ is the sum of the non-sample covariate values in area k.

The first approach (16a) considers the variance ($Var(y_i)$) to be unconditional on the selected small area and it is called the population level residuals method. Instead, in the area level residuals approach (16b) the variance is interpreted as conditionally on the selected small area j.

The estimator of the conditional bias $\hat{B}_k$ of (14) is given by:

$$\hat{B}_k' = N_k^{-1}\left(\sum_j \sum_{i \in s_j} p_{ik} \mathbf{x}_i^T \hat{\boldsymbol{\beta}}(u_i, v_i; \hat{\hat{q}}_j) - \sum_{i \in k} \mathbf{x}_i^T \hat{\boldsymbol{\beta}}(u_i, v_i; \bar{\hat{q}}_k)\right). \tag{17}$$

6. Final remarks

In this paper we extend the M-quantile regression including spatially varying regression coefficients for small area prediction process. The proposed approach, beyond to having property of M-quantile approach, allows to deal with spatial non-stationarity among regressed relationship. Moreover, a method of estimation of MSE under M-quantile GWR is presented.

The main reason for using M-quantile GWR for small area estimation is to make the best use of the available spatial auxiliary information and to take into account the spatial non stationarity in order to obtain the most efficient estimator possible. There are several reasons why we might expect measurements of relationships to vary over space. An obvious ones relates to sampling variation. A second possible cause is that, for whatever reasons, some relationships are intrinsically different across space. Another cause is that sometimes the parametric model from which the relationships are estimated is a gross misspecification of reality (Fotheringham *et al.*, 2002).

The M-quantile GWR model, like M-quantile regression, offers a way of modelling between area variability of the data without explicitly specifying the random components of the model. Moreover M-quantile GWR captures the not stationary over space via area-specific M-quantile GW coefficients, that depend not only on the $\hat{\hat{q}}_k$, but also on the location $(\bar{u}_k, \bar{v}_k)$ of each small area. Finally, the M-quantile GWR is not influenced by Modifiable Areal Unit Problem (MAUP) issues to the same extent as are more traditional global models.

One possible objection to this technique is that much of the spatial variation in the parameters could be removed by the addition of further explanatory variables. But we can state that the spatial non stationarity can be caused by factors that the model can not take into account.

The M-quantile GWR for small area estimation maintains the same drawbacks of M-quantile regression respect to the traditional small area models: the M-quantile GWR modelling will be less efficient than the mixed modelling when the assumption of the traditional small area modelling are true. Further the method requires the knowledge of appropriate a priori information: the ideal information set includes maps with individual location of sampled and not sampled units. However if the non sampled units coordinates are missing, they can be substituted by the coordinates of the small area centroids. In addition the auxiliary variables must be known at unit level. The ideal set of information seems to be demanding: anyway the Geographical Information Systems (GIS) make it available the geographic locations and at the same time they can manage any administrative files that can provide appropriate individual auxiliary information.

Issues beyond those discussed in this article require further theoretical work. For example, standard regression diagnostic which can be informative in understanding various aspects of model performance have to be investigated. In addition, it has to be explored what happens when some explanatory variables influencing the study variable have global effect while others still maintain their local effect.

Empirical studies are also important to gain further experience with the approach that we propose.

Bibliography

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

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

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

Chambers, R., Tzavidis, N. (2005): M-quantile Models for Small Area Estimation, *Working Paper M05/07, Southampton Statistical Sciences Research Institute, University of Southampton.*

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

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

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

Fox, J. (2002) *An R and S-PLUS Companion to Applied Regression,* Sage Pubblications.

Huber, P. (1981) Robust Statistics, Wiley Series in Probability and Mathematical Statistics, New York.

Pratesi, M., Salvati, N. (2004) Spatial EBLUP in agricultural survey. An application based on Census data, *Working Paper 256, Dipatimento di Statistica e Matemtica Applicata all'Economia, Università di Pisa.*

Pratesi, M., Salvati, N. (2005) Small Area Estimation: the EBLUP estimator with autoregressive random area effects, *Working Paper 261, Dipatimento di Statistica e Matemtica Applicata all'Economia, Università di Pisa.*

Rao, J.N.K. (2003) *Small Area Estimation,* John Wiley & Sons, New York.

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

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