Borrowing strength over space in small area estimation using M-quantile Geographically Weighted models with different levels of geo-referenced information

N. Tzavidis[1] N. Salvati [2]

[1]CCSR, University of Manchester
[2]DSMAE, University of Pisa

Outline

1. Review M-quantile and Geographically Weighted Regression (GWR)
2. Extend GWR to outlier robust GWR by combining M-quantile regression and GWR
3. Employ M-quantile GWR for small area estimation
4. MSE Estimation
5. Empirical Investigations: Design-based simulations and applications to environmental and economic data using the M-quantile GWR model with different levels of geo-referenced information

Methods for Small Area Estimation

- Small area estimation is based on methods that are more commonly known as model-based methods
- The idea is to use statistical models to link the variable of interest with covariate information that is also known for units not in the sample
- A class of models suitable for small area estimation is multilevel (mixed/random effects) models
- An alternative approach to small area estimation is based on quantile/M-quantile models

Mixed Effects Models that Include Random Area Effects

Concept

Include random area-specific effects to account for the between area variation beyond what is explained by the variation in model covariates

Notation: (j =area, i =individual)

- Variable of interest: y_{ij}
- Focus on unit level covariate information: $\mathbf{x}_{ij}$
- Area level random effect: γ_j
- Random error: ϵ_{ij}

$$y_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta} + \mathbf{z}_{ij}^T \gamma_j + \epsilon_{ij}, \;\; i = 1, ..., n_j, \;\; j = 1, ...d$$

Estimator of Small Area Mean

$$\hat{m}_j = N_j^{-1} \Big(\sum_{i \in s_j} y_{ij} + \sum_{i \in r_j} \mathbf{x}_{ij}^T \hat{\boldsymbol{\beta}} + \mathbf{z}_{ij}^T \hat{\gamma}_j \Big)$$

M-quantile Models

- With regression models we model the mean of the variable of interest (y) given the covariates ($\mathbf{x}$)
- A more complete picture is offered, however, by modeling not only the mean of (y) given ($\mathbf{x}$) but also other quantiles. Examples include the median, the 25th, 75th percentiles. This is known as quantile regression
- An M-quantile regression model for quantile q

$$y = \mathbf{x}^T \beta(q) + \epsilon(q)$$

M-quantile Models (Cont'd)

- Conventionally q is a-priori chosen.
- Estimates of $\beta(q)$'s are obtained via Iterative Weighted Least Squares (IWLS) :

$$\hat{\boldsymbol{\beta}}_{\psi}(q) = (\mathbf{x}^T \mathbf{W} \mathbf{x})^{-1} \mathbf{x}^T \mathbf{W} \mathbf{y}$$

- **W** is an n by n diagonal weighting matrix that depends on both the influence function and the quantile we are modeling

Extensions M-quantile Models - Small Area Estimation

- Central Idea: Area effects can be described by estimating an area specific q value ($\hat{\theta}_j$) for each area (group) of a hierarchical dataset (Chambers & Tzavidis 2006)
- Estimate the area specific target parameter by fitting an M-quantile model for each area at $\hat{\theta}_j$

$$y_{ij} = \mathbf{x}_{ij}^T \hat{\boldsymbol{\beta}}(\hat{\theta}_j) + \epsilon_{ij}(\hat{\theta}_j)$$

- A mixed effects model uses random effects γ_j to capture the dissimilarity between groups. M-quantile models attempt to capture this dissimilarity via the group-specific M-quantile coefficients $\hat{\theta}_j$

Small Area Estimators under the M-quantile Model

- Under an M-quantile model the following small area estimator of the mean has been proposed (Chambers & Tzavidis 2006),

$$\hat{m}_j = N_j^{-1}\Big[\sum_{i \in s_j} y_{ij} + \sum_{i \in r_j} \mathbf{x}_{ij}^T \hat{\beta}(\hat{\theta}_j)\Big]$$

- We refer to this as the 'naïve' estimator. It has been noticed that this estimator may be biased particularly when outliers are present
- A bias-corrected small area estimator is derived under the Chambers-Dunstan distribution function (Tzavidis & Chambers 2007)

Small Area Estimators under the M-quantile Model (Cont'd)

- Following Tzavidis & Chambers (2007), the bias-adjusted estimator is defined as

$$\hat{m}_j^{MQ/CD} = \int_{-\infty}^{\infty} t d\hat{F}_{CD,j} =$$

$$= N_j^{-1}\{\sum_{i\in s_j} y_{ij} + \sum_{i\in r_j} \mathbf{x}_{ij}^T\hat{\beta}(\hat{\theta}_j) + \frac{N_j - n_j}{n_j}\sum_{i\in s_j}[y_{ij} - \mathbf{x}_{ij}^T\hat{\beta}(\hat{\theta}_j)]\}$$

- An alternative to the CD estimator of the distribution function that can be used is the Rao-Kovar-Mantel (RKM) estimator
- It can be shown that under srs integration of the RKM or the CD estimators will result in the same estimator for the small area mean

Small Area Estimation by Borrowing Strength over Space

- In applications involving economic, environmental and epidemiological data observations that are spatially close may be more alike than observations that are further apart
- This creates a type of spatial dependency or spatial association in the data that invalidates the assumption of independent and identically distributed (iid) observations used by conventional regression models
- One approach to accounting for spatial correlation in the data is offered by specifying models with spatially correlated errors (Anselin 1992; Cressie 1993)
- Small area literature suggests that prediction of small area parameters may be improved by borrowing strength over space (Saei & Chambers 2003; Singh *et al.* 2005; Petrucci & Salvati 2006; Pratesi & Salvati 2007)

Global Vs. Local Models for Modeling Spatial Dependency

- Regression models with spatially correlated errors are global models i.e. they assume that the relationship we are modelling holds everywhere in the study area
- Another approach to modelling a spatially non-stationary process is offered via Geographically Weighted Regression (GWR) (Brunsdon *et al.* 1996; Fotheringham *et al.* 1997)
- GWR models attempt to capture the spatial association in the data by allowing local, rather than global parameters, to be estimated

GWR Models

- Assume that we have n observations on $(y_i, \mathbf{x}_i)$ at a set of Locations (u_i)
- A GWR model is defined as follows

$$y_i = \mathbf{x}_i^T \beta(u_i) + \epsilon(u_i)$$

- GWR models allow for local rather than global parameters to be estimated and will produce estimated local surfaces of the relationship between y and x
- GWR models work by assuming that observed data near to location i will have a greater influence on the estimation of $\beta(u_i)$ than observations farther from i
- Weighted Least Squares (WLS) is used for estimating the GWR parameters

M-quantile Geographically Weighted Models

- We first propose a robust GWR model namely an M-quantile GWR model. This is a locally robust to outliers model
- With this model we attempt to model locally the different quantiles of the conditional distribution accounting at the same time for the spatial non-stationarity in the data
- For estimating the parameters of the M-quantile GWR model we use an Iterative Weighted Least Squares algorithm

Estimation for M-quantile Geographically Weighted Models

- An M-quantile GWR model is defined as follows

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta}(u_i; q) + \epsilon(u_i; q)$$

- The model parameters $\boldsymbol{\beta}(u_i; q)$ are estimated by solving

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

- Estimates of $\boldsymbol{\beta}(u_i; q)$'s are obtained via IWLS:

$$\hat{\boldsymbol{\beta}}(u_i, v_i; q) = (\mathbf{x}^T \mathbf{W}^* \mathbf{x})^{-1} \mathbf{x}^T \mathbf{W}^* \mathbf{y}$$

- $\mathbf{W}^*$ is an n by n diagonal matrix combining the spatial weights with the weights from the influence function and the modeled quantile

M-quantile GWR Models for Small Area Estimation

- Achieved via an extension to the algorithm used for estimating group effects with M-quantile models
- Step 1: Estimate an M-quantile coefficient for each unit in the sample, $\hat{\theta}_{ij}$, using M-quantile GWR models. The $\hat{\theta}_{ij}$'s are now estimated accounting for the spatial structure in the data
- Step 2: Recognize the hierarchical structure of the data and estimate a group specific M-quantile coefficient, $\hat{\theta}_j$, using the unit level M-quantile coefficients, $\hat{\theta}_{ij}$
- Step 3: Estimate the area specific target parameter by fitting an M-quantile GWR model for each area at $\hat{\theta}_j$

$$y_{ij} = \mathbf{x}_{ij}^T \hat{\beta}(u_i; \hat{\theta}_j) + \epsilon_{ij}(u_i; \hat{\theta}_j)$$

M-quantile GWR Small Area Estimators

- Under an M-quantile GWR model a 'naïve' small area estimator of the mean is

$$\hat{m}_j^{MQGWR} = N_j^{-1}\{\sum_{i\in s_j} y_{ij} + \sum_{i\in r_j} \mathbf{x}_{ij}^T\hat{\beta}(u_i;\hat{\theta}_j)\}$$

- A bias-corrected small area estimator derived under the CD or the RKM estimator $\hat{m}_j^{MQGWR/CD}$ of the distribution function is

$$N_j^{-1}\{\sum_{i\in s_j} y_{ij} + \sum_{i\in r_j} \mathbf{x}_{ij}^T\hat{\beta}(u_i;\hat{\theta}_j) + \frac{N_j - n_j}{n_j}\sum_{i\in s_j}[y_{ij} - \mathbf{x}_{ij}^T\hat{\beta}(u_i;\hat{\theta}_j)]\}$$

MSE Estimation

- MSE estimation of the small area mean is based on the ideas described in Chambers, Chandra and Tzavidis (2007)
- To start with we note that the MQGWR CD estimator can be expressed as a weighted sum of the sample y-values

$$\hat{m}_j^{MQGWR/CD} = N_j^{-1}\mathbf{w}_{sj}^T\mathbf{y}_s$$

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

- Given the linear representation, an approximation to the MSE can be computed by applying the ideas of robust mean squared error estimation for linear predictors of population quantities (Royall and Cumberland, 1978)

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

Estimation for Out of Sample Areas

- There are situations where we are interested in estimating small area characteristics for domains with no sample observations
- The conventional approach to estimating a small area characteristic in this case is synthetic estimation:

$$\hat{m}_j^{MX/SYNTH} = N_j^{-1} \sum_{i \in U_j} \mathbf{x}_i \hat{\beta}$$

$$\hat{m}_j^{MQ/SYNTH} = N_j^{-1} \sum_{i \in U_j} \mathbf{x}_i \hat{\beta}(0.5)$$

- One way of potentially improving the efficiency of synthetic estimation is by using the MQ GWR model. A synthetic-type mean predictor for out of sample area j is then

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

Empirical Investigations - Design-based Simulation (1)

- Between 1991 and 1996 researchers from the US Environmental Protection Agency (EPA) conducted an environmental health study for the lakes in the North-eastern states of the US
- Dependent variable is Acid Neutralizing Capacity (ANC), an indicator of the acidification risk of water bodies. 334 lakes were selected from the population of all northeastern lakes (21,026). The total number of measurements is 551
- Region is divided in 113 (86 in sample and 27 out of sample) 8-digit HUCs; we need to estimate mean lake Acid Neutralizing Capacity (ANC) for all HUCs
- For sampled locations we know the exact spatial coordinates of the corresponding location. For non-sampled locations the centroid of the lake is known

Empirical Investigations - Design-based Simulation (1) (Cont'd)

- We first generated a population dataset that had similar spatial structure to that of the EMAP sample data
- A total of 200 independent random samples were then taken from each HUC, with sample sizes set to equal or greater than 5. No samples were taken from HUCs that had not been sampled by EMAP
- We compare the following small area estimators (a) EBLUP, (b) M-quantile CD (MQ), (c) M-quantile GWR (MQGWR) and (d) M-quantile GWR local intercepts model (MQGWR-LI)
- For the M-quantile GWR estimators we investigate the impact of using different types of geo-referenced information i.e. for non-sampled locations we use (a) the centroid of the lake and (b) the centroid of the HUC, which represents aggregated spatial information

Empirical Investigations - Design-based Simulation (1) (Cont'd)

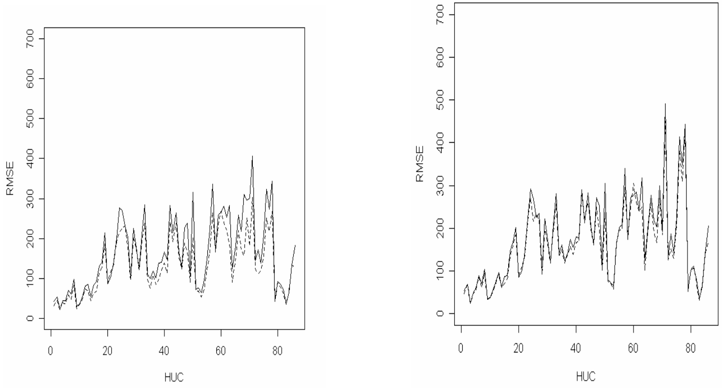
Table: Design-based simulation results using the EMAP data. Results show medians of Relative Bias (RB) and Relative Root Mean Squared Error (RMSE) over areas and simulations.

Predictor	**RB**(%)	**RRMSE**(%)	**RB**(%)	**RRMSE**(%)
	86 sampled HUCs		27 non-sampled HUCs	
EBLUP	8.51	43.41	−36.59	53.76
MQ	−1.15	40.29	−69.29	68.65
MQGWR	−0.25	26.12	−3.69	17.50
MQGWR LI	−0.69	28.52	−23.21	26.82
MQGWR (ag.sp.)	−0.08	30.34	−4.92	20.75
MQGWR LI (ag.sp.)	−0.55	35.10	−22.18	26.55

Empirical Investigations - Design-based Simulation (Cont'd)

Figure: HUC-specific values of actual design-based RMSE (solid line) and average estimated RMSE (dashed line).

Left is MQGWR version and right is the MQGWR-LI version with RMSE estimated using the proposed expression.



Empirical Investigations - Design-based Simulation (2)

- Source: the Living Standard Measurement Study (LSMS, 2002)
- Dependent variable is the household per-capita consumption. The total number of observations is 3600. The target is to estimate the District average of household per-capita consumption
- Albania is divided in 12 Prefectures, 36 Districts
- For sampled households we know the exact spatial coordinates of their corresponding locations. For non-sampled households only the centroid of the district is known

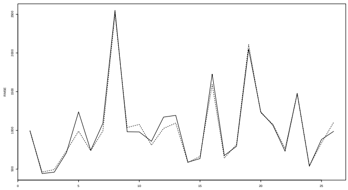
Empirical Investigations - Design-based Simulation (2) (Cont'd)

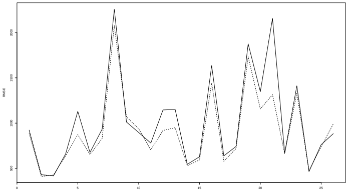
Table: Design-based simulation results using the LSMS data. Results show medians of Relative Bias (RB) and Relative Root Mean Squared Error (RMSE) over areas and simulations.

Predictor	**RB**(%)	**RRMSE**(%)	**RB**(%)	**RRMSE**(%)
	26 sampled Districts		10 non-sampled Districts	
EBLUP	1.42	10.37	6.83	16.65
MQ	0.20	10.94	1.09	14.92
MQGWR (ag.sp.)	1.05	11.80	1.54	15.79
MQGWR LI (ag.sp.)	0.30	11.44	1.30	15.01

Empirical Investigations - Design-based Simulation (2) (Cont'd)

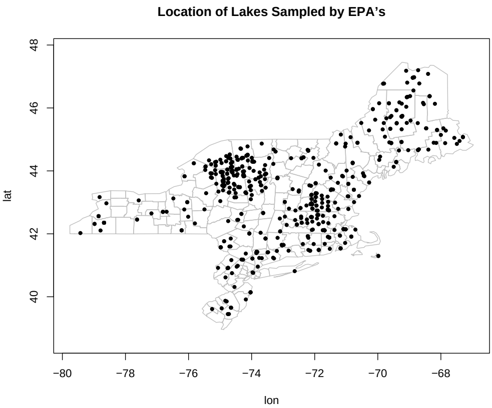
Figure: District-specific values of actual design-based RMSE (solid line) and average estimated RMSE (dashed line). Left is MQGWR version and right is the MQGWR-LI version with RMSE estimated using the proposed expression.





Case Study: Modelling Ecological Data in the North-eastern US

The figure displays the region of interest and the locations of the sampled lakes

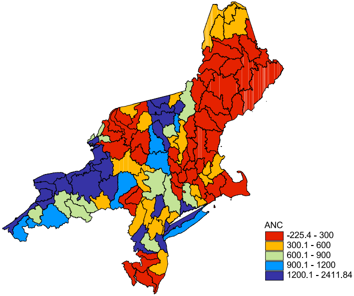


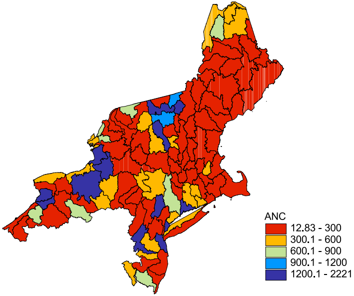
A Case Study (Cont'd)

- Direct estimates are not reliable (small or zero sample sizes) $\longrightarrow$ small area estimation techniques should be employed
- Auxiliary information is available at frame level from remote sensing: LATITUDE, LONGITUDE, ELEVATION
- Potential problem as under conventional models errors are assumed to be normally distributed
- Brunsdon, Fotheringham and Charlton (1999) applied an ANOVA test to the data and rejected the null hypothesis of stationarity of model parameters based on measuring their variability over space

Case Study (Cont'd)

Figure: Maps of estimated average ANC for all 113 HUCs. The left map shows estimates computed using MQGWR and the right map shows estimates computed using the M-quantile model.



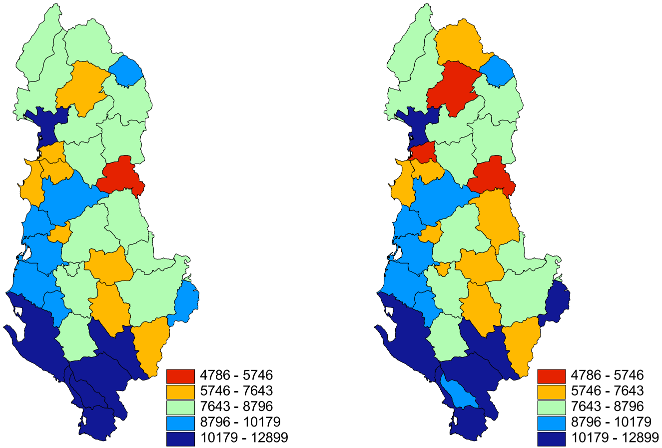


A Case Study: Modelling LSMS data in Albania

- Target parameter is the average household per-capita consumption at District level
- Auxiliary information is available at District level from Census (2001): household size, the presence of facilities in the dwelling (TV, parabolic dish antenna, refrigerator, air conditioning, personal computer), ownership of dwelling, ownership of land and ownership of car
- Potential problem as under conventional models errors are assumed to be normally distributed

Case Study (Cont'd)

Figure: Maps of estimated average household per-capita consumption for all 36 Districts. The left map shows estimates computed using MQGWR and the right map shows estimates computed using the M-quantile model.



SAMPLE project

Small Area Methods for Poverty and Living condition Estimates
EU-FP7- SSH-2007-1- Grant Agreement 217565

- Total money in the grant: $874,000$ Euros
- Starting date: 1^{st} March 2008
- Partners: University of Pisa (Coordinator), University of Siena, University of Manchester, Universidad Carlos III de Madrid , Universidad Miguel Hernandez de Helce, Warsaw School of Economics, Province of Pisa, Simurg Ricerche, Glowny Urzad Statystyczny
- Web-site: www.sample-project.it

SAMPLE project: the goal

The aim of the SAMPLE project is

- to identify and develop new indicators and models that will help the understanding of inequality and poverty with special attention to social exclusion and deprivation
- to develop models and implement procedures for estimating these indicators and their corresponding accuracy measures at the level of small area (NUTS3 and LAU 1 and 2 level).

SAMPLE project: structure of the project

The project is structured in six parts corresponding to six main areas of research or development. Each part consists of a group of tasks (called Work Package - WP) and will be carried out by a set of participant entities.

- WP 1 New indicators and models for inequality and poverty with attention to social exclusion, vulnerability and deprivation (CRIDIRE / WSE / GUS / PP / UNIPI-DSMAE / SR)
- WP 2 Small area estimation of poverty and inequality indicators (UNIPI-DSMAE / CCSR / UC3M / UMH)
- WP 3 Integration of EU-SILC data with administrative data (PP / SR / UNIPI-DSMAE)
- WP 4 Standardisation and application development - Software for living conditions estimates (SR)
- WP 5 Management (UNIPI-DSMAE / ALL)
- WP 6 Information, dissemination of results (SR / ALL)

Essential bibliography

- Chambers, R. and Dunstan, R. (1986). Estimating distribution function from survey data, *Biometrika*. **73**, 597–604.
- Breckling, J. and Chambers, R. (1988). M -quantiles. *Biometrika*, **75**, 761–771.
- Chambers, R. and Tzavidis, N. (2006). M-quantile Models for Small Area Estimation. *Biometrika*, **93**, 255–268.
- Brunsdon, C., Fotheringham, A.S. and Charlton, M. (1996) Geographically weighted regression: a method for exploring spatial nonstationarity. *Geographical Analysis*, **28**, 281–298.
- Fotheringham, A.S., Brunsdon, C. and Charlton, M. (2002) *Geographically Weighted Regression*. John Wiley & Sons, West Sussex.
- 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*].
- Salvati, N., Tzavidis, N., Pratesi, M. and Chambers, R. (2007). Small Area Estimation Via M-quantile Geographically Weighted Regression. [*Paper submitted for publication. A copy is available upon request]*.
- Tzavidis, N. and Chambers, R. (2007). Robust prediction of small area means and distributions. [*Paper submitted for publication. A copy is available upon request*].
- Tzavidis, N., Salvati, N., Pratesi, M. and Chambers, R. (2007). M-quantile models for poverty mapping. To appear in Stat Meth. & Applications.