Two stage non-parametric approach for small area estimation

Pushpal Mukhopadhyay and Tapabrata Maiti
Iowa State University

Abstract

Small area estimators commonly borrow strength from other related areas. These indirect estimators use models (explicit or implicit) that relate the small areas through supplementary data. Various unit-level and area-level small area models are proposed in the literature, but all these models assume the small area mean is linearly related with supplementary information. In this article, we propose an area-level, non-parametric regression estimator based on Nadaraya-Watson kernel on small area mean. In this direction, we adopt a two-stage estimation approach proposed by Prasad and Rao (1990). The asymptotic properties of the proposed estimator are studied and a second order approximation to the mean squared prediction error (MSPE) of the two-stage estimator and the estimator of MSPE approximation are obtained under normality. We perform a simulation study to show the superiority of the proposed estimator and finally we apply this smoothing method to estimate soil loss due to erosion in certain mid-western counties in U.S.

1 Introduction

The term "Small Area" denotes any subpopulation for which direct estimates of adequate precision cannot be produced (Rao, 2003). Some indirect domain estimation procedures are used to gain precision. This problem is not new in survey statistics. Small area statistics existed even in eleventh century England and in seventeenth century Canada (Brackstone, 1987). The use of small area estimation is relatively common in survey sampling; e.g., formulating policies and programs in allocating the government funds and in regional planning . Government planing is just one place small area estimation is used. Recent years have seen an increased demand of small area estimates from the private sector. Small businesses relay heavily on local socio-economic conditions, local environmental conditions and other local conditions. A small area estimate from a large national survey helps them to save a large amount of money. There are several organizations who produce small area statistics. The U.S. National Center for Health Statistics (NCHS)which pioneered the use of synthetic estimation based on implicit models to develop state estimates of disability and other health characteristics for different groups from the National Health Interview Survey (NHIS). The U.S. National Agriculture Service (NASS) publishes model-based county estimates of crop acreage using remote sensing data as auxiliary information. The U.S. Census Bureau produces estimates of small area incomes based on a basic area level linking model. The National Research Council produces model-based county estimates of poor school-age children in the USA (National Research Council, 2000).

All of these small area models use parametric estimation procedures to relate between covariates and unobserved small area means. In this paper, we propose a non-parametric smoothing approach to predict the unobserved small area mean. We will also present MSE of the above predictor and an estimate of MSE. We have conducted a simulation study and have shown that if the linearity breaks then the non-parametric model gives much better results than the

linear prediction. Even when the linear relationship is true the non-parametric prediction is 'as good as' the linear prediction.

Section 2 describes the background and the proposed kernel based non-parametric approach. A simulation study is described to check the performance of the proposed estimator in section 3. A practical application of the proposed method is discussed in section 4 using the National Resource Inventory (NRI) data set. Section 5 discuss some limitations of the proposed method and its remedies. The proof of the theorems are not given here but can be obtained from the authors.

2 Kernel-Based approach

Small area means are usually modeled using a mixed linear model of the form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon} \tag{1}$$

where X is a design matrix, $\mathbf{u}$ is a random vector commonly known as small area effects , and $\boldsymbol{\epsilon}$ is a vector of sampling error (Rao, 2003). In particular, for basic area level model with one covariate, model (1) can be written as

$$y_i = \theta_i + \epsilon_i \tag{2}$$

$$\theta_i = \beta_0 + \beta_1 x_i + u_i \tag{3}$$

where ϵ_i and u_i are distributed independently as $N(0, D_i)$ and $N(0, \sigma_u^2)$ (Fay and Herriot, 1979). Prasad and Rao proposed an estimate of the MSE for the best linear unbiased predictor under model 3 (Prasad and Rao, 1990).

In almost all applications of small area estimation this linear mixed effect model is assumed and the estimates are very sensitive to this assumption. If the assumption of linearity between the small area mean and supplementary information fails then borrowing strength from other areas using a linear model is not very appropriate. In fact, depending on the dependence of the area level mean to the covariate, the relative bias can be as large as 150% (See simulation results). To reduce the relative bias and to get a better estimate of the MSE we propose a model of the form

$$y_i = \theta_i + \epsilon_i \tag{4}$$

$$\theta_i = m(x_i) + u_i \tag{5}$$

where $i = 1, 2, ..., m$ denotes the number of small areas. The function $m(.)$ is a smooth mean function which defines the true relation between x and y. θ_i is the unobserved small area mean, y_i is the observed direct survey estimator of small area mean, u_i is independent and identically distributed random error with $E(u_i) = 0$ and $V(\epsilon_i) = \sigma_u^2$, and the ϵ_i is independent sampling error with $E(\epsilon_i) = 0$ and $V(\epsilon_i) = D_i$. We also assume that all the D_i's are known constants.

To estimate $m(x_i)$ we propose the use of Nadaraya-Watson kernel estimate

$$\hat{m}_h(x) = \frac{\sum_i K_h(x - x_i) y_i}{\sum_i K_h(x - x_i)} \tag{6}$$

where $K_h(.)$ is a kernel function with band width h and is of the form $K_h(u) = \frac{1}{h} K(u/h)$ with $K(.)$ satisfying:
i) $K(.)$ is symmetric.
ii) $K(.)$ is bounded and continuous on the range of x (say, χ)
iii) $\int_\chi K(a)da = 1$
The above estimator is linear in y_i and can be rewritten as, $\hat{m}_h(x) = \frac{1}{m} \sum_{i=1}^{m} W_{hi}(x) y_i$, where $W_{hi}(x) = \frac{K_h(x - x_i)}{1/n \sum_i K_h(x - x_i)}$.
With the above setup it is easy to show that the best predictor of small area mean θ_i can be written as,

$$E(\theta_i | y_i) = \tilde{\theta}_i = \gamma_i y_i + (1 - \gamma_i) \hat{m}_h(x_i) \tag{7}$$

where $\gamma_i = \frac{\sigma_u^2}{\sigma_u^2 + D_i}$ and we assume σ_u^2 is known. Now in the second stage we estimate,

$$\hat{\theta}_i = \hat{\gamma}_i y_i + (1 - \hat{\gamma}_i) \hat{m}_h(x_i) \tag{8}$$

where $\hat{\gamma}_i = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + D_i}$ and $\hat{\sigma}_u^2$ is a consistent estimator of σ_u^2.
From the theory of kernel regression it is now easy to show that $\hat{m}(x)$ is a consistent estimator for $m(x)$ at every point of continuity $m(.)$ under the non-parametric model (5). More formally, we can prove theorem 1.

Theorem 1. Under the non-parametric model (4 and 5) with a one-dimensional predictor variable x and,
(A1) $\int |K(a)|da < \infty$,
(A2) $lim_{|a|\to\infty} aK(a) = 0$,
(A3) $Ey_i^2 < \infty$ for all i,
(A4) $m \to \infty$, $mh \to \infty$,
then, at every point of continuity of $m(x)$,
$m^{-1}\sum_{i=1}^{m} \frac{K_h(x-x_i)y_i}{\sum_i K_h(x-x_i)} \to m(x)$ in probability.

Under certain bound conditions on x_i and $k(.)$ we can find the mean square error for estimating $m(.)$ by $\hat{m}(.)$. Theorem 2 states the complete result to find the estimate of $\hat{m}(.)$.

Theorem 2. Under the non-parametric model (4 and 5) with a one-dimensional predictor x and define $c_k = \int K^2(a)da$, $d_k = \int a^2 k(a)da$ and assume
(A5) $m(.)$ is continuous.
(A6) $max_i|x_i - x_{i-1}| = \bigcirc(m^{-1})$
(A7) $D_i = D$ for all i and is finite
(A8) $m \to \infty$, $mh \to \infty$,
then
$E[\hat{m}_h(x_i) - m_h(x_i)]^2 \approx (mh)^{-1}\sigma^2 c_k + h^4 d_k^2 [m^{''}(x_i)]^2/4$
where $\sigma^2 = \sigma_u^2 + D$.

In theorem 2, we ignore the terms which are higher than the order of m^{-1}. From the expression we see that the MSE of $\hat{m}(x)$ has two parts. One part comes from the bias and the other part is related to the variance. A suitable selection of bandwidth h can compromise between bias and variance. Selection of bandwidth has a very large effect on kernel smoothing. In this paper we will choose a fixed bandwidth of $h \propto n^{-1/5}$. For a more detailed discussion on bandwidth selection see Hardle, 1990. So far, for any given x, we have got the form for $\hat{m}(x)$ and we also know its mean square error. Now to find out $\hat{\theta}_i$ we need to estimate σ_u^2. A method of moment type estimator using weighted sum of squares for residuals is given in proposition 1.

Proposition 1. Under the assumptions of theorem 2 and if x is a point of continuity of $\sigma_u^2(x)$ then, $\hat{\sigma}_u^2(x) = min\{0, \frac{1}{m-1}\sum_{i=1}^{m} W_{hi}(x)\{y_i - \hat{m}(x_i)\}^2 - D\}$. That is, between area variance for any given small area x_i can be estimated as a weighted sum of the residuals. The proposed estimator can be negative but it can be shown that as $m \to \inf$, $P(\hat{\sigma}_u^2 < 0) \to 0$. Hence small area means can now be estimated using model (8).

2.1 MSE of the proposed estimator

In this subsection, we will calculate the mean square error of $\hat{\theta}_i$ and will propose an estimator of the true MSE. To get the MSE of $\hat{\theta}_i$ we break the square difference of $\hat{\theta}_i$ and θ_i into three parts. First we must find the MSE of $\theta^* = \gamma_i y_i + (1-\gamma_i)m(x_i)$ from θ_i. Then we find the average square distance between θ_i^* and $\tilde{\theta}_i$, which is mainly due to the estimation of $m(.)$ by $\hat{m}(.)$. The third term is the average square distance between $\tilde{\theta}_i$ and $\hat{\theta}_i$ and it is due to the estimation of σ_u^2 by $\hat{\sigma}_u^2$. The result is stated in theorem 3.

Theorem 3. Under the assumptions (A1) to (A8) and if (A9) ϵ_i and u_i are independently normally distributed then
$MSE(\hat{\theta}_i) \approx \frac{D\sigma_u^2}{\sigma_u^2+D} + (1-\gamma)^2 MSE[\tilde{m}_h(x_i)] + D^2(\sigma_u^2 + D)^{-4} E[(y_i - m(x_i))(\hat{\sigma}_u^2 - \sigma_u^2)]^2$
where $MSE(m_h(x_i))$ is given by theorem 2.

An estimation of the above MSE can be obtained by plugging in $\hat{\sigma}_u^2$ for σ_u^2, $mse(.)$ for $MSE(.)$, and by taking a first step approximation of the product term. But if we do this, the first term of the expression is estimated with a bias of order $\frac{1}{n}$ and hence a bias adjusted estimator is proposed in proposition 2.

Proposition 2. Under the assumption of (A1) to (A9), mean square of error of $\hat{\theta}_i$ can be estimated as, $mse(\hat{\theta}_i) = \frac{D\hat{\sigma}_u^2}{\hat{\sigma}_u^2+D} + (1-\hat{\gamma})^2 mse[\tilde{m}_h(x_i)] + 2D^2(\hat{\sigma}_u^2 + D)^{-3} mse(\hat{\sigma}_u^2)$.

Therefore unobserved small area means can be estimated using a non-parametric setup and an estimate of MSE and its estimate can be obtained. Moreover, if we put $x_i^T\hat{\beta}$ for $\hat{m}(x_i)$ we will get exactly the same form of linear mixed effect estimates as proposed in Rao, 2003.

3 Simulation

To check the performance of the kernel based estimate of small area prediction over the linear para-

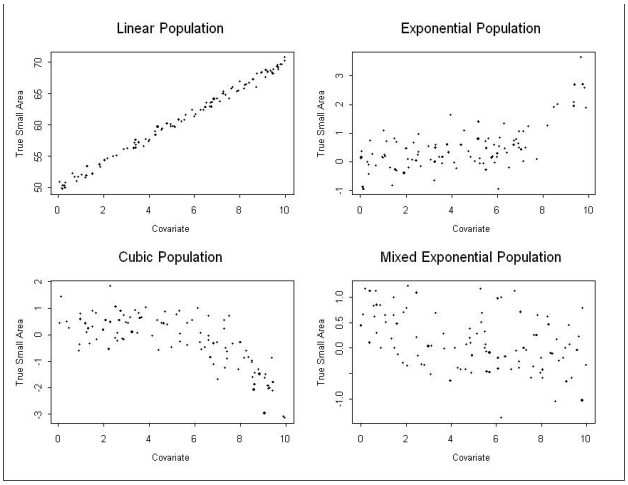

Figure 1: *Plot for generated populations for simulation.*

metric model prediction, we perform some simulation studies. We considered a wide range of smoothing functions as true functions and different ratios of small area variance over error variance.
We consider four mean functions:
i)Linear: $m_1(x) = 50 + 2x$
ii) Cubic: $m_2(x) = .01 + .2x - .005x^3$
iii) Exponential: $m_3(x) = exp(.5x)$
iv) Mixed Exponential: $m_4(x) = \{1 - x + exp((x - 5)^2)\}10^{-6}$
where x_i is generated from uniform (0,10) distribution for $i = 1, 2, ..., 100$ areas. The small area effects σ_u^2 is taken as .25 and D_i are taken to be .1 for one-third areas, .25 for one-third areas and .5 for the rest of the one-third areas.

For all the above populations, we fit both the linear model (3) and the non-parametric model (5). Small area means and an estimate of MSE for each small area is computed using both models. To compare the two models we generate populations R times and we calculate the following fit statistics:
i) Relative bias for i^{th} area :

$$RB(\hat{\theta}_i) = \frac{1}{R}\sum_{r=1}^{R} \hat{\theta}_i^{(r)} - \theta_i^{(r)} \tag{9}$$

ii) True MSE for the estimated mean of i^{th} area:

$$MSE(\hat{\theta}_i) = \frac{1}{R}\sum_{r=1}^{R} \{\hat{\theta}_i - \theta_i\}^2 \tag{10}$$

iii) Relative bias of estimated MSE for i^{th} area:

$$RB\{mse(\hat{\theta}_i)\} = \frac{\frac{1}{R}\sum_{r=1}^{R}\{mse(\hat{\theta}_i)^{(r)} - MSE(\hat{\theta}_i)\}}{MSE(\hat{\theta}_i)} \tag{11}$$

. Where $mse(\hat{\theta}_i)$, and $MSE(\hat{\theta}_i)$ are the estimated and true mean square error for the i^{th} area.
iv) Coefficient of variation of the estimated MSE for the i^{th} area:

$$CV\{mse(\hat{\theta}_i)\} = \frac{\sqrt{\frac{1}{R}\sum_{r=1}^{R}\{mse(\hat{\theta}_i)^{(r)} - MSE(\hat{\theta}_i)\}^2}}{MSE(\hat{\theta}_i)} \tag{12}$$

.
A better model should have smaller values for all these three statistics defined above.

3.1 Results

We gave generated values for 100 small areas with a summary of the statistics given in table (1) to table (4). In these tables, predictions from Fay-Herriot model is denoted as FH and predictions from non-parametric mixed effect model is denoted as NPME. Mean, standard error, first quantile (1^{st} Q.), and third quantile (3^{rd} Q.) are given in the columns. From table (1), for the linear population, predictions from NPME model is 'as good as' the predictions from the FH model. For any other population considered in the simulation, NPME predictions give low relative bias as compared to FH predictions. Although we have less bias using NPME model, we cannot reduce the MSE by a big margin (table (2) to table (4)). One can always expect to reduce the MSE (or to obtain a balance between RB and MSE) from the NPME predictions by changing the bandwidth. The estimation of the MSE using a non-parametric model always has less RB(mse) and less CV(mse) as compared to the estimated MSE using FH predictions. Therefore, for all the populations we have

		Mean	SE	1^{st} Q.	3^{rd} Q.
RB	FH	.0003	.002	-.001	.001
	NPME	.0007	.002	-.001	.002
MSE	FH	0.13	0.08	0.08	0.16
	NPME	0.18	0.12	0.10	0.23
RB(mse)	FH	0.37	0.31	0.19	0.42
	NPME	0.26	0.17	0.14	0.37
CV(mse)	FH	7.19	6.16	3.77	8.21
	NPME	8.14	9.56	3.85	8.76

Table 1: *Linear Population: NPME prediction is performing as good as FH predictions.*

		Mean	SE	1^{st} Q.	3^{rd} Q.
RB	FH	.37	12.05	-0.64	0.38
	NPME	.21	9.94	-0.48	0.14
MSE	FH	0.18	0.12	0.10	0.23
	NPME	0.14	0.09	0.07	0.18
RB(mse)	FH	6.16	5.27	3.23	7.02
	NPME	-3.91	4.59	-4.21	-1.85
CV(mse)	FH	10.64	9.11	5.58	12.14
	NPME	6.83	8.02	3.23	7.35

Table 2: *Cubic Population: NPME prediction performs better than FH predictions.*

considered here, NPME predictions are 'as good as' the linear predictions, and for the populations with a nonlinear trend NPME predictions has less bias as compared to the predictions from the linear model.

4 Application to NRI

The United States National Resource Inventory (NRI) is a large nation-wide survey of the U.S. land area. The current NRI is a longitudinal survey of soil, water, and related environmental resources. The NRI is designed to assess conditions and trends of non-federal US lands on a yearly basis. The data were collected using a two-stage, two-phase supplemented panel longitudinal area sample design at the national level (Nusser and Goebel, 1997 and Fuller, 2003). In some Midwestern states, soil erosion due

		Mean	SE	1^{st} Q.	3^{rd} Q.
RB	FH	0.47	5.95	-0.55	0.48
	NPME	0.24	6.23	-0.51	0.27
MSE	FH	0.16	0.10	0.09	0.20
	NPME	0.14	0.09	0.08	0.18
RB(mse)	FH	0.98	0.84	0.52	1.12
	NPME	-1.07	1.26	-1.15	-0.51
CV(mse)	FH	5.03	4.31	2.64	5.74
	NPME	4.07	4.77	1.92	4.38

Table 3: *Exponential Population: NPME performs better on relative bias but similar with FH on MSE, and estimates of MSE.*

		Mean	SE	1^{st} Q.	3^{rd} Q.
RB	FH	0.42	9.20	-0.62	0.44
	NPME	0.22	7.44	-0.50	0.25
MSE	FH	0.13	0.09	0.08	0.17
	NPME	0.15	0.10	0.08	0.20
RB(mse)	FH	3.52	3.84	2.89	4.47
	NPME	-2.48	3.31	-2.11	-0.99
CV(mse)	FH	7.05	6.98	6.01	8.82
	NPME	5.55	6.37	2.85	6.93

Table 4: *Mixed-Exponential Population: NPME predictions are better over FH predictions except for MSE, where both the predictions are nearly same.*

Total Observations	:	75573
Number of states	:	3
Number of counties	:	276
County with size 0	:	57
County with size < 10	:	114
County with size < 20	:	152

Table 5: *Summary of observed counties*

to wind causes massive problems. It may be beneficial for the local and state governments to estimate this soil loss in their area. Because of the national level design of the NRI, sample size within one county could be as low as 5, but the nature of the problem is similar to all the adjacent counties. We should somehow borrow strength from other counties with similar trends (soil properties, landscape, weather, etc.) to increase precision of our estimation. In this application we will use soil erodibility index (IFact) as our auxiliary information and soil loss due to wind for the year (WEQ03) as our covariate. The response variable WEQ03 is not directly observed in the field but rather it is a function of some other observed variable. There are several advantages of choosing IFact as a covariate. First, we believe that high values of IFact indicates higher erosion (Wind Erosion Research Unit). Second, IFact can be obtained from Natural Resources Conservation Service (NRCS) soil survey database available through the NRCS Soil Data Mart (SDM) at each county level for US. Table (5) shows a summary of observations in each county. There are 152 counties with less than 20 observations. Figure (2) shows the relation between WEQ03 and IFact. Each point in figure (2) represents the observed mean for each county. The county level mean plot for soil loss due to wind suggests a non-linear relationship among WEQ03 and IFact. This motivates us to use a non-parametric, small area model as opposed to a linear model. The model $y_i = m(x_i) + u_i + \epsilon_i$ is fitted with $m(x_i)$ as the Nadaraya-Watson kernel estimate to the NRI dataset and the estimates are discussed in the next subsection.

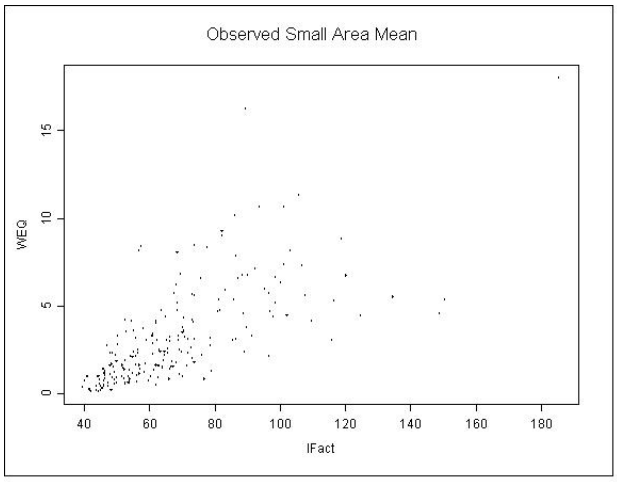

Figure 2: *Scatter plot of WEQ against erodibility index. Each point represents observed mean value for each county. Scatter plot shows a nonlinear pattern or a trend of unequal variance.*

4.1 Estimates

The 2003 NRI data set has not yet been released for public use, so all values are strictly for demonstrational purposes and are in no way related with the original observed values. Table (6) shows a summary of estimated means for all the counties with different sample sizes. The number of counties for a specified size are given within bracket for each size category. From this table we can compare the direct survey weighted estimates (DE), estimates using Fay-Herriot model (FH), and estimates using nonparametric mode l (NP). The predicted means from the NP model is always close to the observed mean as compared to the predicted mean from the FH model. The estimated MSE under NP model is always lower than any of the other two methods when the sample size within a county is less than 20. When the sample size within a county is more than 50, DE gives better precession than NP or FH predictions. For small sample sizes within a county, FH model improves precision over DE but is not better than NP model. This is not very surprising as the data plot suggests a violation from linearity. A plot of esti-

	County Size = 1 (8)	
	Mean	mse
DE	0.93	-
	(0.52, 1.19)	
FH	3.03	0.24
	(1.36, 3.77)	(0.23, 0.26)
NP	1.50	0.22
	(1.31, 1.89)	(0.19, 0.22)
	County Size = 2-10 (52)	
	Mean	mse
DE	3.04	3.53
	(0.44, 4.48)	(0.03, 2.60)
FH	2.94	0.98
	(0.83, 5.12)	(0.07, 1.12)
NP	2.99	0.77
	(0.49, 4.62)	(0.14, 0.99)
	County Size = 11-20 (8)	
	Mean	mse
DE	2.02	0.65
	(0.68, 2.53)	(0.02, 0.53)
FH	2.48	0.18
	(1.35, 2.81)	(0.10, 0.17)
NP	1.97	0.11
	(0.73, 2.41)	(0.09, 0.14)
	County Size > 50 (52)	
	Mean	mse
DE	4.02	0.42
	(2.08, 3.36)	(0.03, 0.45)
FH	3.16	0.74
	(2.55, 3.66)	(0.15, 0.90)
NP	3.74	0.85
	(2.07, 4.61)	(0.24, 1.01)

Table 6: *Distribution of small area means and its estimated MSE based on county size. DE denotes the direct survey estimates, FH is the predicted mean from Fay-Herriot model and NP is the predicted means from non-parametric model. Estimated MSE for each predicted mean is given in the parenthesis.*

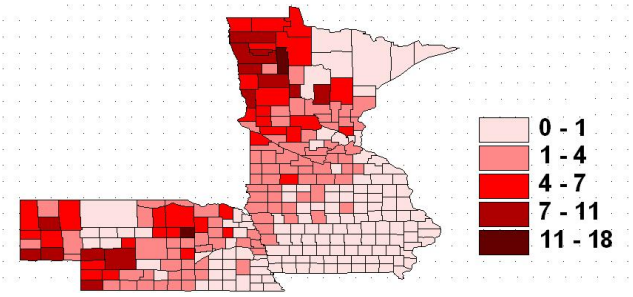

Figure 3: *Estimation of county means using direct survey means. A darker shade of red implies more erosion.*

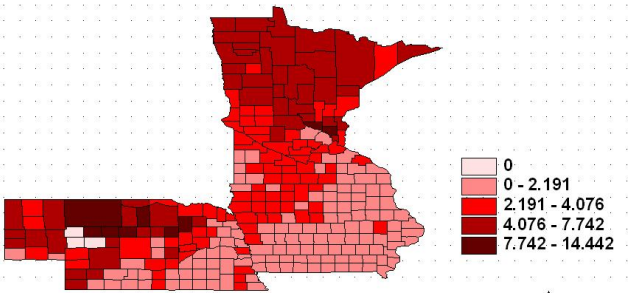

Figure 4: *Estimation of county means using Fay-Herriot model. A Dakar shade of red implies more erosion.*

mated county means (for three states under study) using direct estimates is given in (3), whereas plots of predicted means for each county using FH model and NP model are shown in figure (4), and figure (5) respectively. Dark values of red imply high values for estimated mean for that county. A closer look at these three plots suggests that both FH and NP prediction makes the plot look more smooth (the change of color is not a jump). NP prediction puts more color on the map as compared to the FH prediction, which makes it more smooth.

5 Discussion

In this work, we take a step toward the use of non-parametric regression for small area estimation. Nadaraya-Watson based kernel estimation is used to

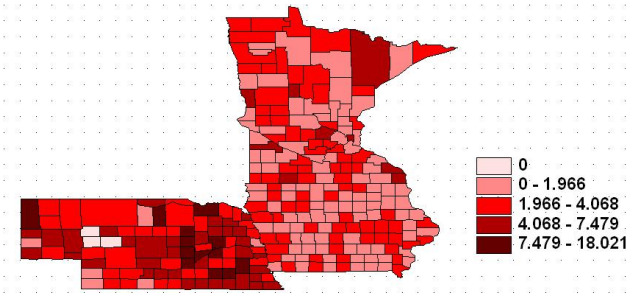

Figure 5: *Estimation of county means using non-parametric model. A darker shade of red implies more erosion.*

estimate small area means. We also find out the MSE and an estimate of MSE for the proposed estimator. A simulation study shows the efficiency of the proposed estimator over its linear counterpart. The proposed estimator is applied to a NRI data set to estimate soil loss due to wind at the county level in three mid-western states in the US. The estimated value shows more smooth estimates than both the direct or linear model based estimates.

Much more rigorous work is required to find out the exact order of approximation for the proposed estimators. All the theorems are stated under the assumption that the sampling variances are the same in each county. Work needs to be done to incorporate unequal sampling variance.

References

[1] Brackstone, G.J. (1987), "Small Area Data: Policy issues and technical challenges, in R. Platek, J.N.K. Rao, C.E. Sarndal and M.P. Singh (Eds.)," *Small Area Statistics*, New York: Wiley, pp. 3-20.

[2] Fay, R.E., and Herriot, R.A. (1979), "Estimation of income from small places: An application of James-Stein procedures to census data." *Journal of the American Statistical Association*, **74**, 269-277.

[3] Fuller, W.A. (2003), "Sample selection for the 2000 NRI-2004 NRI surveys." Unpublished manuscript.

[4] Hardle, W. (2002), "Applied non-parametric regression," *Cambridge University Press.*

[5] Nusser, S.M., and Goebel J.J. (1997), "The national resource inventory: a long-term multi-resource monitoring program." *Environmental and Ecological Statistics*, **4**, 181-204.

[6] Prasad, N.G.N., and Rao, J.N.K. (1990), "The estimation of the mean squared error of the small area estimators." *Journal of the American statistical association*, **85**, 163-171.

[7] Rao, J.N.K. (2003), "Small Area Estimation," *Wiley Series in Survey Methodology.*