SMALL AREA ESTIMATION UNDER INFORMATIVE SAMPLING AND NONRESPONSE

Michail Sverchkov, Bureau of Labor Statistics

&

Danny Pfeffermann, Government Statistician of Israel, Professor, Hebrew University of Jerusalem, Israel & Southampton University (S3RI), UK

The opinions expressed in this paper are those of the authors and do not necessarily represent the policies of the Bureau of Labor Statistics and Israel CBS

BLS
BUREAU OF LABOR STATISTICS
U.S. DEPARTMENT OF LABOR

www.bls.gov

Introduction and Notation

$\{y_{ij}, x_{ij}; i = 1...M, j = 1...N_i\}$ - finite population measurements assumed to follow the **two level population model:**

$$y_{ij} \mid x_{ij}, u_i^U \sim f(y_{ij} \mid x_{ij}, u_i^U),\ i = 1...M,\ j = 1...N_i$$

$$u_i^U \sim f(u_i^U);\ E(u_i^U) = 0,\ V(u_i^U) = \sigma_{u^U}^2 .$$

y_{ij} - target study variable

$x_{ij} = (x_{ij}^1 ... x_{ij}^K)$ - covariates known for entire population.

Target: Estimate small area means $\bar{Y}_i = N_i^{-1} \sum_{j=1}^{N_i} y_{ij}$, based on the two-stage sample.

Two-stage Sampling Scheme:

Select m areas with inclusion probabilities $\pi_i = \Pr(i \in s)$,

Sample n_i units from selected cluster i with probabilities $\pi_{j|i} = \Pr(j \in s_i \mid i \in s)$.

$I_i, I_{ij} \rightarrow$ sample indicators,

$w_i = 1/\pi_i$, $w_{j|i} = 1/\pi_{j|i} \rightarrow$ sampling weights.

Unit non-response: $R_{ij} \rightarrow$ unit response indicators,

$R = \{(i, j) : I_i = 1, I_{ij} = 1, R_{ij} = 1\}$;

$R^c = \{(i, j) : I_i = 1, I_{ij} = 1, R_{ij} = 0\}$. **(No area non-response).**

Observed data

It is assumed that the response occurs independently between units.

The observed sample of respondents can be viewed therefore as the result of a two-phase sampling process where in the first phase the sample is selected from the population with **known** inclusion probabilities, and in the second phase the sample is 'self selected' with **unknown** response probabilities (Särndal and Swensson, 1987).

Model for observed data

Under our sampling scheme and response, the observed data follow the **two level respondents' model:**

$$y_{ij} \,|\, x_{ij}, u_i \sim f_{\boldsymbol{R}}(y_{ij} \,|\, x_{ij}, u_i) = f(y_{ij} \,|\, x_{ij}, u_i, (i,j) \in R),$$

$u_i \sim f(u_i \,|\, i \in s)$; $E(u_i \,|\, i \in s) = 0$, where $u_i = u_i^U - E(u_i^U \,|\, i \in s)$.

$f_{\boldsymbol{R}}(y_{ij} \,|\, x_{ij}, u_i) \neq f(y_{ij} \,|\, x_{ij}, u_i^U)$ **(population model)**

Since the model refers to the ***observed data***, it can be estimated and tested by classical SAE methods.

Let $p(y_{ij}, x_{ij}) = \Pr[(i, j) \in R \mid y_{ij}, x_{ij}, i \in s, j \in s_i]$.

If $p(y_{ij}, x_{ij})$ were known, the sample of respondents could be considered as a two-stage sample from the finite population with known selection probabilities π_i and $\tilde{\pi}_{j|i} = \pi_{j|i} p(y_{ij}, x_{ij})$.

Also, **if known**, the response probabilities could be used for imputation within the selected areas via the relationship between the **sample** and **sample-complement distributions** (Sverchkov & Pfeffermann, 2004)**;**

$$f(y_{ij} \mid x_{ij}, u_i, (i,j) \in R^c) =$$

$$\frac{[p^{-1}(y_{ij}, x_{ij}) - 1] f(y_{ij} \mid x_{ij}, u_i, (i,j) \in R)}{E\{[p^{-1}(y_{ij}, x_{ij}) - 1] \mid x_{ij}, u_i, (i,j) \in R\}}. \qquad \textbf{(1)}$$

(1) refers to the model for the ***observed*** data and therefore can be estimated by classical SAE methods.

Estimation of response probabilities

Assume a parametric model for the **response probabilities** $p(y_{ij}, x_{ij}; \gamma) = \Pr[(i, j) \in R \mid y_{ij}, x_{ij}, i \in s, j \in s_i; \gamma]$ and suppose that p is differentiable with respect to the (**vector**) parameter γ.

If the missing data were observed, γ could be estimated by solving the equations:

$$0 = \sum_{(i,j)\in R} \frac{\partial \log p(y_{ij}, x_{ij}; \gamma)}{\partial \gamma} + \sum_{(i,j)\in R^c} \frac{\partial \log[1 - p(y_{ij}, x_{ij}; \gamma)]}{\partial \gamma}. \qquad \textbf{(2)}$$

Denote the observed data by

$\boldsymbol{O} = \{y_{ij}, \pi_{j|i}, \pi_i, n_i, (i, j) \in R;\ x_{kl}, k = 1...M,\ l = 1...N_i\}$.

Missing Information Principle: since the outcome values are missing for $(i,j) \in R^c$, we propose to solve instead,

$$0 = E\{[\sum_{(i,j)\in R} \frac{\partial \log p(y_{ij}, x_{ij}; \gamma)}{\partial \gamma} + \sum_{(i,j)\in R^c} \frac{\partial \log[1 - p(y_{ij}, x_{ij}; \gamma)]}{\partial \gamma}] \mid O\} =$$

$$\sum_{(i,j)\in R} \frac{\partial \log p(y_{ij}, x_{ij}; \gamma)}{\partial \gamma} +$$

$$\sum_{(i,j)\in R^c} E\{\frac{\partial \log[1 - p(y_{ij}, x_{ij}; \gamma)]}{\partial \gamma} \mid O, (i,j) \in R^c\} \overset{\text{by (1)}}{=}$$

$$\sum_{(i,j)\in R} \frac{\partial \log p(y_{ij}, x_{ij};\gamma)}{\partial \gamma} +$$

$$\sum_{(i,j)\in R^c} E\left(\frac{E\{[p^{-1}(y_{ij}, x_{ij};\gamma)-1]\frac{\partial \log[1-p(y_{ij}, x_{ij};\gamma)]}{\partial \gamma} \mid x_{ij}, u_i, (i,j)\in R\}}{E\{[p^{-1}(y_{ij}, x_{ij};\gamma)-1] \mid x_{ij}, u_i, (i,j)\in R\}} \middle| O \right) = 0 \quad \textbf{(3)}$$

(we assume $f(y_{ij} \mid O, u_i, (i,j)\in R) = f(y_{ij} \mid x_{ij}, u_i, (i,j)\in R)$)

The expectations in **(3)** refer to the model for the ***observed*** data and therefore can be estimated by classical SAE methods.

The parameter γ can be estimated by solving (3).

Note: if $p(y_{ij}, x_{ij}; \gamma)$ is a function of x_{ij} and γ only, (**missing data are MAR**), **(3)** reduces to the common log-likelihood equations,

$$0 = \sum_{(i,j)\in R} \frac{\partial \log p(x_{ij};\gamma)}{\partial \gamma} + \sum_{(i,j)\in R^c} \frac{\partial \log[1 - p(x_{ij};\gamma)]}{\partial \gamma}. \quad \textbf{(4)}$$

Prediction of small area means (P-S, *JASA* 2007)

$$MSE(\hat{\bar{Y}}_i) = E[(\hat{\bar{Y}}_i - \bar{Y}_i)^2 \mid O, I_i) = [\hat{\bar{Y}}_i - E(\bar{Y}_i \mid O, I_i)]^2 + V(\bar{Y}_i \mid O, I_i)$$

$\hat{\bar{Y}}_i = E(\bar{Y}_i \mid O, I_i)$ - Optimal small area predictor for area i.

Optimal small-area predictors for selected areas:

$$\hat{\bar{Y}}_i = E(\bar{Y}_i \mid O, I_i = 1) = N_i^{-1}[\sum_{j:(i,j)\in R} y_{ij} + \sum_{k=1,k\notin R}^{N_i} E(y_{ik} \mid O, I_i = 1)] \cong$$

$$N_i^{-1}(\sum_{j,(i,j)\in R} y_{ij} +$$

$$\sum_{k=1,k\notin R}^{N_i} E\{\frac{E[(\tilde{\pi}_{k|i}^{-1} - 1) y_{ik} \mid x_{ik}, u_i, (i,k)\in R]}{E[(\tilde{\pi}_{k|i}^{-1} - 1) \mid x_{ik}, u_i, (i,k)\in R]} \mid O\}) \cong$$

$$N_i^{-1}\Big(\sum_{j,(i,j)\in R} y_{ij} +$$

$$\sum_{k=1,k\notin R}^{N_i} E\{\frac{E\{[w(y_{ik},x_{ik})-1]y_{ik} \mid x_{ik},u_i,(i,k)\in R\}}{E\{[w(y_{ik},x_{ik})-1] \mid x_{ik},u_i,(i,k)\in R\}} \mid O\}\Big)\textbf{;} \quad \textbf{(5)}$$

$\hat{\tilde{\pi}}_{k|i} = \pi_{k|i} p(y_{ik},x_{ik};\hat{\gamma})$ and $w(y_{ik},x_{ik}) = E[\hat{\tilde{\pi}}_{k|i}^{-1} \mid y_{ik},x_{ik},(i,k)\in R]$.

(Refers to **observed data** and can be estimated by regression or non-parametrically).

Expectations in **(5)** are over the model for the ***observed*** data that was **estimated before**.

Optimal small-area predictors for unselected areas:

$$\hat{\bar{Y}}_i = E(\bar{Y}_i \mid O, I_i = 0) = N_i^{-1}[\sum_{k=1}^{N_i} E(y_{ik} \mid O, I_i = 0)]$$

$$\cong N_i^{-1} \sum_{k=1}^{N_i} \frac{\sum_{l \in s} [(\pi_l^{-1} - 1) K_l(x_{ik})]}{\sum_{l \in s} (\pi_l^{-1} - 1)} \qquad \textbf{(6)}$$

$$K_l(x) = E(y_{lk} \mid x_{lk} = x, (l,k) \in U) =$$

$$E\{\frac{E[w(y_{lk}, x_{lk}) y_{lk} \mid x_{lk} = x, u_l, (l,k) \in R]}{E[w(y_{lk}, x_{lk}) \mid x_{lk} = x, u_l, (l,k) \in R]} \mid O\}$$

(6) depends on $w(y_{lk}, x_{lk})$ and the model for the ***observed*** data.

Example: Logistic Mixed Model with Logistic Response

Let $y_{ij} \sim Bernoulli$.

Working model for **observed data** (**can be identified and tested**):

$$\Pr(y_{ij} = 1 \mid x_{ij}, u_i, (i,j) \in R) = p_y(x_{ij}, u_i) = \frac{\exp(\beta_0 + \beta_1 x_{ij} + u_i)}{1 + \exp(\beta_0 + \beta_1 x_{ij} + u_i)},$$

$$u_i \sim N(0, \sigma_u^2).$$

Working response model (**has to be assumed**):

$$p(y_{ij}, x_{ij}, \gamma) = \frac{\exp(\gamma_0 + \gamma_1 x_{ij} + \gamma_2 y_{ij})}{1 + \exp(\gamma_0 + \gamma_1 x_{ij} + \gamma_2 y_{ij})}.$$

The first expectation in **(3)** can be written as

$$E\{[p^{-1}(y_{ij},x_{ij};\gamma)-1]\frac{\partial \log[1-p(y_{ij},x_{ij};\gamma)]}{\partial\gamma}\,|\,x_{ij},u_i,(i,j)\in R\}=$$

$$p_y(x_{ij},u_i)[p^{-1}(1,x_{ij};\gamma)-1]\frac{\partial \log[1-p(1,x_{ij};\gamma)]}{\partial\gamma}+$$

$$[1-p_y(x_{ij},u_i)][p^{-1}(0,x_{ij};\gamma)-1]\frac{\partial \log[1-p(0,x_{ij};\gamma)]}{\partial\gamma}.$$

- Similarly for the second expectation in **(3)**.

$p_y(x_{ij})$ and $\hat{u}_i$ easily estimated by **SAS PROC NLMIX** and **(3)** is solved for γ by **SAS PROC NLIN**.

$$E[(w(y_{ij},x_{ij})-1)y_{ij}\,|\,x_{ij},u_i,(i,j)\in R]=p_y(x_{ij},u_i)(w(1,x_{ij})-1).$$

- Similarly for other expectations in **(5)** and **(6)**.

Simulation Study

Step 1: Generate finite population from **Population model:**

$y_{ij} \sim Bernoulli$,

$$\Pr(y_{ij}=1 \mid x_{ij}, u_i^U, (i,j)\in R) = p_y(x_{ij}, u_i^U) = \frac{\exp(-1 + x_{ij} + u_i^U)}{1+\exp(-1+x_{ij}+u_i^U)},$$

$u_i^U \sim N(0,1)$.

$M = 300,\ N_i = \text{int}[1000\exp\{\min[2.5, \max(\text{-}2{,}5, u_i^U)]/5\}]$,

$x_{ij} \sim Uniform(0,2)$.

Group areas into **3 sets**,

G1={i=1,..,100}, G2={i=101,..,200}, G3={i=201,..,300}.

Step 2: Sampling scheme:

Select m=150 areas by systematic PPS proportional to area size N_i (**informative sampling**).

Select **20** units from each selected area in G1,

40 units from each selected area in G2,

60 units from each selected area in G3,

by PPS sampling proportional to $z_{ij} = .5 + x_{ij} + 3y_{ij}$ (**informative sampling**).

Step 3: Response:

Each selected unit responds with probability

$$p(y_{ij}, x_{ij}, \gamma) = \frac{\exp(-.5x_{ij} + y_{ij})}{1 + \exp(-.5x_{ij} + y_{ij})}.$$

Step 4: Estimate $\hat{p}_y(x_{ij}, \hat{u}_i) = \hat{\Pr}(y_{ij} = 1 | x_{ij}, \hat{u}_i, (i, j) \in R)$ assuming **Logistic Mixed Model** for the respondents, applying PROC NLMIX with default options (Empirical Bayes).

Step 5: Assume working response model,

$$p(y_{ij}, x_{ij}, \gamma) = \frac{\exp(\gamma_0 + \gamma_1 x_{ij} + \gamma_2 y_{ij})}{1 + \exp(\gamma_0 + \gamma_1 x_{ij} + \gamma_2 y_{ij})}$$

. Substitute $\hat{p}_y(x_{ij}, \hat{u}_i)$ into **(3)** and estimate γ by use of **PROC NLIN**.

Estimate $w(y_{ij}, x_{ij}) = E[\tilde{\pi}_{j|i}^{-1} \mid y_{ij}, x_{ij}, (i,j) \in R]$ as follows:

$$E[\tilde{\pi}_{j|i}^{-1} \mid y_{ij}, x_{ij}, (i,j) \in R] = p(y_{ij}, x_{ij}) E[\pi_{j|i}^{-1} \mid y_{ij}, x_{ij}, (i,j) \in R];$$

$$\pi_{j|i} = n_i z_{ij} / \sum_{j=1}^{N_i} z_{ij} = \frac{n_i}{N_i} z_{ij} \underbrace{\left(N_i / \sum_{j=1}^{N_i} z_{ij} \right)}_{\approx \text{Constant}} \Rightarrow z_{ij}^{*} = \pi_{j|i} \frac{N_i}{n_i} \prec z_{ij}.$$

Fit the model $z_{ij}^{*} = g_{\alpha}(y_{ij}, x_{ij})$ (linear model in our study), estimate the parameters of this model and then estimate,

$$\hat{w}(y_{ij}, x_{ij}) = \hat{E}[\tilde{\pi}_{j|i}^{-1} \mid y_{ij}, x_{ij}, (i, j) \in R] \cong$$

$$\left[\frac{n_i}{N_i} g_{\hat{\alpha}}(y_{ij}, x_{ij}) \right]^{-1} p(y_{ij}, x_{ij}; \hat{\gamma}).$$

Calculate ratio of expectations in **(5),**

$$\hat{p}_y^{R^c}(x_{ik}, \hat{u}_i) = \hat{E}\{\frac{\hat{E}[(\hat{w}(y_{ik}, x_{ik}) - 1) y_{ik} \mid x_{ik}, u_i, (i,k) \in R]}{\hat{E}[(\hat{w}(y_{ik}, x_{ik}) - 1) \mid x_{ik}, u_i, (i,k) \in R]} \mid O\}.$$

Estimators considered (selected areas):

1. $\hat{\bar{Y}}_i^{ign} = N_i^{-1}\{ \sum_{j,(i,j)\in R} y_{ij} + \sum_{k=1,k\notin R}^{N_i} \hat{p}_y(x_{ij}) \}$

2. $\hat{\bar{Y}}_i^{H,MCAR} = \sum_{j,(i,j)\in R} \pi_{j|i}^{-1} y_{ij} \,/ \sum_{j,(i,j)\in R} \pi_{j|i}^{-1}$

3. $\hat{\bar{Y}}_i^{H,MAR} = \sum_{j,(i,j)\in R} \hat{w}(x_{ij}) y_{ij} \,/ \sum_{j,(i,j)\in R} \hat{w}(x_{ij}),\quad \hat{w}(x_{ij}) = [\pi_{j|i} p(x_{ij}, \hat{\lambda})]^{-1},$

4. $\hat{\bar{Y}}_i^{H,new} = \sum_{j,(i,j)\in R} \hat{w}(y_{ij}, x_{ij}) y_{ij} \,/ \sum_{j,(i,j)\in R} \hat{w}(y_{ij}, x_{ij}),$

5. $\hat{\bar{Y}}_i^{new} = N_i^{-1}\{ \sum_{j,(i,j)\in R} y_{ij} + \sum_{k=1,k\notin R}^{N_i} \hat{p}_y^{R^c}(x_{ij}, \hat{u}_i) \}.$

Repeat Steps 1-5 independently 1000 times.

Statistics considered:

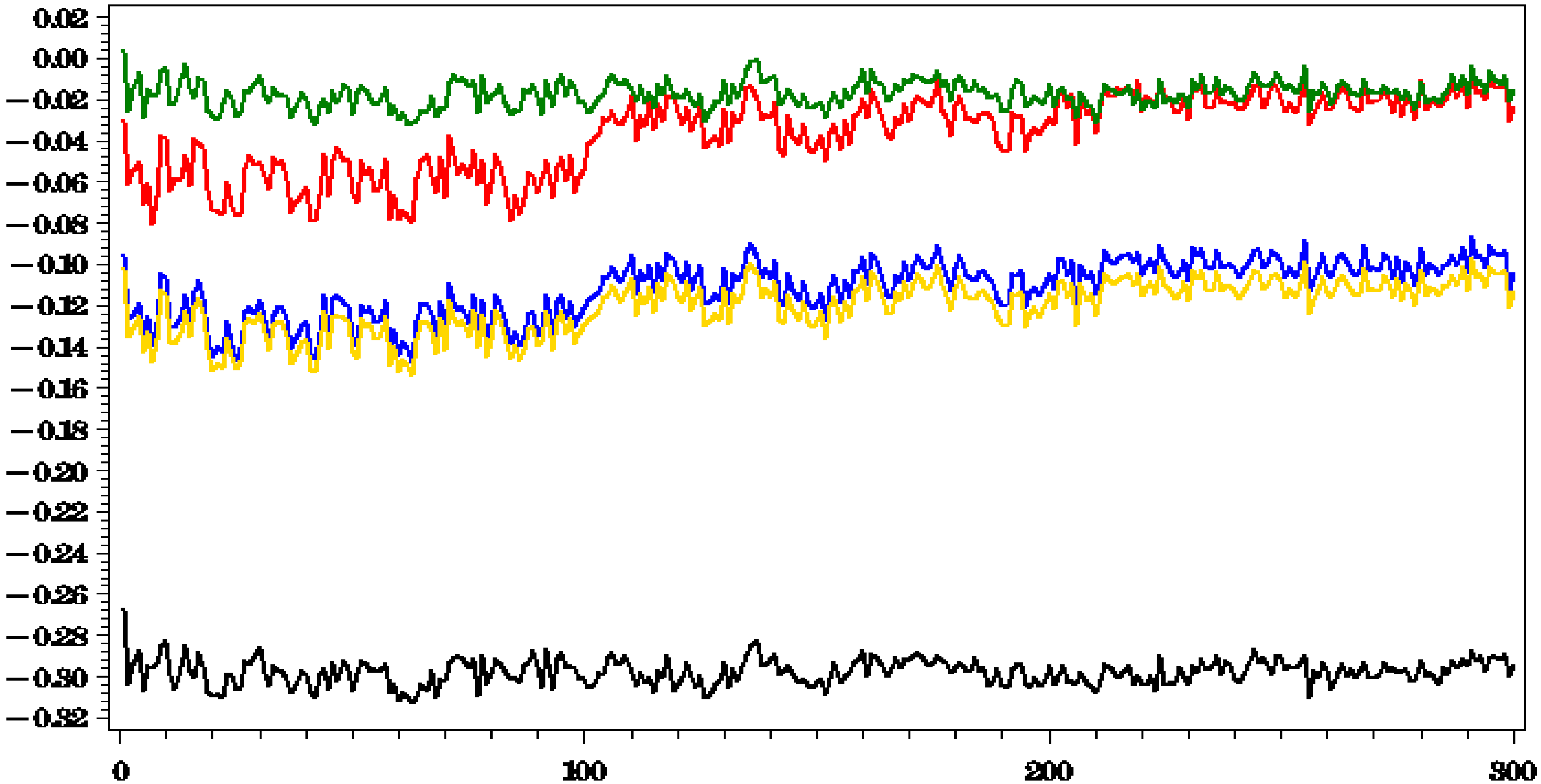
$$Bias_i = \frac{\sum_{r=1}^{1000} D_{ir}(\hat{\bar{Y}}_{ir} - \bar{Y}_{ir})}{\sum_{r=1}^{1000} D_{ir}}$$

$$RMSE_i = \sqrt{\frac{\sum_{r=1}^{1000} D_{ir}(\hat{\bar{Y}}_{ir} - \bar{Y}_{ir})^2}{\sum_{r=1}^{1000} D_{ir}}}$$

$D_{ir} = 1$ if area i **selected** on ***r*-th** simulation.

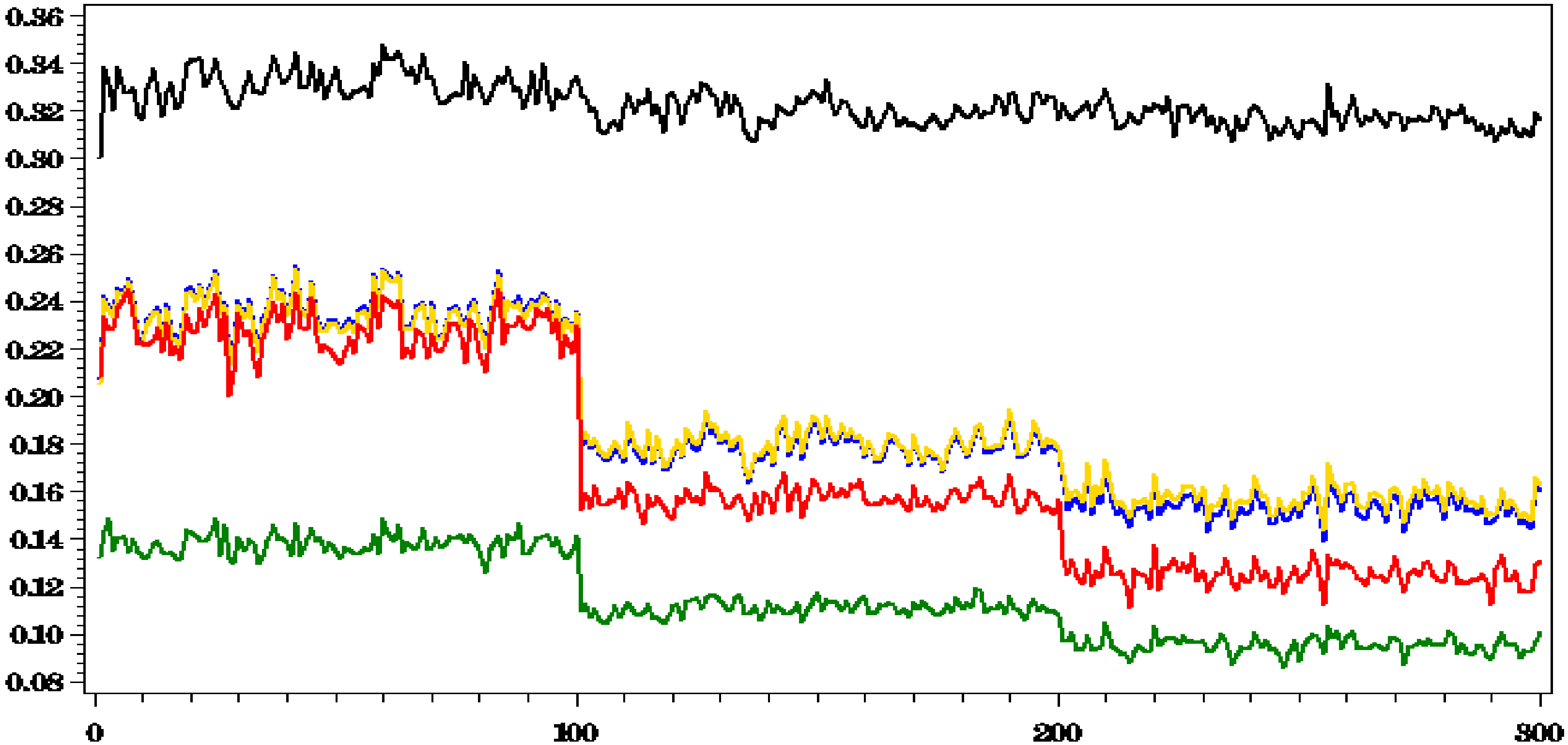
<u>Biases</u>: $\hat{\bar{Y}}_i^{ign}$ - black, $\hat{\bar{Y}}_i^{H,MCAR}$ - **gold,** $\hat{\bar{Y}}_i^{H,MAR}$ **- blue,**

$\hat{\bar{Y}}_i^{H,new}$ **- red,** $\hat{\bar{Y}}_i^{new}$ **- green**



RMSE's: $\hat{\bar{Y}}_i^{ign}$ - black, $\hat{\bar{Y}}_i^{H,MCAR}$ - gold, $\hat{\bar{Y}}_i^{H,MAR}$ - blue,

$\hat{\bar{Y}}_i^{H,new}$ - red, $\hat{\bar{Y}}_i^{new}$ - green



THANKS !!! (Sverchkov.Michael@bls.gov)