IASET: Journal of Applied Mathematics & Statistical Sciences (IASET: JAMSS)
ISSN(P): Applied; ISSN(E): Applied;
Vol. 1, Issue 1, Jan - Jun 2016, 31-42

International Academy of Science, Engineering and Technology
Connecting Researchers; Nurturing Innovations

IMPLICIT SMALL AREA MODELS WITH APPLICATIONS ION AGRICULTURE

NAGEENA NAZIR, S. A. MIR, T. A. RAJA & M. IQBAL JEELANI

Division of Agricultural Statistics, Skuast-K, Shalimar, India

ABSTRACT

In small area estimation the question is often about the trade of between bias and variance. With small sample sizes the unbiasedness of the direct estimators may be of no practical value due to large variance of the estimator. The model-based estimators are prone to bias, but they have the advantage of small variances compared to the design-based estimators. There is evidence that the model-based small area estimators outperform the direct estimators with respect to the estimation accuracy measured with mean squared error (MSE) (Torabi and Rao, 2008). This is possibly why the model-based approach is widely accepted as the framework for small area estimation. In this paper we have obtained direct, synthetic and composite estimators on real agricultural data set and results obtained from these estimators are compared in terms of average relative bias, average squared relative bias, average absolute bias, average squared deviation as well as the empirical mean square error. It has been found that composite estimator works better than direct and synthetic estimators. The above discussed methods are illustrated practically with the help of SAS and R software on the basis of newly developed functions piest (), composite (), relativebias (), absolute bias ().

KEYWORDS: Model-Based Estimation Methods, Synthetic Estimates

1. INTRODUCTION

Small area model-based estimation methods can be broadly divided into two groups methods based on implicit linking models and methods based on explicit linking models. Indirect estimators produced by implicit linking models (synthetic and composite estimators) are based on the assumption that there is an adequate direct estimator for a larger area that one can “borrow strength” from to produce indirect estimators for the small areas. These estimators are typically design-based in the sense that survey weights are used and the sample design induces the probability distribution that is used for determination of confidence intervals and standard errors. The major drawback of implicit linking models is the assumption that small areas possess the same characteristics as larger areas. Typically this is not true and the resulting estimators will be exposed to bias (Jiango et al., 2013).

2. DIRECT ESTIMATOR

Direct estimator provides estimates based only on the local data assuming that the sample is large enough which seldom happens in practice. Direct estimator is the most basic estimator and can only be used when all the areas have been sampled. For the area mean value it is as follows:

$$\hat{\bar{Y}}_{i,DIRECT} = \sum_j w_{ij} y_{ij} / \sum_j w_{ij} \qquad 2.1$$

The weights w_{ij} have been taken as the inverse of the probability of an individual to be in the sample. Note that

since all areas are sampled independently and with replacement, the probability of selecting individual *j* in area *i* is 1/*Ni*, where *Ni* is the number of individuals in area *i*. Thus the weight w_{ij} may be interpreted as the number of elements in the population represented by the sample element. The choice w_{ij} satisfies the unbiasedness condition and leads to the well known Horvitz Thompson (H-T) estimator. If the sample size in region *i* is n_i, the probability of selecting an individual at least once is $1-\left(1-\frac{1}{N_i}\right)^{n_i}$ This is the inclusion probability and we will use weights

$$w_{ij}^{-1} = w_i^{-1} = 1-\left(1-\frac{1}{N_i}\right)^{n_i} \qquad 2.2$$

Direct estimators are generally used when the sample size for each small area is sufficiently large to give reasonably accurate estimates. However, as the sources of data are usually sample surveys designed to give national and regional statistics, sample sizes for the small areas (usually sub domains of the original domains of study) are usually unduly small. Consequently, the associated variances are likely to be unacceptably large since the conditional variances (as can be seen above) are of the order n_i^{-1}. Moreover, if information from a national sample is used to make estimates for small areas and there are no sample units in the small area of interest, then obviously direct estimation cannot be used.

The variance of the direct estimator, which is also known as design variance, can be estimated to assess the uncertainty about the estimates. This can be used to provide approximate confidence intervals. The design variance of the direct estimator (2.1) is

$$V[\hat{\bar{Y}}_{i,DIRECT}] = (1-1/N_i)S_i^2/n_i \qquad 2.3$$

Here, S_i^2 is the variance of the sample obtained from area *i*. The variance can be estimated by

$$\hat{V}[\hat{\bar{Y}}_{i,DIRECT}] = (1-1/N_i)\hat{S}_i^2/n_i \qquad 2.4$$

That is, we substitute the variance of a generic sample S_i^2 by the actual variance of the observed data $\hat{S}_i^2$

3. SYNTHETIC ESTIMATOR

The term "synthetic estimates" was first used by the U.S. National Centre for Health Statistics (1968) of the United States when it calculated estimates of long and short term physical disabilities based on the National Health Interview Survey. Since then, synthetic estimation has been used to generate small area statistics from a number of surveys. More recently, small area synthetic estimates of literacy rates, health and morbidity statistics and income have been generated for purposes of local planning.

The method of synthetic estimation has been described by Gonzales (1973) as follows:

"An unbiased estimate is obtained from a sample for a large area; when this estimate is used to derive estimates for subareas on the assumption that the small areas have the same characteristics as the larger area, we identify these

estimates as synthetic estimates."

This is the method of "borrowing information from related subareas in order to increase the effective sample size for estimation and hence the accuracy of the resulting estimates" (Smith and Tomberlin, 1979). In contrast with the earlier mentioned methods which assume the availability of a known or estimated total for the variable of interest *y* at the subgroup level, the method proposed by Rao and Choudry (1995) uses an auxiliary variable *x* with known or estimated small area totals. The synthetic estimator is based on assuming a (linear) model for the data so that the values of the areas that have not been sampled are estimated from the model using only information for available covariates. For the mean, the synthetic estimator is based on the following model:

$$\overline{Y}_i = BX'_i + u_i \qquad 3.1$$

Where *ui* is an area-based random error, which is normally distributed with zero mean and variance σ_u^2 If the domain specific auxiliary information is available in the form of known totals Xi, then the regression synthetic estimator $X_i'\hat{B}$, can be used as an estimator of domain total Y_i.

$$\hat{Y}_{i,SYNTH} = \hat{B}X_i' \qquad 3.2$$

Where $\hat{B}$ is given by $(\hat{B}_1,...,\hat{B}_p)^T$ the p-bias of Y is approximately equal to $x_i^T B - y_i$ where B is the population regression coefficient. This p-bias will be small relative to yi is the domain specific regression coefficient Bi= $(\sum x_j x_j' / c_j)^{-1}(\sum x_j y_j / c_j)$ is close to B and $y_i = x_i' B_i$ Thus the synthetic regression coefficient will be very efficient when the small area does not exhibit strong individual effect with respect to the regression coefficient. Since that this estimator doesn't make any use of the random effects *ui* and that for this reason it may lead to biased estimates of the area means.

An advantage of the synthetic estimation is its ease of calculation. The variance of the synthetic estimator is of order *n* -1 and, hence, is smaller than that of the direct estimator. However, the synthetic estimates are biased estimates for two reasons. First, the underlying assumption of homogeneity of rates or proportions is often hard to satisfy, i.e., estimated rates for the larger area (for a particular subgroup *j*) may differ from that of one or more subareas. In other words, the "model assumption" that relations observed in large areas must hold for the small domains may not be always valid. Second, the structure of the population may have changed since the previous census. The synthetic method also fails to account properly for local factors. Unless the grouping variables are highly correlated with the variable of interest, the synthetic estimates will tend to cluster near the mean for the larger area, and fail to reflect the actual effects of local area factors.

4. COMPOSITE ESTIMATOR

When small area samples are relatively small, the synthetic estimators outperform the simple direct estimators; however, when small area sample sizes are large, the direct estimators outperform the synthetic estimators. Thus it was concluded that a weighted sum of these two (2) estimators would be better than choosing one over the other.

The composite estimator is constructed as a weighted sum of the direct estimator and the synthetic estimator

(Ghosh and Rao, 1994; Suciu *et al.*, 2001). The weights are defined so that if the sample size is "large" the direct estimate is given more weight than the synthetic one and when the sample is not reliable, the synthetic estimate will be given more weight. Thus a natural way to balance the potential bias of a synthetic estimator, say $\hat{Y}_{i,SYNTH}$ against the instability of a direct estimator, say $\hat{Y}_{i,DIRECT}$ is to take the weighted average of $\hat{Y}_{i,DIRECT}$ and $\hat{Y}_{i,SYNTH}$. Such composite estimator of small area total may be written as:

$$\hat{\bar{Y}}_{iCOMP} = \phi_i \hat{\bar{Y}}_{iDIRECT} + (1-\phi_i)\hat{\bar{Y}}_{i.STNTH} \quad 4.1$$

For a suitably chosen weight $\phi_i (0 \le \phi_i \le 1)$ which controls the shrinkage of the two estimators. That is, depending on how large is the sample in the small area it will give more weight to the direct estimate (if the sample is large) or to the synthetic estimate (if information is needed from other areas). The design MSE of the composite estimator is given by

$$MSE_p(\hat{Y}_{i,COMP}) = \phi_i^2 MSE_p(\hat{Y}_{iDIRECT}) + (1-\phi_i)^2 MSE_p(\hat{Y}_{iSYNTH}) + 2\phi_i(1-\phi_i)E_p(\hat{Y}_{iDIRECT} - Y_i)(\hat{Y}_{iSYNTH} - Y_i)) \quad 4.2$$

By minimizing (2.2.4) with respect to ϕ_i, we get the optimal weight ϕ_i as

$$\phi_i^* \approx MSE_p(\hat{Y}_{iSYNTH})/[MSE_p(Y_{iDIRECT}) + MSE_p(\hat{Y}_{iSYN})] \quad 4.3$$

The approximate optimal weight ϕ_i^* depends only on the ratio of the MSEs

$$\phi_i^* = 1/(1+F_i) \quad 4.4$$

Where $F_i = \dfrac{MSE_p(\hat{Y}_{iDIRECT})}{MSE_p(\hat{Y}_{iSYNTH})}$

It is easy to show that $\hat{Y}_{i,COPM}$ is better than either component estimator in terms of MSE when $\max(0,\ 2\phi_i^* - 1) \le \phi_i \le \min(2\phi_i^*, 1)$. the latter interval reduces to the whole range $0 \le \phi_i \le 1$. When $F_i = 1$, and it becomes narrower as Fi deviates from 1. The optimal weight ϕ_i^* will be close to zero or one when one of the component estimators has a much larger MSE than the other that is when Fi is either large or small. In this case the estimator with large MSE adds little information and therefore it is better to use the component estimator with small MSE in preference to the composite estimator. In practice we use either a prior guess of the optimal value of ϕ_i^* or estimate it from the sample data. Royall (1978) stipulates that the mean square error of the composite estimator is smaller than the larger of the mean squared errors of the two component estimators. Thus mean squared error of the composite estimator is smaller than that of either component estimator when an "appropriate weighting system" is used.

5. NUMERICAL ILLUSTRATION

In this section, we make an empirical comparison between direct, synthetic and composite estimators. The

performance of different estimators is examined from the accuracy of the point estimates standpoints. This is considered through the relative bias and absolute relative bias of different estimators. The different estimators mentioned above are compared according to four different criteria recommended by the panel on small area estimates of population and income set up by the United States committee on National Statistics (1978), Ghosh *et al.* (1996), Datta *et al.* (2002) viz., average relative bias, average squared relative bias, average absolute bias and average squared deviation.

Data collected through pilot survey conducted by the Division of Agri-Statistics on estimation of area and yield of apple in District Baramulla has been used for the purpose of our proposed small area estimation. The district Baramulla comprises of 12 blocks viz., Zanigeer, Boniyar, Tangmarg, Wagoora, Sopore, Baramulla, Uri, Pattan, Rohama, Singphora, Rafiabad and Kunzer. Each block consists of different number of villages. A fixed number of five villages were selected at random from each block by simple random sampling. The data set was named apple-1 for analysis and modeling in R/SAS software's. The same data set has further been condensed by taking average over all the villages to obtain block-wise data and the new data set obtained is named as Apple-2 for analysis and modeling in R/SAS software's. This data set has 13 rows and 6 columns and has been used for Area level modeling. The columns names are Blocks, N, Yield, Area, Trees, Actual Yield for names of blocks, total number of villages in each block, yield of apple from each block in metric tons, area under apple orchards, total number of apple trees in each block and actual yield obtained as per departmental records.

Suppose act_i denotes the true value of the variable for the ith small area, and est_i is any estimate of act_i $i = 1,2,...,m$.

Then average relative bias

$$ARB = \frac{1}{m}\sum_{i=1}^{m}\left|\frac{est_i - act_i}{act_i}\right|$$

Average squared relative bias

$$ASRB = \frac{1}{m}\sum_{i=1}^{m}\left(\frac{est_i - act_i}{act_i}\right)^2$$

Average absolute bias

$$AAB = \frac{1}{m}\sum_{i=1}^{m}\left|est_i - act_i\right|$$

Average squared deviation

$$ASD = \frac{1}{m}\sum_{i=1}^{m}\left(est_i - act_i\right)^2$$

Now using the above four criteria on the apple data set discussed above the results obtained are summarized as:

Table 1: Comparison of Estimators using Different Criteria

Criteria / Estimators	ARB	ASRB	ABS	ASD
Direct	0.1322	0.0221	182.25	48505.82
Synthetic	0.1068	0.0150	141.86	29167.39
Composite	0.0851	0.0099	109.57	18239.42

The results in Table 1 report the values of ARB, ASRB, AAB, and ASD for the Apple data set. It is clear from the value that composite estimate performed significantly better than the synthetic and direct estimates in terms of the entire four criterion. Also the percent Average relative bias is 8.51% with composite compared to 10.68% for synthetic and 13.22% for direct estimator. Similar the value of average absolute bias is 109.57 for composite estimator compared to 141.86 and 182.25 for synthetic and direct estimator respectively.

An Empirical comparison of direct, synthetic and composite estimators for all the 12 small areas separately using Percent Absolute Relative Bias and Absolute Bias is shown in the Table 2.

Table 2: Empirical Comparison of Estimators for all the 12 Small Areas Considered

Estimator	Direct		Synthetic		Composite	
Small Areas	ARB	AB	ARB	AB	ARB	AB
1.	5.97	179.66	5.31	159.66	4.27	128.53
2.	9.42	65.89	10.58	73.89	11.36	79.44
3.	18.20	104.95	16.98	97.95	15.67	90.11
4.	12.01	105.08	10.64	93.08	9.61	84.04
5.	6.70	246.66	5.18	190.66	4.45	163.90
6.	11.92	84.98	10.14	72.29	5.80	41.35
7.	4.83	30.30	1.61	10.04	1.01	9.39
8.	25.54	433.21	23.00	390.21	18.00	305.40
9.	18.70	335.62	12.28	220.56	10.95	196.61
10.	20.10	350.03	11.14	194.05	6.46	112.61
11.	4.91	118.18	3.46	83.175	0.67	16.14
12.	20.32	132.78	17.87	116.78	13.02	85.08

From the Table 2 it is evident that composite estimates exhibit smaller errors and a lower incidence of extreme error than either of the Direct and Synthetic estimates. The value of percent absolute relative bias and absolute bias for composite is also low as 0.67% and 9.39 in comparison to 3.46%, 4.91% and 10.04, 30.30 in synthetic and direct estimator respectively.

Figure 1 shows the comparison of the values of percent relative bias and absolute bias for composite, synthetic and direct estimators.

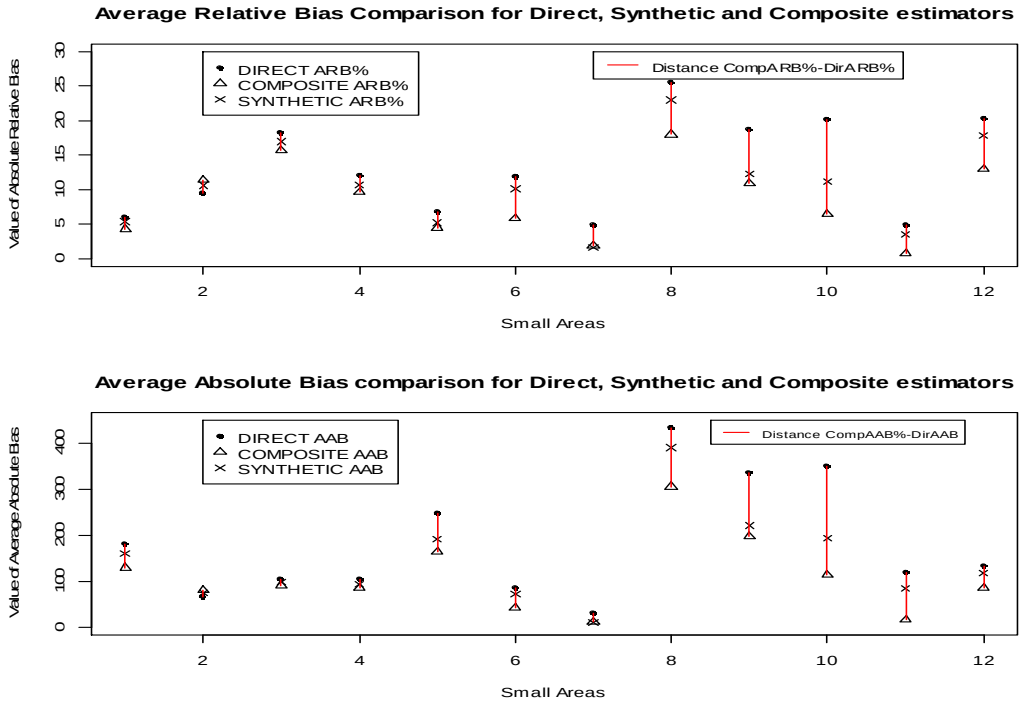

Figure 1: Comparison of Percent ARB and AB of Composite, Synthetic and Direct Estimators

Figure1 displays the deviations of Synthetic and Direct estimators from the composite estimator. Significant disparity is observed among the three estimators. The performance of the composite estimator is the best as it provides the lowest value of both %ARB and AB for each of the small area s compared to the other two estimators.

Table 3: Mean Square Error (MSE) of Estimators of Variance Components for 12 Small Areas

Estimators / Small Areas	Direct	Synthetic	Composite
1.	32277.72	25491.32	16519.96
2.	4341.49	5459.73	6310.71
3.	11014.50	9594.20	8172.16
4.	11041.81	8663.88	7062.77
5.	60841.16	36351.24	26863.21
6.	7221.6	5225.84	1709.82
7.	901.80	100.80	129.73
8.	187670.9	152263.8	93269.16
9.	112640.8	48646.71	38655.49
10.	122521.0	37655.40	12681.01
11.	13966.51	6918.08	260.49
12.	17630.53	13637.57	7238.60

Table 3 reports the different MSE estimates for each of the 12 small areas and it is clear that in terms of MSE the performance of composite estimator is the best. In other words we can emphatically say that the composite estimator performs better than the two estimators.

Table 4: Direct, Synthetic and Composite Estimates of Population Parameters and Their Associated Standard Errors (S.E)

Estimators / Small Areas	Direct		Synthetic		Composite		Actual
	Estimate	S.E	Estimate	S.E	Estimate	S.E	
1	3186.46	179.66	3166.46	159.66	3135.33	128.53	3006.80
2	633.05	65.89	625.05	73.89	619.50	79.44	698.94
3	471.59	97.95	478.59	104.95	486.14	90.4	576.54
4	769.36	105.08	781.36	93.08	790.40	84.04	874.44
5	3429.94	246.66	3485.94	190.66	3512.70	163.90	3676.60
6	627.82	84.98	640.51	72.29	671.45	41.35	712.80
7	591.15	30.03	611.14	10.04	609.79	11.34	621.18
8	2129.35	433.21	2086.35	390.21	2001.54	305.4	1696.14
9	1459.12	335.62	1574.18	196.61	1598.13	220.56	1794.74
10	2090.89	350.03	1934.91	194.05	1853.47	112.61	1740.86
11	2285.22	118.18	2320.22	83.17	2387.26	16.14	2403.40
12	520.42	132.78	536.42	116.78	568.12	85.08	653.20

Table-4 reports the Direct, Synthetic and Composite estimates and their associated standard errors for all the 12 small areas separately. As can be from the values, the Composite estimates are close to the actual values as compared to Synthetic and Direct estimates. Thus it can be concluded that composite estimator performed better than synthetic and direct estimators, same is true for the associated standard errors of the three estimators.

Figure 2 plots the point estimates of θ_i against the small areas and also provides a comparison of these values with the actual value of yield obtained in each of the small area.

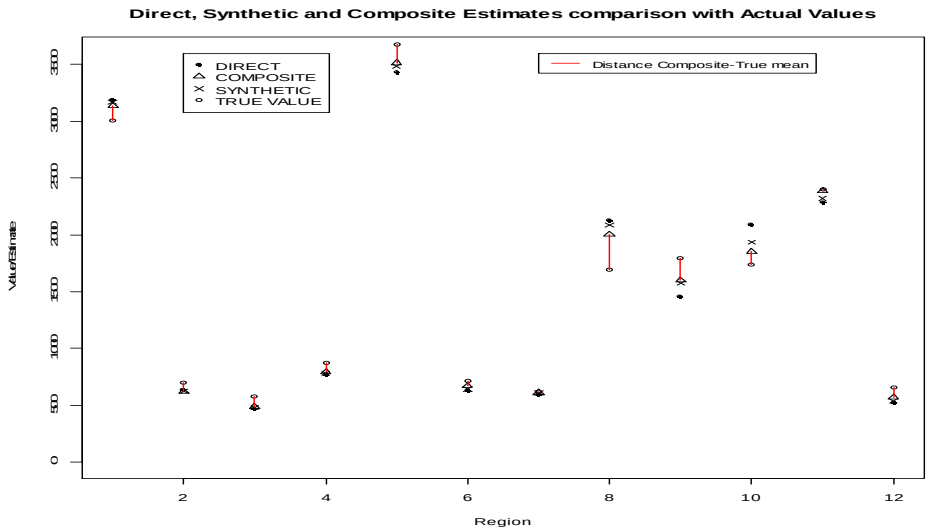

Figure 2: Composite, Synthetic and Direct Estimates Compared to the True Means

Figure 2 displays Composite, Synthetic and direct estimates and their deviation from the actual mean. Here we can see that the values of θ_i obtained by composite estimator are closer to actual values as compared to direct and synthetic estimators. Thus for the plot also we conclude that among the three techniques discussed the composite is the best technique for obtaining the estimates.

6. CONCLUSIONS

In this paper we have provided a broad overview of small area estimation, its usefulness and application in a wide variety of settings, model based approaches and several methods for estimation of variance components which plays an important role in obtaining reliable small area estimates and the associated measure of uncertainties. And it has been found that composite estimator worked better than direct and synthetic estimators.

REFERENCES

1. (Torabi and Rao, 2008 (Smith and Tomberlin, 1979).
2. Becker, R.A. Chambers, J.M. and Wilks, A.R. 1988. *The New S Language*. Chapman and Hall, New York.
3. Bell, W. 1999. Accounting for uncertainty about variances in small area estimation. *Bulletin of the International Statistical Institute* **52**: 25-30.
4. Chambers, R., Chandra, H., Salvali, N. and Tzaidis, N. 2014. Outlier robust small area estimation. *Journal of the Royal Statistical Society**:** Series B* **76**(1): 47-69.
5. Datta, G.S. and Ghosh, M. 2012. Small area shrinkage estimation. *Journal of Statistical Science* **27**(1): 98-114.
6. Datta, G.S., Bell, W.R. and Ghosh, M. 2012. Benchmarking small area estimators. *Journal of Statistical Sciences* **28**(1): 1-134.
7. Datta, G.S., Rao, J.N.K. and Smith, D.D. 2005. On measuring the variability of small area estimators under a basic area level model. *Biometrika* **92**: 183-196.
8. Ghosh, M., Nangia, N. and Kim, D. 1996. Estimation of Median Income of Four-person Families: A Bayesian Time Series Approach. *Journal of the American Statistical Association* **91**: 1423-1431.
9. Gonzalez, M.E. 1973. "Use and Evaluation of synthetic Estimates". **In:** *Proceedings of the Social Statistics Section, American Statistical Association*, USA, pp. 33-36.
10. Haslett, S.J., Isidro, M.C. and Jones G. 2010. Comparison of survey regression techniques in the context of small area estimation of poverty. *Survey Methodology* **36**(2): 157-170.
11. Jiang, J. 2007. *Linear and Generalized Linear Mixed Models and their Applications*. Springer.
12. Jiango, V.D., Haziza, D. and Duchasne, P. 2013. Controlling the bias of robust small area estimators. *NATSEM* **9**: 23-30.
13. Pfeffermann, D. 2013. New important development in small area estimation. *Journal of Statistical Sciences* **28**(1): 40-68.
14. Pinho, L.G.B., Nobre, S.J. and Freilas, S.M. 2012. On linear mixed models and their influence diagnostics applied to an actuarial problem. *Chilean Journal of Statistics* **3**(1): 57-73.
15. R Development Core Team, 2008. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.
16. Rao, J.N.K. 2003. *Small Area Estimation*, New Jersey, John Wiley & Sons, Inc.

17. Rao, J.N.K. and Choudry, G.H. 1995. Small Area Estimation: Overview and Empirical Study. **In**: *Business Survey Methods*. [Eds. B.G. Cox *et al.*]. John Wiley & Sons, New York, pp. 527-542.

18. Royall, R.M. 1971. Linear regression models in finite population sampling theory. **In**: *Foundations of Statistical Inference.* (Proc. Sympos., Univ. Waterloo, Waterloo, Ont., 1970). Rinehart and Winston of Canada, Toronto, Ont.: Holt, pp. 259-279.

19. Sharma, S.D., Srivastava, A.K. and Sud, U.C. 2004. Small area crop estimation methodology for crop yield estimates at Gram Panchayat level. *Journal of Indian Society of Agricultural Statistics* **57**: 26-38.

20. Suciu, G., Hoshaw-Woodard, S., Elliott, M. and Doss, H. 2001. Uninsured Estimates by County: A Review of Options and Issues (Tech. Rep.). Ohio: Ohio Department of Health, Center for Public Health Data and Statistics.

21. Venables, W.N. and Ripley, B.D. 2004. *Modern Applied Statistics with S-PLUS*, 4th edition, Springer Verlag, New York.

APPENDIX

Direct (data,N,n)

```
Direct<-function (data,N,n)
{
apple<-as.dataframe(apple)
probs<-1/N
probs1<-1-(1-probs)^n
weight<-1/probs1
yij<-by[apple$yield,apple$n,sum)
DE<-as.vector(yij)*(weight/(n*weight))
VD<-matrix(as.numeric(tapply(apple$yield,apple$n,var))*(1- 1/N)/n,ncol=1))
List(Direct Estimator=DE, Variance Direct=VD)
}
```

Composite Est(est(D), est(S), ActMean)

```
Composite<-function(est(d),est(s),Yt,var(yd))
{
Mse(d) <-1/n*(Yt-est(d))^2
Mse(s) <-((Yt-est(s))^2)*var(Yd)
Phi <- Mse(s)/((Mse(D)+Mse(s))
Est(c) <-(phi*est(d))+((1-phi)*est(s))
```

```
List(CompositeEstimate=est(c),MseDirect=Mse(D),Mse Synthetic=MSE(s))}
```

Relative Bias(est,act)

```
Rb<-function(est,true)
{
M<-length(est)
Arb<-formatC((est-true)/m,digits=2)
asrb<-formatC(sum((est-true)^2)/m,digit=2)
list(Average Relative Bias=arb, Average Squared Relative Bias=ASRB)}
```

Absolute Bias(est,act)

```
AB<-function(est,true)
{
M<-length(est)
AAB<-formatC((est-true),digit=2)
AASB<-formatC(sum((est-true)^2),digit=2)
list(AverageAbsoluteBias=asb,AverageSquaredDeviations=ASRB)}
```