# GENERATING SMALL AREA STATISTICS FOR HOUSEHOLD INCOME IN SOUTHERN PROVINCE OF SRI LANKA

By

Anoma S. Jayasekera [1] and W. N. Wickremasinghe [2]

[1]*Sample Surveys Division*
*Department of Census and Statistics*
*Maitland Crescent*
*Colombo 7, Sri Lanka*

[2] *Department of Statistics*
*University of Colombo*
*P. O. Box 1490*
*Colombo, Sri Lanka*

[This article was written in 2004 based on the Masters thesis submitted in 2003 by author (1) ]

## INTRODUCTION

Human development is a central objective of economic activity and forms the main purpose of social investment in any country. The Human Development Index (HDI) defined by the UNDP [1] is based on three dimensions: namely, opportunity to lead a long and healthy life, ability to acquire knowledge and learning, and power to access adequate resources. Sri Lanka's HDI is by far the highest in South Asia [2] and exceeds that of some developed countries. This high level of achievement is due to provisions made for accessing health and education, and continued investment in the social sectors. Country's economic development on the other hand is lagging consistently behind, and appears to have bypassed the rural sector where nearly 80% of the population live.

Human poverty [1] is a result of inability to meet the three social dimensions mentioned above. Sri Lanka's level of poverty is moderately high, with a high proportion (56%) having no electricity, a moderate proportion (28%) without access to safe drinking water, and a considerable proportion (24%) without adequate sanitation facilities.[3] The country has gone through several poverty-alleviation programs in the past. However, the success of any such program would largely depend on the availability of up-to-date and reliable poverty related information at small geographic levels. According to latest information [4]

88% of the poor resides in the rural areas compared to 8% in urban, and 4% in the estate. Another study [2] concluded that, this 4% is among the poorest in Sri Lanka. These figures tell us that more work is involved in getting reliable information from the rural sector, if these programs are to be effective.

"Samurdhi" is the largest single welfare program, which accounts for nearly 1% of the GDP, aimed at poverty alleviation in Sri Lanka.[2] Southern Province is reported to have a high percentage of Samurdhi recipients and therefore, this study is focused on the Southern Province. The effectiveness of any such program would largely depend on the reliability of systems that provide information at the smallest geographic levels that need food assistance. Censuses conducted every ten years and surveys conducted more often than that, are the only sources of information on poverty variables. Even though censuses provide reliable information on the smallest of geographic areas such as Divisional Secretariat (DS) or Grama Niladari (GN) levels, the information at these levels is updated only once in a decade. Surveys do not bridge this gap either, as information is available only at larger administrative units such as provincial or at most district, due to budgetary constraints.

The present work is motivated by the fact that in Sri Lanka no attempt has been made in the past to come up with reliable and up-to-date estimates of poverty related variables at small geographic levels. Therefore, in this paper we demonstrate how small area statistics could be generated by a method called small area estimation [5, 6, 7, ..] which has been successfully used elsewhere. The method combines information from censuses and surveys through a statistical approach. We have focused on an important variable, household income.

## METHODS AND MATERIALS

### *Small Area Estimation*

There is no unique methodology recommended for obtaining small area statistics for a given situation, and relevant theories are being developed constantly [6,7]. Many terms such

as "local area"[7], "small domain"[7], "sub-domain"[7], "small group"[7], "indirect"[7], and "synthetic"[6,7] are also found in literature as alternatives to "small area". What ever the name used, the objective is to generate statistics for small domains for which up-to-date and precise information is not available through censuses or surveys.

Consider for example an island-wide Household Income and Expenditure Survey (HIES). Suppose a sample of $n_i$ households from the $i^{th}$ district, i = 1, 2, …, 17 (excluding North and East) is contained in a total sample of size n = $\sum n_i$ . Let $s_i$ be the number of DS divisions per $i^{th}$ district where DS is the level at which reliable estimates are needed. It may happen that $n_{ij}$ ( $\geq 0$) households from the $j^{th}$ DS unit in $i^{th}$ district are included in the sample. However, the number $n_{ij}$ may be insufficient to represent the $j^{th}$ DS unit. As a result, any direct estimate for the $j^{th}$ DS unit based on this sample may have a large variance. In this paper, we use the composite estimator that is a combination of the two broad types of estimators, direct (or sample-based), and indirect (or model-based)[6,7] as given below.

### *1. Direct estimator*

A direct estimator uses values of the variable of interest only from the time period of interest and only from the units in the domain of interest.[6,7] For example, in the HIES carried out in 2002, a direct estimate of household income for the $j^{th}$ DS unit in $i^{th}$ district is given by the mean (or the median) of the $n_{ij}$ income values. Direct estimates are usually unbiased but suffer from large variances.

### *2. Indirect estimator*

Indirect estimators use values of the variable of interest from a domain and/or time period other than the domain and time of interest. [6,7] A common type is the regression estimator. Regression estimates are obtained subject to several conditions. [5,7, 8] Suppose average household income for $j^{th}$ DS unit in $i^{th}$ district is needed for year 2005. As there is no up-to-date estimate available, the information available from the last School Census (2002) and the last sample survey (2002) could be combined to get a reasonably good estimate. In particular, a multiple regression model of the form

$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + \varepsilon_i$, would be fitted after checking all diagnostics, where $Y_i$ is the average household income for i$^{th}$ district (i= 1, 2, …, 17) computed from the survey and $X$'s are the related k independent variables computed from the census; $\varepsilon_i$ is a random error. [6,7] Now the equation $\hat{Y}_i = b_0 + \sum_{\ell=1}^{k} b_\ell X_{\ell i}$ with the estimated parameters can be used to obtain an up-to-date estimate of household income, by plugging in the X values from the census corresponding to j$^{th}$ DS unit. One basic assumption used here is that the relationship between $Y$ and $X$'s at district level, holds true for DS levels as well. This may be a reasonable assumption, but in case it is violated the method has to be modified accordingly. These regression estimators are usually unbiased.

### *3. Composite estimator*

A composite estimator combines direct and indirect estimates to obtain a more precise estimate.[7] For two independent small area estimators $T_1$ and $T_2$ for estimating an unknown parameter $\theta$, a composite estimator is a weighted average of the two, given by $T_c = \omega T_1 + (1-\omega)T_2$, with weights $\omega$ and $1-\omega$. The result that $T_c$ too is unbiased when both $T_1$ and $T_2$ are unbiased, is straight forward. Further it can be easily shown that $T_c$ is better than both $T_1$ and $T_2$ in the sense of mean square error. [6,7,8,9]This follows from the fact that variance, which is the same as mean square error in this case, of $T_c$, is minimized when $\omega = \frac{V_2}{V_1 + V_2}$, and that $\mathrm{V}(T_c) = \frac{V_1 V_2}{V_1 + V_2}$, where $V_1$ and $V_2$ are the variances of $T_1$ and $T_2$, respectively. In our work, we use a composite estimate based on the direct and regression estimates discussed above, to generate small area statistics, average household income in particular, at DS level.

## RESULTS AND DISCUSSION

The nine variables from the School Census (2002) used in the regression model, are described in Table 1.

**Table 1 : Selected School Census Variables as Explanatory Variables at District Level**

| Description | Variable Name |
|---|---|
| 1. Percentage of schools with electricity facility | Electric |
| 2. Percentage of students who failed their classes | Failpct |
| 3. Percentage of schools with telephone facility | Telepho |
| 4. Percentage of schools with safe drinking water facility | Water |
| 5. Percentage of schools of type 1AB and 1C | Bettersc |
| 6. Percentage of primary students who went to pre-schools | Prepct |
| 7. Percentage of schools having photocopy machine | Photo |
| 8. Percentage of schools having computer facility | Comp |
| 9. Percentage of schools having library facility | Plib |

This section summarizes the results of modeling the nine explanatory variables given in Table 1 with mean monthly household income from the Household Income and Expenditure Survey (2002) as dependent variable. Using all possible regression models[10], best model was selected by examining criteria such as residual mean square, coefficient of determination and Mallows Cp Statistics[10] and other residual plots.

The estimated model was found to be

Income = 9618 + 8854 electric + 4188 water - 9579 prepct + 34408 comp, with an $R^2$ of 92.5%. Further, all coefficients of the model were significant at 5% level.

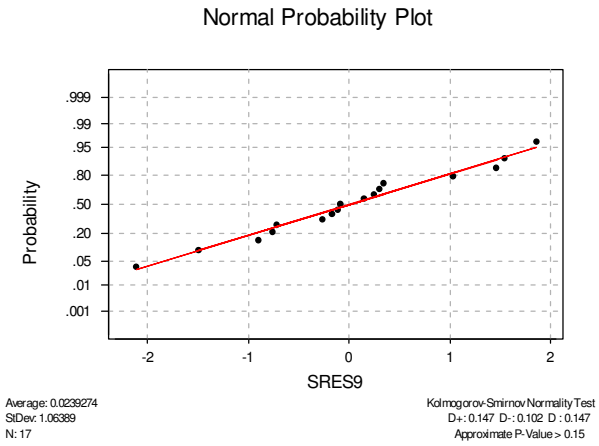
### Model Adequacy Checking

Normal probability plot of residuals (Figure 1) was used to check the assumption of normality for the regression model. It was observed that normality assumption is not violated. Further, it was found that the explanatory variables were moderately correlated. One major assumption in a multiple regression model is that, there is no linear dependency among the explanatory variables. When such dependency exists, we say that there is multicollinearity present. Most statistical packages automatically test for

multicollinearity using the Variance Inflation Factor (VIF). For our data, no significant presence of multicollinearity was found.

The two measures Leverage[10,11] and DIFITS[10,11] were calculated to check the goodness of fit of the estimated model. These values showed no evidence of any extreme cases according to the definitions. All diagnostics suggested that the final fitted model using the four explanatory variables, electric, water, prepct, and comp, was adequate.

Figure (1) : Normal probability plot of residuals




**Efficiency of the composite estimator**

In our study, $T_1$ is the regression estimate of the mean income and $T_2$ is the direct estimate obtained by sample values. Tables 1(a), 1(b) and 1(c) of the Annex 1 show standard errors of both direct and composite estimates and the corresponding values of Coefficient of Variation (CV) [12] .

As expected, the composite estimate had a mean square error that is smaller than that of the direct estimate. It is also evident that CV values of composite estimate is less than that of direct estimate's. Annex 2 shows both direct ($T_1$) and composite ($T_c$) estimates of household income for the small areas, DS-divisions, in the Southern province of Sri Lanka.

This technique may provide the best alternative for small area estimation in the absence of any sophisticated method. In Sri Lanka, the approach discussed above has not been attempted before. Therefore, this work is expected to be important and helpful for generating statistics for the population at domains smaller than what is provided by different surveys conducted in Sri Lanka. We hope that the methods discussed above can be easily extended to obtain estimates for the entire country for poverty related variables.

## REFERENCES


1. United Nations Development Program (1998), " National Human Development Report 1998", Sri Lanka.
2. Ministry of Finance (2001), "Poverty Reduction Strategy", Unpublished Policy Document, Colombo, Sri Lanka.
3. United Nations World Food Program (2000), "Vulnerability Analysis and Mapping Analysis of the Revised Secondary Data set", Colombo, Sri Lanka.
4. Department of Census and Statistics (1995/96), Household Income & Expenditure Survey, Sri Lanka.
5. Emiel B. Barrios (1998), Small Area Estimation of selected Socio-economic Indicators, MIMAP Research Paper No.36, Phillippines.
6. Emiel B. Barrios (1996), Generating Small Area Statistics from Household Surveys Conducted by National Statistics Office, MIMAP Research Paper No.29, Phillippines.
7. Gabriel S., Cike E., and Hani, D. (2001), "Uninsured Estimates by County: A review of Options and Issues", Center for Bio Statistics, The Ohio State University, Ohio.
8. U.S. Census Bureau, "Small Area Income and Poverty Estimates – 1995 Details", Internet: http://www,census.gov/hhes/www/saipe/techdoc/1998/98cntymod.html

9. U.S. Census Bureau, “Small Area Income and Poverty Estimates – 1993 Details”, Internet: http://www,census.gov/hhes/www/saipe/techdoc/1998/98cntymod.html
10. Shaible, W.L., (1996), Indirect Estimators in U.S. Federal Programs, Lecture Notes in Statistics 108, Springer-Verlag, New York.
11. N.R. Draper and S. Weisberg (1981), Applied Regression Analysis, Second Edition, John Wiley & Sons. Inc.

## Annex 1

## Standard Errors and Respective CV Values of Direct and Composite Estimates

**Table 1(a): Galle District**

| DS division | Standard Error | | CV (%) | |
|---|---|---|---|---|
| Galle District | Direct Estimate ($T_1$) | Composite Estimate ($T_c$) | Direct Estimate ($T_1$) | Composite Estimate ($T_c$) |
| Akmeemana | 1434.3 | 847.8 | 14.0 | 7.9 |
| Amblangoda | 1049.1 | 742.5 | 8.5 | 7.1 |
| Baddegama | 2374.7 | 961.1 | 19.6 | 8.4 |
| Balapitiya | 2916.6 | 988.8 | 18.9 | 5.5 |
| Bentota | 1745.0 | 900.3 | 13.7 | 10.6 |
| Bope-poddala | 1871.5 | 916.4 | 9.0 | 5.9 |
| Elpitiya | 976.2 | 715.3 | 8.9 | 5.9 |
| Galle | 4701.4 | 1025.7 | 19.4 | 11.9 |
| Habaraduwa | 978.3 | 716.1 | 8.2 | 5.8 |
| Hikkaduwa | 874..6 | 672.3 | 9.7 | 6.8 |
| Imaduwa | 385.8 | 362.2 | 3.8 | 3.6 |
| Karandeniya | 1823.3 | 910.6 | 18.1 | 8.2 |
| Nagoda | 877.4 | 673.6 | 13.2 | 7.2 |
| Neluwa | 552.0 | 488.8 | 5.3 | 4.7 |
| Niyagama | 3282.0 | 1000.9 | 21.1 | 9.9 |
| Thawalama | 1909.5 | 920.8 | 23.1 | 12.2 |
| Welvitiya-Divithura | 480.9 | 437.3 | 3.9 | 3.7 |
| Yakkalamulla | 1677.3 | 890.6 | 14.7 | 5.6 |

**Table 1(b): Hambantota District**

| DS division | Standard Error | | CV (%) | |
|---|---|---|---|---|
| Hambantota District | Direct Estimate ($T_1$) | Composite Estimate ($T_c$) | Direct estimate($T_1$) | Composite estimate($T_c$) |
| Ambalantota | 1164.4 | 780.2 | 13.3 | 7.1 |
| Angunukolapallassa | 1285.3 | 813.7 | 13.2 | 7.3 |
| Beliatte | 1769.3 | 903.6 | 19.6 | 8.7 |
| Hambantota | 1187.8 | 787.1 | 10.2 | 6.7 |
| Katuwana | 2157.7 | 944.9 | 27.2 | 15.4 |
| Lunugamvehera | 3054.8 | 993.9 | 27.9 | 11.8 |
| Okewela | 986.75 | 719.4 | 11.7 | 7.5 |
| sooriyaweva | 694.47 | 579.4 | 7.4 | 6.4 |
| Tangalle | 1345.9 | 828.4 | 14.2 | 6.4 |
| Thissamaharama | 984.57 | 718.6 | 13.0 | 5.8 |
| Weeraketiya | 1580.8 | 875.2 | 12.7 | 7.1 |

**Table 1 (c): Matara District**

| DS Division<br>Matara District | Standard Error | | CV (%) | |
|---|---|---|---|---|
| | Direct Estimate ($T_1$) | Composite Estimate ($T_c$) | Direct estimate($T_1$) | Composite estimate($T_c$) |
| Akuressa | 672.6 | 566.5 | 9.9 | 7.3 |
| Athuraliya | 4781.3 | 1026.5 | 32.8 | 9.2 |
| Devinuwara | 2039.5 | 934.3 | 16.5 | 9.5 |
| Dickwellla | 552.55 | 489.1 | 6.1 | 5.7 |
| Hakmana | 2692.3 | 979.1 | 35.0 | 9.2 |
| Kamburupitiya | 1640.1 | 884.9 | 12.6 | 8.7 |
| Kirinda | 1576.2 | 874.5 | 16.9 | 11.6 |
| Kotapola | 2147.3 | 944.0 | 18.2 | 8.7 |
| Malimboda | 1867.2 | 915.9 | 16.7 | 7.9 |
| Matara | 1605.4 | 879.4 | 11.2 | 7.1 |
| Mulatiyana | 1625.0 | 882.5 | 16.2 | 6.8 |
| Pasgoda | 1069.2 | 749.5 | 12.4 | 7.6 |
| Pitabeddara | 828.55 | 650.7 | 8.6 | 7.4 |
| Thihagoda | 1745.0 | 900.3 | 20.9 | 9.8 |
| Weligama | 1400.9 | 840.7 | 12.6 | 9.0 |
| Welipitiya | 855.59 | 663.5 | 12.4 | 8.8 |

## Annex 2

## Direct and composite estimates of household income

**Table 2(a): Galle District**

(Rupees)

| DS division | Direct Estimate ($T_1$) | Composite Estimate ($T_c$) |
|---|---|---|
| **Galle District** | | |
| Akmeemana | 10231 | 16572 |
| Amblangoda | 12310 | 16463 |
| Baddegama | 12148 | 12332 |
| Balapitiya | 15457 | 18529 |
| Bentota | 12733 | 13051 |
| Bope-poddala | 20641 | 21793 |
| Elpitiya | 11016 | 9000 |
| Galle | 24191 | 27573 |
| Habaraduwa | 11893 | 12735 |
| Hikkaduwa | 9022 | 14166 |
| Imaduwa | 10143 | 13193 |
| Karandeniya | 10070 | 10754 |
| Nagoda | 6659 | 11105 |
| Neluwa | 10392 | 12051 |
| Niyagama | 15578 | 8316 |
| Thawalama | 8282 | 8311 |
| Welvitiya- Divithura | 12104 | 7882 |
| Yakkalamulla | 11390 | 10135 |

**Table 2 (b) : Hambantota District**

(Rupees)

| DS division | Direct Estimate ($T_1$) | Composite Estimate ($T_c$) |
|---|---|---|
| **Hambantota District** | | |
| Ambalantota | 8764 | 9339 |
| Angunukolapallassa | 9763 | 8087 |
| Beliatte | 9027 | 7280 |
| Hambantota | 11685 | 8601 |
| Katuwana | 7929 | 10337 |
| Lunugamvehera | 10940 | 7504 |
| Okewela | 8445 | 11149 |
| Sooiyaweva | 9341 | 5445 |
| Tangalle | 9463 | 11017 |
| Thissamaharama | 7546 | 15826 |
| Weeraketiya | 12437 | 9582 |

Table 2 (c) : **Matara District**

**(Rupees)**

| DS division | Direct Estimate ($T_1$) | Composite Estimate ($T_c$) |
|---|---|---|
| **Matara District** | | |
| Akuressa | 6821 | 11092 |
| Athuraliya | 14579 | 12952 |
| Devinuwara | 12359 | 13971 |
| Dickwellla | 9104 | 14418 |
| Hakmana | 7691 | 11391 |
| Kamburupitiya | 13027 | 12937 |
| Kirinda | 9316 | 8364 |
| Kotapola | 11826 | 7696 |
| Malimboda | 11156 | 17908 |
| Matara | 14373 | 20264 |
| Mulatiyana | 10037 | 8188 |
| Pasgoda | 8632 | 10622 |
| Pitabeddara | 9683 | 9764 |
| Thihagoda | 8364 | 11760 |
| Weligama | 11110 | 9611 |
| Welipitiya | 6887 | 10324 |