Seminar Component

Name of authors:

1. Dr. Gamini De Silva (UNESCAP, consultant)
2. Dr. Chitraka Wickramarachchi (University of Sri Jayewardenepura, Sri Lanka)
3. Dr. Piyasena Hapuarachchi (University of Sri Jayewardenepura, Sri Lanka)
4. Dr. Darshi Thoradeniya (University of Colombo, Sri Lanka)
5. Mr. Priyath De Silva (Institute of Applied Statistics Si Lankaamini de Silva

Organization*:*

1. UNESCAP, consultant
2. University of Sri Jayewardenepura, Sri Lanka
3. University of Sri Jayewardenepura, Sri Lanka
4. University of Colombo, Sri Lanka
5. Institute of Applied Statistics Si Lankaamini de Silva

Contact address: Institue of Applied Statistics Si Lanka, OPA Building, Colombo 07, Sri Lanka

Contact phone: *94777541191, 94112927412*

Email: apgsdesilva@yahoo.com

Title of Paper

Improving data availability for economic empowerment of women in Sri Lanka:

A study on data integration for monitoring the SDGs

Abstract

Due to economic, social and cultural discrimination which limits women's economic participation, Female Headed Households (FHH) seem to be one of the most vulnerable groups and can be considered as a priority to women's empowerment in Sri Lanka. Thus, in this study, we decided to focus on FHHs in developing a methodology to improve data availability to quantify the SDG indicator 1.2.1 (Proportion of population living below the national poverty line) in relation to women's economic empowerment. In Sri Lanka, indicators related to poverty and labor force are mostly gained from household surveys, the HIES and the Labour Force Survey (LFS). However, surveys usually fail to provide information for all necessary levels of disaggregation. Here we demonstrate how to enhance the usability of LFS data for analyzing selected women empowerment issues through data integration. By combining LFS database with income

variable modeled through HIES we also present the issues on data integration using two independent data sources.

Keywords*: data integration, Household Income and Expenditure Survey, Labour Force Survey, women economic empowerment, poverty-related indicators.*

The views expressed here are those of the author and should not be considered as reflecting the views or carrying the endorsement of the United Nations

Contents

1. Preamble

Historically Sri Lanka is considered an outlier among other developing countries.[1] Many social indicators led by HDI and its component indicators are well above the standards of other developing countries (HDI rank 76, Life expectancy at birth 75.5 years, expected years of schooling 13.9 years and Gross National Income (GNI) per capita at 2011 PPP $ is 11326). However, progress is not in par with above on few other indicators such as gender equity in economic front, etc. For example, female population amounts to 51.9% of the total population in Sri Lanka (in 2016), of which only 36% are engaged in labour force.[2] On one hand, female graduates outnumber male graduates (53%) at the tertiary level in Sri Lanka, but constitute only 35.9% of the graduate labour force.[3] On the other hand, due to the protracted ethnic conflict, chances of women participating in labour force declined over the last three decades as they are confronted with problems relating to child bearing, child rearing, household chores and taking care of other family members. The most recent Household Income and Expenditure Survey (HIES) estimates that 1.4 million households (25.8 percent of households) in Sri Lanka are female-headed.[4] More than 50 percent of women who are head of household are widows, 4 percent have never married, or are divorced and rest are married but husbands live away from family[5]**.** Due to economic, social and cultural discrimination which limits women's economic participation, Female Headed Households (FHH) seem to be one of the most vulnerable groups and can be considered as a priority to women's empowerment in Sri Lanka.

Minister of Women and Child Affairs identified FHH as a priority policy issue related to women's empowerment in Sri Lanka at the 61st Session of the Commission on the status of women held in March 2017 at the UN.[6] With the technical support of the UNFPA, the Ministry of Women and Child Affairs in Sri Lanka developed a National Action Plan for female headed households in 2016. It will be effectively implemented in cooperation with a wide range of stakeholders.[7]

Thus, in this study, we decided to focus on FHHs in developing a methodology to improve data availability to quantify the SDG indicator 1.2.1 (Proportion of population living below the national poverty line) in relation to women's economic empowerment.

---

[1] W.D. Lakshman and Clement A. Tisdell (eds), *Sri Lanka's Development Since Independence: Socio-economic Perspectives and Analysis*, (New York: Nova Science Publishers, 2000), p. 9.
[2] (https://tradingeconomics.com/sri-lanka/population-female-percent-of-total-wb-data.html)
[3] (http://nation.lk/online/2017/09/23/the-overlooked-barrier-in-womens-economic-empowerment.html)
[4] *Household Income and Expenditure Survey 2016*, 2018, Colombo, Dept. of Census and Statistics, p. vi.
[5] *Household Income and Expenditure Survey 2016*, 2018, Colombo, Dept. of Census and Statistics, p. 80.
[6] https://www.un.int/srilanka/statements_speeches/national-statement-sixty-first-session-commission-status-women-csw61-0 (Accessed on the 23 Feb. 2018).
[7] file:///C:/Users/User/Downloads/20%20Sri%20Lanka%20CPD%20-%20LKA.9%20-%206Jun17.pdf (Accessed on 12 March 2018)

2. Study Objectives

i) a. National Issues/Priorities

In Sri Lanka, women are not subjected to extreme forms of oppression, but there are many adverse structures and forces, such as the patriarchal social structure, traditional values, rituals and myths, the division of labour, unequal pay and women's lack of participation in politics contribute to the continuing subordination of Sri Lankan women (Herath, 2015)

NGOs working with FHHs consider war widows, disabled women and elderly women as economically most vulnerable groups. NGOs and government both agree that war widows have distinct socio-economic needs and issues which need to be recognized and addressed.[8]

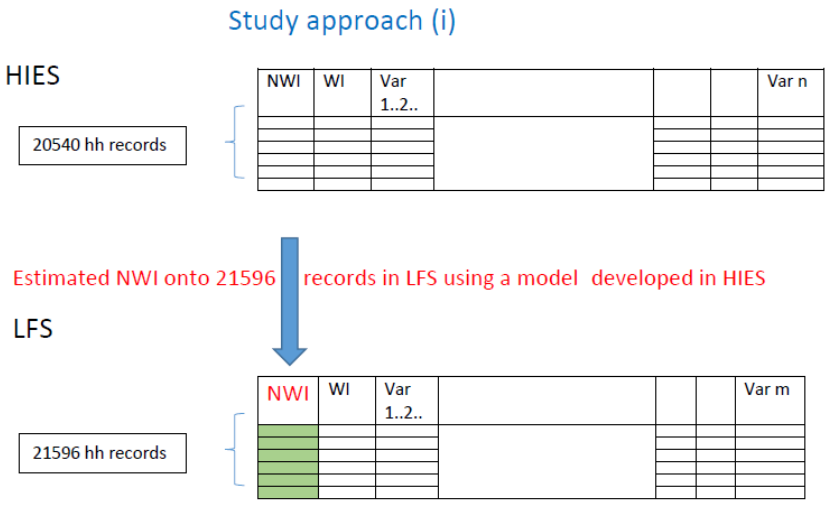
Beside the indicators of SDG5 which are directly related to women's empowerment, indicators of SDG-1 (Poverty) and SDG-10 (Inequality) when being disaggregated by sex can provide a tool to reflect women's economic empowerment in any society, if the data available at necessary disaggregated levels. Keeping in line with the ambition of SDGs – "leaving no one behind (LNOB)"- and under the light of national priorities, research team decided to focus on FHH to address poverty and inequality in relation to economic empowerment of women.

Given the low labour force participation of women in Sri Lanka and the general assumption of poverty being linked with FHH, research team sought a way to provide richer data to help better understanding of how FHH fall between the cracks of national poverty line. To do so, we selected HIES and Labour Force Survey (LFS) and tried to generate new statistical data that could be used to strengthen the national action plan of FHH.

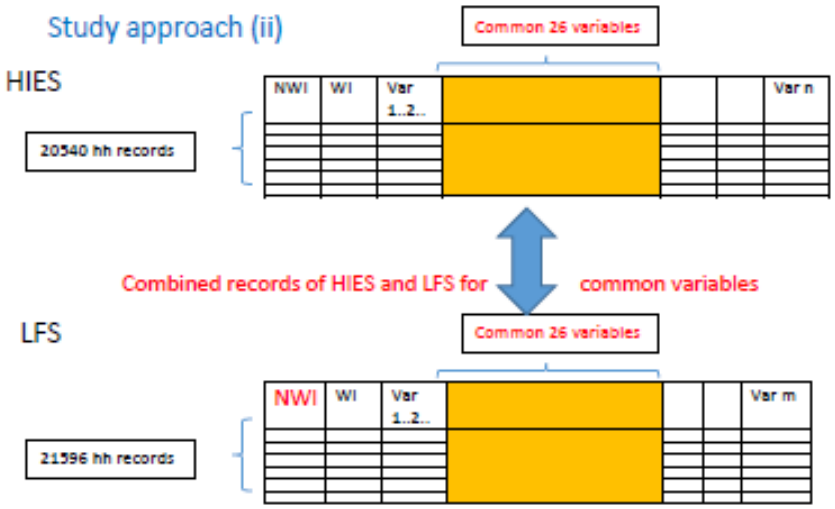
[8] *Mapping of Socio-Economic Support Services to Female Headed Households in the Northern Province of Sri Lanka*, United Nations, 2015, p. 37

b. Study Approach

i. In this study we demonstrate how to increase the potential of using LFS data by combining LFS database with income variable modeled through HIES, thereby enhancing the usability of LFS data for analyzing selected women empowerment issues as described in below section (ii) Research Question. (See diagram below)



ii. Subsequently, we combine the two datasets LFS (updated) and HIES using variables common to both surveys. We identified and/or generated 26 variables common to both surveys. Combined file contains only 26 variables. Nevertheless, sampled HHs in each survey (i.e. LFS and HIES) conducted in 2012-13 are combined to increase the number of total sampled observations. (See diagram below)



ii) Research Question

The main research question is that “can the data availability and its disaggregation be improved by combining two sources of information such as LFS and HIES?”. There are also specific questions that can be addressed when the data is available:

1. Can the LSF data updated with NWI (and Total HH Income) through proposed Study Approach (i) above be used to address certain issues related to FHH more efficiently?
   Example: Are poor female headed households (FHH) more vulnerable in the labour market than male headed households?
2. Can the combined data generated through proposed Study Approach (ii) above be used to disaggregate at least certain identified aspects of FHH more effectively?
   Example: Do heads of FHHs experience greater inequality than heads of Male Headed Households (MHHs) at District/Sector levels?
3. Can we identify the limitations and make recommendations to National Statistics Office (NSO) on improving existing data sources for better utilization?

iii) Selected SDG Indicator to be Focused

We intend to improve data availability in order to quantify the SDG indicator 1.2.1 (Proportion of population living below the national poverty line) by sex and age in relation to women’s economic empowerment. This will help identify nested disaggregated data needs as well.

iv) Desired disaggregation levels of characteristics of the head of the household for Exploratory Data Analysis

- Sex of the Head of the household
- Poor/Non-poor status
- Education level
- Age
- Disability status
- District
- Sector
- Selected employment characteristics

3. Combining National Data Sources to Improve Data Availability

Focus of our study is on providing data for Goal 1-Poverty indicators while considering women economic empowerment issues through addressing the FHHs. For measuring poverty, the main indicator used is the household based per capita income which is available through HIES conducted by the Department of Census and Statistics (DCS). However, the information base to analyze the female empowerment using Labour Force characteristics available in HIES is somewhat weaker compared to the ones in LFS.

The LFS is a nationally representative quarterly survey of the size around 26,000 households at each quarter, which is conducted by the DCS and covers labour force characteristics in a comprehensive scale in addition to basic demographic details. Characteristics on economic activity status include, among other things: individual occupation, wages, hours of work, industry, absence from work, unemployment and underemployment, job permanency, full time/part time status, primary and secondary employment details, reasons and methods for seeking another job, and unemployment benefits. This information seems very useful in exploring poverty effects on FHHs in Sri Lanka. However, the LFS is not the ideal source to get total household income, the key variable in analyzing poverty. The Labour force survey collects the information of monetary income and in-kind receipts from the monthly and daily wage/salary earners for paid employees and also the information on gross monthly income of employers and own account workers (gains from their enterprises). As such, it only collects information on employment/wage income (WI). In HIES, however, in addition to employment income, the information on non-wage income (NWI) which is other cash receipts and by chance/ad hoc gains is also collected.[9] (See Annex III for details). In both cases, wage income (WI) from employment is collected while NWI is missing in LFS.

HIES, which is conducted every three years, surveys a sample of 25,000 housing units throughout the country to facilitate the information be given at district level. Data is collected at the field in twelve consecutive monthly rounds to capture seasonal variations in income, expenditure and consumption of households. The HIES gathers information related to demographic characteristics of the household members, household expenditure on food and non-food items, and household income (received by each household member from all the different sources). The sampling frame, which is the collection of all the census blocks prepared in Census of Population and Housing (CPH) 2011 in Sri Lanka, is used for selection of PSU[10]s at the first stage of sample selection. The PSU selection is done within all the independent selection domains (25 Districts and 3 Sectors: Urban, Rural and Estate) that are assigned different sample size allocations to total the targeted sample size of 2,500 PSUs. The method of selection of the PSUs is systematic with a selection probability proportionate to the number of housing units available in each census block within the selection domains (PPS).

As mentioned earlier Total Household Income is comprehensively reported in HIES. It covers the income a household receives as WI and NWI, described earlier. According to the HIES report 2016, the wage income

[9] See Annex II for overview of HIES and LFS
[10] Primary Sampling Unit

amounts to 78% of total household income. The balance 22% is the NWI. Therefore, if the income source of LFS is used for this analysis, on average, 22% of the income is under estimated thus, the extent of poverty would be overstated.

As such, it requires the use of statistical techniques either to impute NWI data in the LFS or to match the HH wage income obtained from LFS with that of HIES and adjust LFS income to reflect the Total HH income to effectively estimate the income data. However, the latter option requires both surveys to be identical, but it is not always/usually true. In Sri Lanka, it occurred only in 2016, when, the sample selected for HIES was used for the LFS up to household level – meaning the sample household units were exactly the same. However, in 2012/13 surveys, sample household units were substantially different. Therefore, we opted to estimate the NWI in LFS using a statistical model developed using HIES data 2012/13.

Section 5 below describes the modelling of NWI using the HIES data which will be used to impute the NWI into LFS households.

4. Limitations of Data from the Selected Household Surveys

Data disaggregated not only by sex, but by age, income, province, sector, and other socioeconomic and demographic characteristics are also important for monitoring the SDG Goal 1.

The foremost characteristics of any data are its timeliness of availability, validity, accuracy, reliability, frequency and comprehensiveness. The frequency of data generation/production plays the key role in integrating different data sources. As indicated, Sri Lanka LFS is a quarterly survey while HIES collects data once every three years. As such, the technique used for this study cannot be repeated annually but once in every three years.

5. Statistical Techniques Used to Combine Data from Different Sources

i) Literature Review on Estimation using Regression Models

$\ln(y_i) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \epsilon_i$ is a well-known multiple regression model for income where $\ln(y_i)$ is the logarithm of income, y_i, x's are predictors and β's are the model parameters. This model is widely used in national statistical agencies such as Statistics Canada, Wendt (2007), US Department of Labour, Eltinge et al (2003), Fisher (2006), Montaquila et al (2014) and Australian Bureau of Statistics Starick (2005), as an imputation technique to compensate for missing data. Kalton (1982) has suggested that regression imputation can be used even if there is more than 10% non-response. In that respect, application of this technique to estimate 100% data for a variable is a new application to Sri Lanka. As such, we cannot cite any previous work related to 100% of imputation. However, there are several studies available for combining survey data from independent samples as a cost-efficient method to secure better statistics (Wendt, (2007), Dorfman, (2008), Qi Dong et al (2014)

Kim and Rao (2012) proposed a model-assisted projection method of estimation based on a working model to combine two surveys. They generated synthetic or proxy values of a variable of interest by first fitting the

working model relating the variable of interest to the auxiliary variables from an independent survey and then predicting the variable of interest using the auxiliary variables observed in the survey of interest. Our exercise also to a great extent resembles what Kim and Rao (2012) suggested.

ii) Methodology

We initially focus on estimating the SDG indicator 1.2.1 (Proportion of population living below the national poverty line) by sex and age and eventually identifying nested disaggregated data availability/needs to study women's economic empowerment.

We checked different regression models for NWI based on number of selected 22 variables (such as age, sex of head of household, education index, income from employment, main industry (agriculture or non-agriculture), sector, ethnicity and religion, etc.) from the list provided in the Annex I excluding identification and weight variables using the information from the HIES database. Then this model was used to estimate total household income in LFS sample HHs. The estimated NWI in LFS can be combined with existing LFS WI data to get the total HH income in LFS which can be used to measure poverty and women's economic empowerment. In this methodology, we used predictor variables common to both surveys as mentioned above. A brief description of the regression procedure is given below.

We define following variables:

Y2 = Non-wage-income (NWI) in a given household in HIES.
Y1 = Wage income in a given household in LFS.

After conducting preliminary regression analysis, we found that $\sqrt{Y_2}$ is a better representation of the income variable than $\ln(Y_2)$ based on the correlation analysis between actual and predicted values. (See Figure 1 in section 6.c)

Hence, we propose the following regression model to estimate NWI in LFS.

$$\sqrt{Y_2} = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_{22} + \epsilon_i$$

where x's are the common variables which are listed in the Annex I.

Initially, we used all 22 predictor variables (after excluding X_3-X_5 which are identification variables and X_{26}, which was used as a weight not as a predictor) in the regression using HIES 2016 data. We then selected the best predictor model based on adjusted R^2 value. The model contains six predictors which are highly significant, and this model was used to estimate NWI in LFS. Let this estimate be $\widehat{Y_2}$. Then the total income

in each household in LFS is estimated by Y= Y1+$\widehat{Y_2}$. Using 2016 data, we evaluated the model validity and efficiency. This was possible since ***both LFS and HIES used the same sample of households in 2016.*** Having the experience of above procedure, we developed a similar model using HIES 2012/13 to estimate NWI in LFS 2012/13. We then applied this model to estimate NWI in LFS 2012/13 survey to obtain the complete survey database which includes the total HH income to be used to assess poverty. As a secondary output, we combined the LFS 2012/13 and HIES 2012/13, to increase the total sample size encompassing common variables as mentioned in the section on 'Study Approach', to increase data availability for analyzing women's economic empowerment factors at disaggregated levels. Since the samples were different in 2012/13 for HIES and LFS, we could combine the samples to increase the size, subject to common variables listed in the Annex I.

iii) Steps Followed

a. Estimated NWI in HIES 2016 for HHs using the above model. We used 22 variables from the list of 26 common variables (X_1-X_{26}), including direct and generated ones as listed in the Annex I for analysis (many other variables may be generated as needed using original survey variables).[11] However, adjusted R^2 values were not promising to estimate the NWI. Therefore, we limited our exercise only on people under the National Poverty Line (NLP).

b. We fixed the poverty level at LKR 4,166.00 per person per month (NPL) as set by the DCS for 2016 and selected HHs with total per capita income less than NPL from HIES 2016 for estimating the model. Also, to get estimates with better precision, we dropped the observations where the per capita income falls below zero (This was due to financial losses incurred by HHs on agricultural and non-agricultural self-employment activities) to model the NWI component of the households. (Initially 3958 observations were used in the analysis but only 2021 observations were retained due to missing or partially missing values and negative income values)

Note - *For regression models, large variability present in the NWI seems less suitable to produce estimates with high precision.*

c. Estimated NWI in LFS 2016 for HHs with less than Rs.4166 'wage income' using the above model (this is to ensure that all poor HHs are covered, including zero NWI).[12]

[11] See Annex I for variables list.
[12] To comply with the SDG 1.2.1 indicator which was selected for the analysis, income cut off point of national poverty line was used to develop the model. If all income levels are considered for the model, the model fit gets much weaker.

Note: It may be appropriate to estimate different models at different levels of wage income. However, in this study we have estimated only a model for people under poverty line according to NPL. (In accordance with definition used in SDG 1.2.1)

d. Evaluated the model efficiency comparing the 'estimated' and 'actual' values of NWI data in the HIES 2016 (using correlation analysis). After calculating the total income using HIES 2016 data, the responses were matched against the actual values as available in the HIES 2016. (Both survey samples were identical in 2016).

e. Repeated similar exercise for 2012/13 surveys (HIES and LFS). To identify poor HHs in 2012-13, per capita total income of Rs.3624 per month was used (NPL for 2012/13 as set by the DCS for 2012/13).[13]

f. Estimated the total income of LFS HHs by adding NWI to wage income.

g. Updated LFS database with the new total income variable which enables identification of poverty to analyze LFS data with the focus of FHH as part of women's economic empowerment.

h. Combined two data sets, HIES 2012/13 and LFS 2012/13 for HHs below poverty line using the 26 common variables for supplementary analysis of data.

i. Thus, demonstrated that we have increased the sample size for disaggregation by district, sex, etc. at the level of poor HHs. In table 1, for each survey, number of poor sample households by district, and sex of the household head is presented.

[13] This poverty line is based on expenditure, which is accepted proxy for income especially for lower income HHs.

Table 1 – Number of poor sample households by district and sex of household head in HEIS and LFS

Sample Tabulation of poor FHH counts

District by Survey & Gender of HH head

	HIES		LFS	
DISTRICT	Male	Female	Male	Female
11	31	13	317	47
12	34	16	366	39
13	52	17	209	22
21	23	16	362	69
22	17	9	151	22
23	18	7	94	15
31	36	16	283	39
32	45	20	247	31
33	15	11	358	41
41	18	10	162	21
42	6	2	123	4
43	7	4	102	15
44	5	4	131	6
45	23	2	119	17
51	17	8	161	45
52	21	8	208	30
53	14	2	129	19
61	41	15	597	97
62	14	6	169	33
71	13	4	254	60
72	12	1	177	26
81	31	8	284	38
82	7	7	186	20
91	54	18	316	36
92	38	16	250	33

Determination of the new weights (inflation factors) of the combined sample of PSUs integrating the two samples were done based on the assumption that both samples were drawn using Probability Proportional to Size (PPS) method at a single selection. Coinciding months from both LFS 2012 and 2013 surveys were selected to maintain the reference period of the HIES survey, which is, July 2012 to June 2013.

Distribution of PSUs by District and Sector were obtained for each survey and recalculation of weights was done. Within each Sector and District, the adjusted weight for combined survey is derived as:

$$Ajd.Weight = \frac{Weight(survey_i) \times (\#\ of PSUs\ \text{in}\ survey_i)}{\#\ of PSUs\ \text{in}\ combined\ data\ file}$$

where *i* = 1 (HIES), 2 (LFS 2012) and 3 (LFS 2013)

Sample PSUs for both surveys are drawn for the whole year and allocated for different months and quarters. Therefore, weights calculated refers to annual sample.

6. Results and Discussion

An analysis of results obtained from the regressions and a comparative study given in the methodology section is produced in this section. Following steps were taken:

a. Identification and generation of common variables:

   As indicated earlier, 26 common variables were identified in both LFS and HIES (except disability which will be available from 2018). Relevant data files containing 26 variables were extracted for LFS 2016, HIES 2016, HIES 2012/13 and LFS 2012/13[14] .

   There were 2540 poor HH records for model building.

b. Validating the model, following the steps below:

   i. The predictor variables found to be significant are listed in Table 2.

   ii. Regression was forced through the origin. (Intercept is made equal to zero).

   iii. Sampling weights were considered when fitting the model.

   iv. The model was fitted using 2021 observations (excluding records with partially missing variables).

Table 2 - Model Parameters, their standard errors and p-values for 2016 model

Variable	Estimate	Std. Error	t value	Pr(>|t|)
Sex HH Head	5.83	1.29	4.50	0.000
Other No	8.14	0.38	21.42	0.000
Is_Active_HH_Head	12.06	1.12	9.88	0.000
Is_Active median	9.24	1.14	8.08	0.000
AGRI INCOME HH	-0.0012	0.0001	-11.58	0.000
EMP INCOME HH	-0.0018	0.0001	-17.11	0.000

Mean square error is 404.3. Adjusted R^2 is 84.78% and F value is 1733 with 6 and 1903 degrees of freedom with a p-value of 0.0000. This model is highly significant and suitable for predicting NWI.

[14] These data files are submitted to ESCAP as attachments.

The model is:

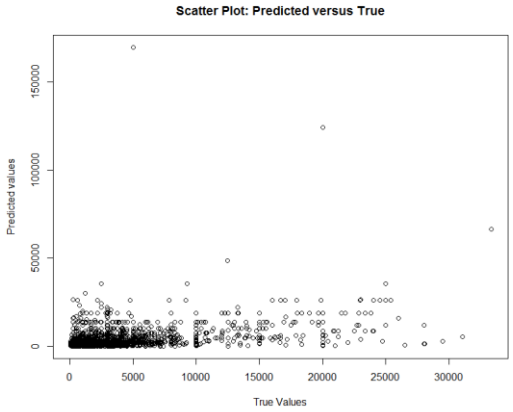
$$\sqrt{Y_2} = 5.83(Sex\ HH\ Head) + 8.14(Other\ No) + 12.06(Is - Active - HH - Head) + 9.24(Is - Active - median) - 0.0012(AGRI\ INCOME\ HH) - 0.0018(EMP\ INCOME\ HH)$$

The model developed refers to the whole country. Although it was expected to develop models for 25 districts due to limitations of observations available in the sample district level estimations were not feasible.

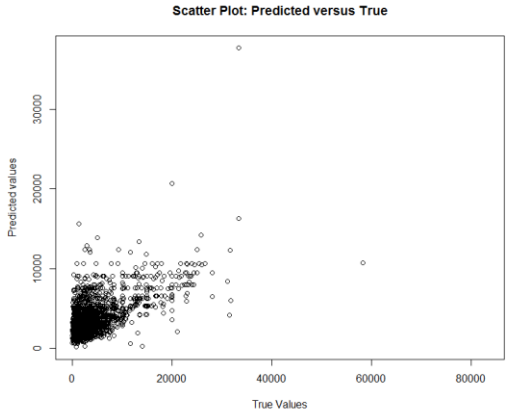
c. Estimation of NWI (Y_2) for LFS 2016 and checking efficiency of the model:

After calculating the total income using HIES 2016 data, the responses were matched against the actual values as available in the HIES 2016. (Both survey samples were identical in 2016). Following chart gives the correlation between the actual values and the predicted values using the model under both transformations.

Figure 1 - Plots of predicted and true values of NWI



$\ln(Y_2)$ Transformation



$\sqrt{Y_2}$ Transformation

Pearson product moment correlation coefficient between predicted and true values of $\sqrt{Y_2}$ is 0.58 and a 95% confidence interval for the correlation coefficient is [0.55, 0.61]. As mentioned in the methodology section the corresponding correlation value using $\ln(Y_2)$ as the dependent variable was only 0.32. Thus, $\sqrt{Y_2}$ transformation was preferred over $\ln(Y_2)$ transformation due to higher acceptance level of the model through correlation.

A new model was developed using HIES 2012/13 data based on the experience gained from the estimation of the model and analysis of HIES 2016 data which gave idea about the variables that could contribute in predicting Y_2. Here we used the same six predictor variables given in Table 2. Estimated model is:

$$\sqrt{Y_2} = 4.87(Sex\ HH\ Head) + 4.89(Other\ No) + 9.76(Is - Active - HH - Head) + 9.46(Is - Active - median) - 0.0079(AGRI\ INCOME\ HH) - 0.0013(EMP\ INCOME\ HH)$$

Summary of results of this regression is given in Table 3.

These are the results based on 2012/13 data; we observed there is a little difference between 2013 and 2016 models thus confirming consistency of estimations.

Table 3 - Model Parameters, their standard errors and p-values for 2012/13 model

Variable	Estimate	Std. Error	t value	Pr(>|t|)
Sex HH Head	4.87	1.21	4.03	0.000
Other No	4.89	3.41	14.36	0.000
Is_Active_HH_Head	9.76	1.08	9.06	0.000
Is_Active median	9.47	9.23	10.26	0.000
AGRI INCOME HH	-0.0008	0.0001	-5.89	0.000
EMP INCOME HH	-0.0013	0.0009	-13.06	0.000

Mean square error is 345.7. Adjusted R^2 is 76.87% and F value is 1462 with 6 and 2632 degrees of freedom with a p-value of 0.0000.

The predictions were obtained for the households in HIES 2013 using the model. Pearson product moment correlation coefficient between Predicted values and True values of NWI for poor HHs is 0.45. A 95% confidence interval for the correlation coefficient is [0.42, 0.48]. As there are 20 other

variables which might influence the NWI, though not significant in this model, the achieved correlation can be considered adequate.

d. Estimation of Y_2 for LFS 2012-13:

Using the above model (Table 3) we then obtained the predicted values of NWI for LFS 2012/13. The variable name is **Predicted Value** which is given in the combined data file.

e. Updating LFS data using NWI for poverty in HHs:

The estimated Y_2 value was used to estimate the Total Income as:

$$Y = Y_1 + Y_2$$

and thus, estimated variable Y was added to LFS database enabling the assessment of indicator 1.2.1 under SDG- 1. This database enables to analyze SDG 1.2.1 in relation to the research questions as set in the Section 2-ii. (This is a demonstration of a value addition for LFS data).

f. Combining two data files for poor households using common variables X_1-X_{26}:

The HIES 2012/3 survey period was 1 July 2012 – 30 June 2013. The relevant monthly data were extracted (surveyed PSUs during the reference period) from LFS-2012 and LFS-2013 to formulate a combined data file to coincide with the above HIES survey period.[15] Suitable Inflation Factors (weights) were calculated for the combined data file. Since the NWI was estimated only for poor HHs, at this stage we could only combine the data for poor HHs. As this exercise is a demonstration of the technique for data improvement with regard to poverty indicators, it is not intended to expand this for non-poor HHs which would require more advanced techniques for estimating NWI due to its huge variation/range of values.

Now, the data is available and for research purposes, poor FHH can be disaggregated by district, employment status and Industry, income, education, HH size and sector (urban/rural/estate).

7. Conclusions and Recommendations

a. Main Outcome

Through this approach, we have estimated an additional variable (NWI) in the LFS to compute Total HH Income and it enabled the identification of poverty within the context of labour force characteristics collected in the LFS (2012-13).

[15] The combined data set for poor HHs is provided to ESCAP.

b. Secondary Outcome

By combining HIES 12/13 with LFS 12/13 using 26 common variables, we increased the sample size by integrating the two databases for poor HHs, providing supplementary space for women's economic empowerment issues.

However, in 2016 sample of PSUs and the households are exactly the same in LFS and HIES and therefore, improving the count of observation was impossible. However, we have used 2016 data to estimate the model and to evaluate the efficiency of the model and predictions as mentioned in the introduction.

c. Comment on Survey Data

In LFS and HIES, there are several other proxy variables, but they cannot be utilized to combine these two data sets because of the difference in structuring of questions. We, therefore, recommend that DCS (The National Statistics Office of Sri Lanka) maintains coherence among surveys in terms of definitions of variables and consistent designs in sample surveys enabling combining data across different surveys.

Future studies may be encouraged to improve the techniques demonstrated in this study to capitalize the availability of many different data sources to increase disaggregated data (by sex, age district and sector) for monitoring SDG's and other development indicators.

8. Study Limitations, Potential Risks and Challenges

In this study we have modelled only the poor HHs according to NPL. Due to weak correlation between dependent variable and predictors at the non-poor HHs modelling for non-poor HHs was not straightforward as in the case of poor HHs where the sample size is relatively small. Nevertheless, as the main focus of the study was on SDG 1.2.1 which is related to HHs within NPL, the objective of shedding more light on the target group was achieved.

Sri Lanka LFS (quarterly year basis) and HIES are separate surveys conducted by the Department of Census and Statistics. They are designed to provide a range of information and data for specific purposes. We believe that it is extremely difficult to link these survey data for the purpose of disaggregation for gender equality and equity and economic empowerment of women at various sub-group levels. Although there are number of common linking variables available in both data sets, the samples in the two surveys are completely independent and hence the chance of linking the two data sets will be futile. Also, the comparability of data in terms of definition, coverage, reference period and frequency, etc. is questionable. Few other challenges are also present, for example the accuracy of data and small cell counts for disaggregated groups and possible weak significance levels in model formulations due to possible effects of multi co-linearity among variables.

Additional surveys to gather data increase response burden. As such effective use of administrative data produced by various statutory agencies and ministries with required modifications could lessen the burden of the need for new surveys. Another approach to minimize the burden on respondents and cost of doing large-scale surveys is to maintain the uniformity of questions and response categories across surveys (e.g. LFS and HIES) enabling the integration of such information to analyze measures poverty, equality and economic empowerment of women, etc. in relation to many other socio-economic indicators of interest. Another risk is the lack of willingness of decision makers to use estimated data through sample surveys in place of total enumerations. It is the practice of all stakeholders to gather data whenever data become necessary to make even their day to day decisions. The reason is that sample survey estimates are seen as incomplete by administrative professionals who do not have sufficient statistical literacy to appreciate statistics and to make decisions based on estimated data with sufficient accuracy. Reluctance by such groups to use the data generated through more sophisticated methods would pose a risk of making such processes popular.

References

Dong, Q., Elliot M.R., et al. (2014), Combining Information from Multiple Complex Surveys, *Survey Methodology*.

Dorfman, A.H. (2008), The Two Sample Problem, Bureau of Labour Studies.

Drew, J.D., M. P. Singh and G. H. Choudry (1982). Evaluation of small area estimation techniques for the Canadian Labour Force Survey, Proceedings of American statistical Association.

Durant, G.B. (2005). Imputation Techniques to handling item-nonresponse in the social sciences: A methodical Review, national Centre for Research Methods Working Paper Series, ESRC National Centre for Research Methods and Southampton Statistical Sciences Research Institute (SRI), University of Southampton.

Eltinge, J.L., R. A. Koslow and D. M. Luery (2003). Imputation in Three Federal Statistical Agencies, US Department of Labour.

Fisher, J.D. (2006). Income imputation and the analysis of consumer expenditure data, Monthly Labour Review.

Foster, J, J. Greer and E. Thorbecke (1984). A class of decomposable Povery measures, Econometrica, 52, 761-766.

Herath, H. M. W. A., L. H. P. Guneratne and Nimal Sanderatne (2015), Impact of microfinance on women's empowerment: a case study on two microfinance institutions in Sri Lanka, Sri Lanka Journal of Social Sciences, 38 (1): 51 – 61

Jae Kwang Kim, J. N. K. Rao; Combining data from two independent surveys: a model-assisted approach, *Biometrika*, Volume 99, Issue 1, 1 March 2012, Pages 85–100, https://doi.org/10.1093/biomet/asr063

Kalton, G. and Kasperzy (1982). Imputing for missing survey responses, Istat.

Lombardini, S., K. Bowman and R. Garwood (2017). A 'How To' Guide to measuring Women's Empowerment, Oxfam.

Molina, I. and J. N. K. Rao (2009). Small area estimation of poverty indicators, Working Paper 09-15, Statistics and Econometrics Series 05, Universidad Carlos III De Madrid.

Montaquila, J.M. and C. H. Ponikowaski (2014), An evaluation of alternative imputation methods, Westat and Bureau of Labour Statistics.

Rao, J.N.K., (2003). Small Area Estimation, Wiley.

Starick, R. Income Imputation in the Household, Income and Labour Dynamics in Australia (HILDA) Survey, Australian Bureau of Statistics. (Unknown year)

Wendt, M. (2007), Considerations before Pooling Data from Two Different Cycles of the General Social Survey, Social and Aboriginal Statistics Division, Statistics Canada.

Annex I: Common Variable List Identified in both LFS and HIES Surveys

	Variable defined	Description
1.	District	Administrative unit
2.	Sector	Urban/Rural/Estate sector
3.	PSU (Identification variable)	Sampled PSU number
4.	HU no (Identification variable)	Housing Unit number
5.	HH no (Identification variable)	Household number
6.	Age_HH_Head	Age of HH Head
7.	Sex_HH_Head	Gender of HH head
8.	Curr_Educ_HH_Head	Current Education of HH head
9.	Education_HH_head	Education level of HH Head
10.	Ethnicity_HH_Head	Ethnicity of HH head
11.	Religion_HH_Head	Religion of HH head
12.	Other_No	No of members (excluding HH head)
13.	Is_Active_HH_Head	Economically active or not HH Head
14.	Main_Occupation_HH_Head	Occupation of the HH
15.	Industry_HH_Head	Industry of HH Head
16.	Employment_Sta_HH_Head	Employment Status of the HH Head
17.	Monthly_Employment_HH_Head	Monthly Employment Income of HH Head
18.	Monthly_Agri_IN_HH_Head	Monthly Agri Income of HH Head
19.	Monthly_NonAgri_Income_ HH_Head	Monthly non Agri Income of the HH Head
20.	Age_mean_Others	Mean age of others
21.	Education_median	Median Education of Others
22.	Is_Active_median	Median Active or Not
23.	EMP_INCOME_HH	EMPLOYMENT INCOME HH
24.	AGRI_INCOME_HH	AGRI INCOME HH
25.	NONAGRI_INCOME_HH	NONAGRI INCOME HH
26.	Final weight	Weight

1. It needs to be mentioned here that items 3-5 above are Identification variables which could be used for possible comparisons at PSU or HH levels
2. The median values of few variables were also generated to reflect the average level of the variable value at household level. Example: Education median (median of the values of individual members' education levels of the same HH. The education level is an ordinal variable).
3. Another example is item 22 (Is active median). This is a median value of a binary variable (economically active=1; inactive=2) reflecting the majority status.
4. Weight Final is an inflation factor to be used when weighted results are desired.

Annex –II: Survey Briefs

A brief of LFS:

Department of Census and Statistics (DCS) designed a Labour Force Survey (LFS) on a quarterly basis to measure the levels and trends of employment, unemployment and labour force in Sri Lanka on a continuous basis. Two stage stratified sampling procedure is adopted to select a sample of 25,750 housing units to be enumerated at the survey. The sampling frame prepared for 2012 Census of Population and Housing is used as the sampling frame for the sample selection of LFS in 2016. Primary sampling units are the census blocks prepared at the Census of Population and Housing - 2012. In 2016, 2575 Primary sampling Units (PSU"s) were allocated to each district and to each sector (Urban, Rural and Estate) by using the Neymann allocation method which considers the variance of unemployment rate as usually. The allocated sample for each district then equally distributed for 12 months. Secondary Sampling Units are the housing units in the selected 2575 primary sampling units (census blocks). From each selected primary sampling unit, 10 housing units (SSU) are selected for the survey using systematic random sampling method.

A brief of HIES:

The Department of Census and Statistics (DCS) conducts the Household Income and Expenditure Survey (HIES) since 1990/91 and continued once in every five years until 2006/07.Thereafter once in every three years starting from 2009/10 due to rapidly changing economic conditions demanded far more frequent monitoring of the household income and expenditure patterns in the country. The latest survey was in 2016 covered all 25 districts in the country. HIES provides the most important socio-economic indicators for development and evaluation of the socio economic development policies and plans and finalization of SDG development goals

Generally, the HIES surveys a sample of 25,000 housing units throughout the country to facilitate the information be given at district level. Data is collected at the field in twelve consecutive monthly rounds to capture seasonal variations in income, expenditure and consumption of households.

The HIES gathers information related to demographic characteristics of the members of the surveyed households, expenditure on food and non-food items and income received by each household member from all the different sources in a compulsory manner.

The sampling frame, which is the collection of all the census blocks prepared in CPH 2011 in Sri Lanka, is used for the selection of the PSUs at the first stage of the selection.

The PSU selection is done within all the independent- selection domains that are assigned different sample size allocations to total the targeted sample size of 2,500 PSUs. The method of selection of the PSUs at the first stage is systematic with a selection probability given to each census block proportionate to the number of housing units available in the census blocks within the selection domains (PPS).

Annex III: Information on WI and NWI

Average monthly HH income by main source of income –HIES 2016 and 2013

Source of income	2016		2012/13	
	Mean (Rs.)	Percentage Share of income (%)	Mean (Rs.)	Percentage Share of income (%)
Sri Lanka	**62,237**	**100.0**	**45,878**	**100.0**
Monetary Income	**52,979**	**85.1**	**39,300**	**85.7**
Wages/Salaries	23,790	38.2	16,134	35.2
Agricultural activities	4,753	7.6	5,213	11.4
Nonagricultural activities	10,813	17.4	7,990	17.4
Other cash income	8,029	12.9	5,230	11.4
Income by chance/adhoc gains	5,594	9.0	4,733	10.3
Non-monetary Income	**9,257**	**14.9**	**6,578**	**14.3**
Income in- kind	2,964	4.8	2,381	5.2
Estimated rent value of own occupied housing unit	6,293	10.1	4,197	9.1

Other cash Income categories

Pension Payment	Disability payments	Samurdi	Elderly payments	Tuberculosis/kidney diseases payment	Educational & Scholarships	School food program	Triposha food program	Dividends /Interests	Rent from properties/boarding fees	Other Income	Current remittance & transfers	
											Income from foreign country	Income from the country

Income by chance /ad-hoc gains

Loans taken from banks / money lenders, etc (*including credit cards).	Sale/ pawning of assests (Land, house, jewellary)	Withdrawals from savings, bank deposits, Grativity, Provident fund	Income receives from assosiations/ welfare societies for births, deaths, marriages /, etc	Seettu / settlement of loans	Health and medical aids	Compensation / Insurance etc	Other (lottery & other adhoc gains)	Foods and other commendations	Disaster/other relief payments