DLSU RESEARCH CONGRESS 2020
"Building Resilient, Innovative, and Sustainable Societies"
June 17-19, 2020

# A Procedure for the Generation of Small Area Estimates of Philippine Poverty Incidence: The Case of Region 8 (Eastern Visayas)

Nelda Atibagos Nacion[1,2,*], Frumencio F. Co[1], and Arturo Y. Pacificador, Jr.[3]

[1]*Mathematics and Statistics Department, De La Salle University*
[2]*Mathematics and Statistics Department, De La Salle University-Dasmariñas*
[3]*Institute of Statistics, University of the Philippines Los Baños*
**Corresponding Author: nanacion09@gmail.com*

**Abstract:** This study proposes an alternative procedure in generating small area estimates of poverty incidence using imputation-like procedures coupled with a calibration of estimates to ensure coherence in the regional estimates. Specifically, this study applied Deterministic Regression Approach, Stochastic Imputation-like procedure similar to Stochastic Regression, and applied the calibration techniques to ensure that the small area estimates conform to the known regional estimates. This study used the Family Income and Expenditure Survey of 2009 and the Census of Population and Housing (CPH form 2) 2010 to come up with reliable estimates of poverty incidence by municipal level. Since the CPH is conducted in the Philippines every 10 years, the CPH 2010 is the latest data that was used. Small area estimates of poverty in the Eastern Visayas (Region 8) at the municipal level were produced by combining survey data with auxiliary data derived from census. Region 8 was chosen because the poverty in this region is quite alarming. Nearly 4 out of 10 persons in this region, or 3 in every 10 families are poor. The study fitted different models and by comparing the two methods of imputation, it was found that the Stochastic Regression Imputation performed better than Deterministic Regression in attaching the income in census. The error used in Stochastic Regression was estimated using the non-parametric Kernel Density Estimation method.

**Key Words:** poverty; small area estimation; imputation; calibration

## 1. INTRODUCTION

Poverty, as defined by Encyclopedia Britannica, is the state of one who lacks a usual or socially acceptable amount of money or material possessions. Looking at the bigger picture, poverty is a general indicator of how well a country is moving to ensure economic growth that would provide the citizen a better quality of life (Global Policy Forum, 2010). Poverty has always been a challenge in many countries in the world. In fact, according to the 2015 World Bank estimates, 9.6% (700 million) of the world's population are living on less than $1.90 a day and are considered poor. Because of this, poverty alleviation has always been a part of each country's development programs. The first Millennium Development Goal (MDG 1) is to eradicate poverty and hunger with three targets: (i) to halve the proportion of people whose daily income is less than $1.25 a day, (ii) to achieve full and productive employment, and (iii) to halve the proportion of individuals suffering from hunger between 1990 and 2015. Thus, a study about poverty is deemed important.

The programs of the government should be in response to the issues mentioned by the MDGs. The local government should be involved in the different programs formulated by the national government to be able to reduce poverty. For this to be possible in the smaller level or domains, reliable data is important. In the Philippines, provinces and municipalities are considered small area levels. To be able to devise a plan, there should be a picture of the poverty condition of these small areas. And to be able to get a picture, a reliable estimate is necessary. Hence, the importance of small area statistics.

There are many small area estimation (SAE) techniques available. So far, the official methodology done in the Philippines to generate small area statistics is the one conducted by the National Statistical Coordination Board (NSCB) in their 2005 paper, "Local Estimation of Poverty in the Philippines" with a modification in attaching the income. The said paper estimated the poverty and expenditure in the provincial and municipal level. The procedure basically utilized the Elber's, Lanjouw and Lanjouw (ELL) Methodology, which utilized both the

Census of Population and Housing (CPH) and the Family Income and Expenditure Survey (FIES). Regression models were utilized from the FIES for the purpose of predicting income based on variables common to both FIES and CPH. Once the model has been identified, it is used to predict the income values which is attached to CPH from where a household is identified as poor or non-poor. And thereby allows estimate of the poverty incidence. However, based on the results, the municipal level estimates do not conform to the regional estimates.

In this regard, a similar approach was done by Pacificador et al. (1996) in the study "Attaching the Income and Expenditure Dimension to the 1990 Census of Population and Housing (CPH)," which is in response to the call for a more in-depth analysis of the 1990 Census of Population and Housing (CPH) data. This was an initial attempt in developing appropriate file merging technique also called Record Linkage. The income and expenditure variables were attached to the CPH data using Deterministic Regression. The methodologies done by the NSCB and Pacificador are like an imputation approach in coming up with attaching income which can be used to generate small area estimates of poverty incidence. Both methods employed Deterministic Regression approach in imputing data. However, the downside of using Deterministic Regression as a model in predicting income and expenditures is that it is the same as the class mean imputation, wherein the predicted values are the average values of the dependent variable and the fitted values will have grouping effects. Additionally, there are three problems that can be encountered in using this type of imputation: it reduces the variance of the imputed variables; it shrinks standard errors which invalidates most hypothesis tests and the calculation of the confidence interval; and it does not preserve the relationship between variables such as correlations. Thus, the model was not able to preserve the distribution of the error term.

The disadvantages of these two methodologies are: they will not replicate the distribution because of the grouping effect and there is no guarantee that the estimates will be coherent with the regional estimates of which direct estimates are available of adequate precision. In the proposed procedure of the researcher, calibration techniques were used to ensure that the estimates conform with the regional estimates. Building up on the weakness of these methodologies, this study proposed an alternative procedure in estimating the poverty incidence of the municipalities in the Philippines. The procedure is the same as the procedure used by Pacificador (1996) but borrowed strength from the imputation; the Deterministic and Stochastic Regression to address the weakness of using Deterministic Regression only. The imputation procedure is that of a unit level and not area level. In addition, the final estimates were calibrated so that they conform to the regional estimates of poverty incidence in the Philippines.

## 2. METHODOLOGY

### *2.1 Data*

This study utilized the 2009 Family Income and Expenditure Survey (FIES), which is conducted every three years in the Philippines by the Philippine Statistical Authority (PSA), formerly NSO. According to the 2005 World Bank report, in cooperation with the NSCB, the FIES contains information on household income, expenditure, and consumption, in addition to socio-demographic characteristics.

Along with the FIES 2009, the 2010 Census of Population and Housing (CPH) was also used in this study. The CPH provides data on which the government planners, policy makers, and administrators base their social and economic development plans and programs (2010 CPH). This full census is conducted every 10 years, with a Census of Population at 5-year intervals.

### *2.2 Statistical Models*

This study utilized two models: The Deterministic Regression and the Stochastic Regression to predict the income value to be attached to CPH. A comparison was done to ensure the reliability of estimates. The Stochastic Regression was used in order to address the weakness of the Deterministic Regression. Stochastic Regression is the same as that of the Deterministic Regression but with the addition of the error term. The estimation of error term is crucial in this study. The Kernel Density

Estimation (KDE) or histogram method was used in estimating error since the data is not normally distributed. KDE is a non-parametric way of estimating error. Non-parametric approaches are more appropriate if it is not possible to make strict assumptions about the form of the underlying density function. This method subdivides the domain into bins and counts the number of samples $n_b$ which fall into each bin. The local probability density is obtained by dividing the number of samples in each bin by the number of samples $N$ and the bin width $h$. It can be expressed as $\hat{f}(x) = \frac{n_b}{Nh}$, for $x_b \leq x < x_{b+1}$, where $x_b$ and $x_{b+1}$ are the extents of bin $b$, and $h = x_{b+1} - x_b$. The $\hat{f}$ is used to denote a density estimate of the probability density function $f$. This smoothed rendition connects the midpoints of the histogram, rather than forming the histogram as a step function, it gives more weight to the data that are closer to the point of evaluation.

In this study, different models were fitted in the region which were built using the first three FIES replicates and validated in the last replicate. A total of four model building sets and four model validation sets were used. This part is the modification made by the researcher from the methodology employed by NSCB to ensure the accuracy of the models using Deterministic Regression Imputation (DRI) and Stochastic Regression Imputation (SRI) techniques.

In this paper, the model used to attach the income is of the form $\hat{Y}^* = \mathrm{X}\hat{\beta} + \mathrm{e}$. Here, $\beta$ represents the regression coefficients giving the effect of the X's or auxiliary variables on Y (the total income of the household), and $\hat{e}$ is a random error term representing that part of the income that cannot be explained using the auxiliary information.

Some implementations of ELL methodology have fitted separate models for each stratum defined by the survey design. The advantage of this is that it tailors the model to account for the different characteristics of each stratum, but it might increase the problem of over-fitting if the strata is small.

Another way of validating the estimates is through the bootstrap method, introduced by Efron (1979), which is a very general resampling procedure for estimating the distributions of statistics based on independent observations. The bootstrap, which is shown to be successful in many situations, is accepted as an alternative to the asymptotic methods. In fact, it is better than some other asymptotic methods such as the traditional normal approximation and the Edgeworth expansion (NSCB, 2005). Bootstrap methodology was used to determine the error term ê to be considered in the model. A total of 1,000 independent samples were drawn, and the mean error was considered.

After predicting the income using the model or imputation method, another variable was created to compute for the per capita income of the households. The per capita income is the ratio of the income and the number of members in a household. This step is necessary to determine the number of poor households per municipality.

According to the PSA website, the annual per capita threshold in the Philippines for 2009 is Php16,871 at the national level. A family living below this value annually is considered poor. In this study, the provincial threshold was used in determining whether a family is poor or not. The total number of poor households per municipality was determined by collapsing the new created data set by municipalities. But before this was done, the variable municipality in the CPH was recoded in such a way that the municipal code is unique for each municipality.

In producing the final estimates, the poverty incidence was computed as $P_R^b = \frac{\sum_{n_{ij}} I(E_{ij}^b < z)}{\sum_{ij \epsilon R} n_{ij}}$, where $n_{ij}$is the size of household $ij$ in $R$ and $I(E_{ij}^b < z)$ is an indicator function (equal to 1, when the per capita income is below the poverty line/threshold and 0, otherwise).

After identifying if a household is poor or non-poor, the number of poor households per municipality was obtained and was simulated by using bootstrap methodology. In this paper, the simulated values for the number of poor was obtained by parametric bootstrap. The mean and standard error for the 1000 bootstrapped values served as the estimates for mean number of poor households and standard error for

each municipality. After producing the poverty estimates, the values were calibrated in order to conform to the regional estimates obtained from FIES.

The term calibration estimation was introduced by Deville and Sarndal (1992) as a procedure of minimizing a distance measure between initial weights and final weights subject to calibration equations. In this study, the bootstrapped total number of poor was calibrated using the formula

$$\hat{Y}_m^* = \hat{Y}_m * \left[ \frac{\hat{Y}_R}{\hat{Y}_R^*} \right].$$

where $\hat{Y}_m^*$ = calibrated municipal estimates

$\hat{Y}_m$ =municipal estimates from CPH

$\hat{Y}_R$ =regional estimates from FIES

$\hat{Y}_R^*$ =total municipal estimate per region from CPH

The rescaled value of poor is expected to correspond to the total number of poor households in the region. The regional estimates were used in the calibration since FIES was designed for regional level estimates.

## 3. RESULTS AND DISCUSSION

### *3.1 Common Variables*

After careful checking on the issues between FIES and CPH, the common variables were found. The common variables, denoted by X, are called the auxiliary variables. The common variables are: the type of building where the family reside, the construction materials of the roof, construction materials of the wall, the floor area, sex of household head, age of the household head, marital status of the household head, highest grade completed by the household head, province, municipality, and barangay. Among these 11 common variables, six were recoded to ensure that the variables were measured in the same way between the two data sets for the modeling purposes. The type of building where the family reside, the construction materials of the roof, construction materials of the wall, the floor area, marital status of the household head, and the highest grade completed by the household head were the six variables recoded and the remaining five variables were measured in the same way.

### *3.2 Association of Common Variables with Income*

After identifying the common variables, a first step regression was done in the original FIES data in order to determine the significant variables in predicting the income. In this case, Y is the total income (dependent variable) and the auxiliary variables are the X variables (independent). The result shows that all the variables were found to be significant (generally) except some of the dummy variables. All the variables were used in modeling for the purpose of predicting the total income.

The original FIES has 38,400 observations. The model was found to be significant ($p < 0.0001$). However, the r-squared value is just 0.2461. Since the model is used for determining significant predictors of income only and not for explaining the relationships, the r-squared is not expected to be high. Survey regression in STATA was used in order to include the survey weights in the analysis. The sampling weight or survey weight includes the inverse of the probability that the observation is included because of the sampling design in the model. The same set of variables were used in each region to ensure validity of the model.

### *3.3 Statistical Matching*

After running the first regression, all the variables from the CPH that are common to FIES were extracted from the whole data set and statistical matching was done. This part is an important step in any estimation method. It is important that the common variables were measured equivalently. Kolmogorov-Smirnov (K-S) test was done in order to assure that the variables are statistically matched. It was shown that the maximum differences are all less than the critical values from the K-S table, with k-1 degrees of freedom. This causes the failure of rejection of the null hypothesis that the distribution between the two data sets are the same. Thus, the variables can be used to model the income in the CPH data since they have the same distribution in FIES. CPH data is divided into two: rt1 and rt2. The person level characteristics were encoded in rt1 while the household level characteristics were encoded in rt2.

*3.4 Model Building*

In constructing the model in FIES, the FIES data set per region were extracted from the whole FIES 2009 data set. For each region, the data set was divided into four replicates and the models built in the first three replicates were then validated in the last replicate. The common variables which were denoted by X are called the auxiliary variables. The common variables are: family size, the type of building where the family reside, the construction materials of the roof, construction materials of the wall, the floor area, sex of household head, age of the household head, marital status of the household head, highest grade completed by the household head, province, municipality, and barangay. All variables were used in predicting the income except for the location variables. These variables are categorical except for the age of the household head and the family size; thus dummy variables were created in the utilization of regression.

Each region has four building sets and four validation sets. A total of 153 data sets were constructed for the purpose of modeling the total income. Before fitting the model in the regional data, the variable total income (toinc) was tested for normality. The Shapiro-Wilk's (S-W) test for normality was utilized and showed that the errors are not normally distributed ($p < 0.0001$).

Since most of the regression methods rely on the assumption of normality, transformation of dependent variable was done to ensure the aptness of the model using Deterministic Regression. The total income in FIES is expected to have a positively skewed distribution because of the nonnegativity of the values. Thus, the log transformation is the most appropriate transformation. Logarithmic transformation is often used to stabilize the variation in the data. This made the data ready for fitting the model for income. After the transformation, the model was fitted in the first three replicates of FIES and validated in the last replicate.

The Relative Cumulative Frequencies for each model set and validation set was done. The summary of the K-S tests done between the predicted and actual values of the total income in FIES using Deterministic Regression shows that replicates 1 and 3 were not able to preserve the distribution. Thus, it can be concluded that the Deterministic Regression was not able to predict the income that is similar to the actual income in FIES. Since the Deterministic Regression was not able to preserve the distribution of the income, another type of regression imputation was used: The Stochastic Regression. In this case, the estimation of the error term is very crucial. Before fitting the Stochastic Regression, the error term in FIES was tested for normality after fitting the Deterministic Regression.

Since the error is not normally distributed, some transformations were done to normalize the data. However, the transformations did not normalize the error term. In this case, the non-parametric technique called Kernel Density Estimation (KDE) in estimating error was utilized.

After generating the error, the data is now ready for attaching income and generate poverty incidence by adding the error term in the deterministic regression. The generation of uniform random numbers between 0 and 1 is important in many numerical simulations. To ensure randomness of the generated values, the first 1,000 iterations were ignored. The process is called burn-in. The burn-in is a term that describes the practice of ignoring some iterations at the beginning of the generation of random numbers. Since 1,000 errors were generated, 1,000 different models were also produced as bootstrapped values.

*3.5 Attaching Income to CPH Data*

After attaching the errors in the CPH data set, the 1,000 bootstrapped logarithmic incomes were produced for each observation in the data set. Another set of 1,000 bootstrapped columns were produced for the exponential values since income was log-transformed at the start of the modeling. The next 1,000 bootstrapped columns were produced for the per capita income, which is the total income divided by the family size. Another 1,000 bootstrapped columns were produced as an indicator whether a household is poor or not based on the per capita threshold in each province. If the per capita income of a household is less than the per capita threshold of the corresponding province, then the household is considered poor (denoted by "1"), otherwise non-poor (denoted by "0").

After producing 5,000 variables for the estimation of poor households, the data was then collapsed in municipality level to attain the municipal level of poor households in the region. The mean number of poor households out of the 1,000 bootstrapped estimates together, with the standard errors, were used as estimates of the municipality level. The estimates were also calibrated so that they conform to the regional level estimates. Table 1 shows the comparison of distribution of income between the deterministic and stochastic regression.

Table 1 shows that the distribution of income using SRI is closer than the distribution using DRI. The K-S test shows that the maximum difference is less than the critical value which leads to the non-rejection of the null hypothesis. Thus, the SRI produced income that has the same distribution as the true income value in FIES.

**Table 1.** Region 8 Relative Frequency Distribution

| | Income Class | True Value(FIES) | Fitted Value (CPH Deterministic) | Fitted Value (CPH Stochastic) |
|---|---|---|---|---|
| 1 | 12000 to 111999 | 60.79 | 38.11 | 35.57 |
| 2 | 112000 to 211999 | 21.17 | 16.63 | 21.42 |
| 3 | 212000 to 311999 | 7.41 | 23.69 | 14.00 |
| 4 | 312000 to 411999 | 3.58 | 12.94 | 9.02 |
| 5 | 412000 to 511999 | 2.68 | 5.24 | 5.59 |
| 6 | 512000 to 611999 | 1.44 | 1.98 | 3.60 |
| 7 | 612000 to 711999 | 0.65 | 0.78 | 2.31 |
| 8 | 712000 to 811999 | 0.60 | 0.33 | 1.52 |
| 9 | 812000 to 911999 | 0.45 | 0.14 | 1.03 |
| 10 | 912000 to 1011999 | 0.35 | 0.07 | 0.72 |
| 11 | 1012000 to 1111999 | 0.15 | 0.03 | 0.51 |
| 12 | 1112000 to 1211999 | 0.10 | 0.02 | 0.39 |
| 13 | 1212000 and above | 0.65 | 0.02 | 4.32 |
| | **Total** | **100** | **100** | **100** |

Table 2 shows the summary of the number of poor households, poverty incidence, and their corresponding standard errors using SRI with and without calibration. The result shows that the number of poor households produced with calibration is almost the same as the number of poor households in FIES 2009 which is 290,391. The SRI is off by only four households as compared to SRI without calibration.

**Table 2. Summary of Estimates for Stochastic Regression With and Without Calibration**

| Estimates | Without calibration | With calibration |
|---|---|---|
| Number of poor households | 178817 | 290387 |
| Standard error of poor households | 1.6907 | 2.7503 |
| Poverty incidence | 25.11 | 34.23 |
| Standard error of poverty incidence | 2.8479 | 0.0507 |

The poverty incidence is higher with calibration, but the standard error is smaller. Notably, the estimates without calibration are quite far from the estimates with calibration. This does not mean that the SR models are "wrong," since the FIES estimates are subject to sampling error and may, in some cases, be further from the true values. FIES estimates were used to calibrate the produced estimates of SR in CPH.

## 4. CONCLUSIONS

The SRI is better than the DRI in attaching income to CPH. The SRI was able to preserve the distribution of the income as compared to DRI. Since the errors in fitting the DRI in CPH do not follow a well-known distribution such as the Normal distribution, the non-parametric method of estimating error called KDE was used to generate the errors attached in SRI and was found to be effective in using the SR. The calibration technique achieved municipal estimates that conforms to the regional estimates.

As claimed by the PSA in their official poverty statement on December 6, 2019, the official poverty statistics show significant progress in increasing overall income. Accordingly, the poverty incidence for Eastern Visayas dropped by 10.3% in 2018. However, there is still a need in re-evaluating and enhancing the poverty alleviation programs in this region by targeting the poor efficiently using small area estimation (municipal level), especially the procedure utilized in this study. Appendix A shows the municipal estimates of poverty incidence and number of poor households in this region. Through this, the regional and provincial leaders should modify the poverty program depending on the poverty level of a certain municipality because despite the drop, the region remains one of the poorest in the country.

Since the CPH focuses mainly on the socioeconomic variables, the researcher highly recommends that health variables should be included in the small area estimation models. This is because small area estimates based on poverty may not always provide the best possible estimates on health.

## 5. ACKNOWLEDGMENT

The researcher would like to express her deepest gratitude to everyone especially the following persons and institutions for contribution, guidance,

and financial assistance: Dr. Arturo Y. Pacificador, Jr., Dr. Rechel G. Arcilla, Mr. Frumencio F. Co, Mr. Marcus Jude P. San Pedro, De La Salle University, De La Salle University-Dasmariñas (DLSU-D), and Philippine Statistical Research and Training Institute (PSRTI).

## 6. REFERENCES

Global Policy Forum. (n.d.). Retrieved March 29, 2019, from https://www.globalpolicy.org/social-and-economic-policy/poverty-and-development/economic-growth-and-the-quality-of-life.html

Efron, B. (1979). Bootstrap methods: Another look at the Jackknife. *Ann. Statist.* 71-26. Retrieved from http://www.math.ntu.edu.tw/~hchen/teaching/LargeSample/notes/notebootstrap.pdf

Millennium Development Goals. (n.d.). Retrieved March 29, 2019, from http://www.ph.undp.org/content/philippines/en/home/library/mdg/fast-facts-MDGs-in-the-Philippines.html

National Statistical Coordination Board. (2005). Estimation of local poverty in the Philippines. A World Bank Project in cooperation with the National Statistical Coordination Board. *Estimation of Local Poverty in the Philippines. A World Bank Project in Cooperation with the National Statistical Coordination Board*. Retrieved 2017, from https://psa.gov.ph/sites/default/files/NSCB_LocalPovertyPhilippines_0.pdf

Pacificador, A.Y., et al. (1996). Attaching Income and Expenditure Dimension to the 1990 Census of Population and Housing. A research study funded by the UNFPA-NSO Project PHI/93/P01-Utilization and Dissemination of Demographic Data.

Sarndal, C. E., Swenson, B, and Wretman, J. H. (1989). The weighted residual technique for estimating the variance of the general regression estimator of the population total. *Biometrika*, **76**, 527-537.