A Procedure for the Generation of Small Area Estimates of Philippine Poverty Incidence

Nelda A. Nacion
Mathematics and Statistics Department, De La Salle University

Arturo Y. Pacificador
Institute of Statistics, University of the Philippines Los Baños

The main purpose of this study is to propose an alternative procedure in the generation of 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. The difference of this methodology as compared to the ELL is that the error terms used for predictions is based on the empirical distribution of such residuals and thereby a protection against misspecification of the error model. At the same time, the procedure is simpler with available computing resources. In addition, the proposed methodology only utilized data from the census short form which is a 100% percent sample. Thus, eliminating another source of variation as compared to using the census long-form which is collected from a sample of households. This study used the Family Income and Expenditure Survey (FIES) 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. The researcher was able to produce small area estimates of poverty in the Philippines at municipal level by combining survey data with auxiliary data derived from census. The study fitted different models for each region. Considering the results of this paper, the following conclusions were derived: The Stochastic Regression Imputation (SRI) is better to use as compared to Deterministic Regression Imputation (DRI) in attaching income to CPH. The SRI was able to preserve the distribution of 82% of the total number of regions. The DRI was able to preserve only 10.22% of the validation sets. Since the error in fitting the DRI in CPH does not follow a well-known distribution (such as the Normal distribution), the non-parametric way of estimating error was used to generate the errors attached in SRI. The technique is called Kernel Density Estimation (KDE) or the histogram method, which was found to be effective in using the SR. Using the calibration technique achieved municipal estimates that conforms to the regional estimates. The estimates of the poor households in CPH reflects the bottom 30% of the wealth index.

Keywords: *small area estimation, deterministic regression, stochastic Regression, calibration, poverty incidence*

1. Introduction

As defined by World Bank, poverty is living on less than $1.90 a day. According to the estimates released in 2017, 9.2% (689 million) of the world's population is living on less than $1.90 a day and is considered poor. 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. Because of this, poverty alleviation has always been a part of each country's development programs. In fact, the first Sustainable Development Goal (SDG) is to eradicate poverty in all its forms by 2030.

The programs of the government should be in response to the first goal of SDG. 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, the provinces and municipalities are considered small area level in the context of income-based poverty measures using FIES and CPH. Although, the FIES was redesigned to provide reliable estimates at the provincial level already. For some years, the NSCB has been producing estimates of poverty incidence at regional level. However, there has been an increasing demand from policy makers for a more disaggregated set of poverty statistics so that the poverty alleviation programs could target the poor areas efficiently. To address this, the NSCB released estimates of poverty incidence at provincial level in 2003 based on FIES. However, due to small sample sizes at this level, large standard errors were evident. The small area estimation allows the possibility of having improved estimates even for a finer level of disaggregation to municipality level by combining FIES with information from a recent census. For the government 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.

Small Area Estimation (SAE) refers to a set of statistical techniques used to improve sample survey estimates through the use of auxiliary information (Gosh and Rao 1999). The process starts by identifying a variable to be estimated over a range of small subpopulations corresponding to small geographical areas such as provincial or municipal. Due to a very small sample size in the area, the direct estimates will have a large standard error and therefore not reliable. Thus, it is necessary to use indirect estimates to have a precise measurement of the variable of interest. The basic idea is to borrow strength by using variables of interest from the related areas, thus increasing the effective sample size. These values are used into the estimation process through a model that provides a link to related areas through the use of supplementary information related to the variables of interest, such as recent census counts and current administrative records (Rao 2003).

Estimates are required for a diversity of domains and the types of domains should influence the choices both of design and estimation (Kish1980). Further, the sizes of domains also influence the choice of methods for design and estimation. There are three types of domains: the main domain, the minor domains, and mini domains. Most national surveys are designed to only provide reliable estimates of the main domain and at some instance, at minor domains. In the Philippine set up, main domains are regional levels and to some extent at the provincial level. Thus, no estimates are available at minor domains or municipal level. The national surveys cannot provide a direct estimate at municipal level because of the large sample size requirement which the government cannot afford. Hence, small area estimation is necessary.

There are many SAE techniques available. So far, in the Philippines, the methodology done to generate small area statistics is the one conducted by the National Statistical Coordination Board (NSCB) which is now part of the Philippines Statistical Authority (PSA) 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. The ELL methodology utilized the Census of Population and Housing (CPH), the Family Income and Expenditure Survey (FIES), and the Labour Force Survey (LFS). 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 the latest development of SAE, there are parallel initiatives that use satellite imagery for mapping poverty. Asian Development Bank (ADB), in their September 2020 issue published an article about Integration of Big Data (particularly in the form of geospatial data and mobile phone data) in Small Area Estimation Framework. According to Eagle et al. (2020); Data2X (2017), it has a potential of enhancing the compilation of a wide range of development statistics. Marchetti et al. (2015) explain three possible approaches to integrating big data into small area estimation framework. However, these approaches have disadvantages as well.

The Asian Development Bank, in collaboration with the Philippine Statistics Authority and the World Data Lab, conducted a feasibility study to enhance the granularity, cost-effectiveness, and compilation of high-quality poverty statistics in the Philippines through poverty mapping. This poverty mapping explored the potential use of satellite imagery in enhancing the geographical disaggregation of government-compiled poverty and population data in the Philippines, where the government releases poverty indicators at the municipal and/or city level every 3 years and population data at the municipal or city and barangay level every 5 years. The findings of this study using Philippine datasets are promising, despite employing nonproprietary images with resolutions that are not as refined as those in proprietary images. The poverty predictions were generally consistent with government-published poverty data, and the methodology produced more geographically disaggregated estimates of poverty.

An initiative called Thinking Machines utilized similar satellite-image-based methodology in predicting wealth (Tingzon et al. 2019). However, since they used data that are solely based on surveys in training machine-learning algorithms, the poverty estimates used during the training are less prone to large sampling error. A similar study was done in Thailand that focused on poverty correlating on night lights (Dorji 2019). However, the geographic disaggregation results in even less reliable poverty statistics. Thus, the estimates derived was referred to as government-published estimates. Generally, the government's effort to explore the feasibility of using imagery as an alternative data source for poverty does not intend to substitute conventional sources of poverty data but addresses limitations associated with traditional techniques.

In this regard, a similar approach as ELL was done by Pacificador et al. (1996). They conducted the study "Attaching the Income and Expenditure Dimension to the 1990 Census of Population and Housing (CPH)." The study 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 imputation approach in coming up with attaching income which can be used to generate small area estimates of poverty incidence. The ELL also uses Stochastic Regression while the latter uses Deterministic Regression. 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. In the class mean imputation, 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 the 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.

Building up on the weakness of the previous 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 difference of this methodology as compared to ELL is that the error terms used for predictions is based on the empirical distribution of such residuals and thereby a protection against misspecification of the error model. At the same time, the procedure is simpler with available computing resources. In addition, the proposed methodology only utilized data from census short form which is 100% sample. Thus, eliminating another source of variation as compared to using long-form which is collected from a sample of households. The imputation procedure is that of a unit level and not area level. Moreover, the final estimates were calibrated so that they conform to the regional estimates of poverty incidence in the Philippines. Some statistical tests were also done to assess the precision of the estimates.

2. Common Variables

As mentioned, small area estimation starts by identifying common variables between the two sources FIES and CPH which were denoted by **X** and then the target variable **Y** (total income) was modelled when the same auxiliary information is available for both surveyed and census households. A matrix relationship between **Y** and **X** can be estimated using the survey data, for which both the target variable and the auxiliary variables are available, either at household level or as subgroup means at a higher level of aggregation. β represents the regression coefficients giving the effect of the **X**'s or auxiliary variables on **Y** (the total income of the household), and $\boldsymbol{u}$ is a random error term representing that part of the income that cannot be explained using the auxiliary information.

Auxiliary Variables is defined as the set of variables which are not part of the main analysis that can help to make estimates on incomplete data (Collins et al. 2001). Since the objective of the study is to come up with an estimate of income to be attached to the CPH, the auxiliary variables came from FIES. These variables were used to estimate poverty and the number of poor households.

Common variables between the survey and the census that were measured in common were identified and the appropriate statistics were compared. For categorical variables, such as type of house/building the family reside, construction material of the roof, construction material of the outer wall, proportions were used to compare for each category. In the case of numerical data such age of the household head, means and standard deviations were compared. To test whether the variables are measured equivalently, or of the same distribution since they came from two different years, the Kolmogorov Smirnov test (K-S) was utilized. Time invariant variables were not considered since the time difference is just one year and the assumption is that the estimates of poverty are not statistically different from each other.

After identifying the common variables between the FIES and the CPH, a richer data set was used in building a model for the income. It should be taken into consideration that the set of auxiliary variables were measured in a consistent way in both data sources. In a case where the auxiliary variables are not measured in the same way, some recoding and transformation of the variables were done. The distribution was also tested using the Kolmogorov Smirnov Test (K-S test).

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. The dummy variables that were created was attached in rt1, the person level characteristics.

After identifying the common variables, a first step regression was done in the original FIES data to determine the significant variables in predicting the income. Some measures of association were also done to determine the association of the auxiliary variables to the total income in FIES. In searching for possible relationships, statistical significance of the results of regression were considered. If there exists a non-linear relationship based on plots, squares or other transformations of the numerical variables was also considered. Careful selection of appropriate predictor was done to ensure that the model will not over-fit, so the number of predictors included in the model should be smaller than the number of observations in the survey. After identifying the significant predictors, the same variables were created in the CPH data so that the two data sets are statistically matched. When the comparison is done and the variables are already matched, rt1 and rt2 data sets were merged based on the person level characteristics to create household level variables that can be found in FIES. When rt1 and rt2 were merged, the number of observations is the same as the observations in rt2. When the merging is done, and the variables are matched between FIES and CPH, the FIES data is ready in modeling the income for the purpose of prediction. Since the regression models were used for prediction of income and not for explaining the phenomenon, the model fit (r-squared) is not expected to be high.

3. The ELL Methodology

The Elbers, Lanjouw and Lanjouw (ELL) methodology was designed specifically for the small-area estimation of poverty measures based on per capita household expenditure. The first step is to identify a set of auxiliary variables X that are in the survey and are also available for the whole population. The model $Y_{ij} = X_{ij}\beta + C_i + e_{ij}$ is then estimated for the survey data, by incorporating aspects of the survey design for example through use of the "expansion factors" or inverse sampling probabilities. The residuals u_{ij} from this analysis are used to define cluster-level residuals $\hat{c}_l = \hat{u}_l$, the dot denoting averaging over j, and household-level residuals $e_{ij} = c_i + u_i$.

It is usually assumed that the cluster-level effects c_i all come from the same distribution, but that the household-level effects e_{ij} may be heteroscedastic. This can be modelled by allowing the variance σ_e^2 to depend on a subset Z of the auxiliary variables

$$g\left(\sigma_e^2\right) = Z_\alpha = r,$$

where g(.) is an appropriately chosen link function, α represents the effect of Z on the variance and r is a random error term. Fuji (2004) uses a version of the more general model of ELL involving a logistic-type link function, fitted using the squared household level residuals. Fuji's model is

$$\ln\left(\frac{\widehat{e_{ij}^2}}{A - \widehat{e_{ij}^2}}\right) = Z_{ij}\alpha + r_{ij} \quad .$$

From this model, the fitted variances $\widehat{\sigma_{e_{ij}}^2}$ can be calculated and used to produce standardized household-level residuals $\widehat{e_{ij}^*} = \frac{\widehat{e_{ij}}}{\widehat{\sigma_{e_{ij}}^2}}$. These can then be mean-corrected or mean-centered to sum to zero, either across the whole survey data set or separately within each cluster.

In standard applications of small-area estimation, the estimated model $Y_{ij} = X_{ij}\beta + C_i + e_{ij}$ is applied to the known X values in the population to produce predicted Y values, which are then averaged over each small area to produce a point estimate, the standard error of which is inferred from appropriate asymptotic theory (Elbers, Lanjouw, and Lanjouw 2003).

4. Imputation Procedures

Aside from the ELL Methodology, this paper was modified in such a way that the generation of estimates (poverty incidence and number of poor households) came from the Imputation-like Methods. The Imputation Methods used are the Deterministic and Stochastic Regression. Few statistical problems can be regarded as missing-data problems which lead to the evolution of missing-data analysis, contributing to new insights to data fusion (file matching) or post stratification and weighting (Meinfelder 2009). Meinfelder further explained that there are several techniques in handling missing values in data sets, ranging from partially or completely erasing incomplete cases to filling-in the gaps. Filling-in the gaps is called imputation.

Imputation can be used for a general-purpose analysis if carried out in a sensible way (Meinfelder 2009). There are two approaches of imputation. Single and multiple imputation (MI). According to Rubin (1987a), as stated in the paper of Meinfelder (2009), the MI method is recently the most popular technique in countering insufficient information. However, majority of the studies about MI is confounded only on fully parametric variants with the corresponding distributional assumptions. Conversely, surveys most of the time produce mixed-scale data and the predictors are usually non-conforming to any statistical distributions. Furthermore, some multivariate methods attempt to model continuous data that focuses on cell combinations of qualitative variables. Because of these problems, some authors like Schafer (1997) suggested to impose some restrictions on those cell combinations via log-linear models.

Deterministic Regression Imputation (DRI)

Deterministic Regression Imputation (DRI) replaces missing values with the exact prediction of the regression model and the random variation (or the error term) is not considered. This leads to an overestimation of the correlation between X and Y because the imputed values are often too precise. Moreover, DRI can also underestimate the variance of the estimates. It can also distort the distribution of the data. One of the major disadvantages of this method is it can produce negative values which are out of range or unfeasible values. The model for DRI is given by:

$$\widehat{y_k} = \widehat{\beta_0} + \sum \widehat{\beta_1} X_{ik},$$

where $\widehat{y_k}$ is the predicted value under the *k*th nonresponding unit to be imputed, $\widehat{\beta_0}$ and $\widehat{\beta_1}$ are the parameter estimates X_{ik} is the auxiliary variable that can either be a quantitative variable or a dummy variable under the *k-th* nonresponding unit.

Stochastic Regression Imputation (SRI)

In Stochastic Regression Imputation (SRI), the error term is added to the predicted value and is therefore able to produce correlation of *X* and *Y* more appropriately. This method was developed in order to solve the issue of DRI. The model for SRI is given by:

$$\widehat{y_k} = \widehat{\beta_0} + \sum \widehat{\beta_1} X_{ik} + \widehat{e_k},$$

where $\widehat{y_k}$ is the predicted value under the *k*th nonresponding unit to be imputed, $\widehat{\beta_0}$ and $\widehat{\beta_1}$ are the parameter estimates, X_{ik} is the auxiliary variable that can either be a quantitative variable or a dummy variable under the *k-th* nonresponding unit and $\widehat{e_k}$ is the randomly chosen residual for *k-th* nonresponding unit.

5. Estimating Error

In this paper, the error is defined as the difference between the fitted value and the actual value given by

$$e = y - \hat{y}.$$

Different methods of SRI were utilized depending on the error component. Since the identification of the distribution of the error terms is crucial in this analysis, it was assessed using the following steps: (1) Standard Diagnostic Plots was utilized such as the Normal quantile-quantile (Q-Q) plots and histograms. The Q-Q plot is a plot of quantiles for the observed residuals against those computed from a theoretical normal distribution. (2) K-S test and Shapiro Wilk Test: the formal tests of normality were used. (3) Since the errors do not follow a well-known distribution such as normal distribution, transformation of the dependent variable was also done to address the case of nonnormality of residuals and the heterogeneity of the residual variances. Transformations used was the logarithm. 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.

The method used in estimating the errors in SRI was done addressing the nonnormality of residuals: the non-parametric estimation called Kernel Density Estimation (KDE). 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 and the bin width . 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.

Let $f(x)$ denote the density function of a continuous random variable. The kernel density estimates of $f(x)$ at $x = x_0$ is then

$$\hat{f}(x_0) = \frac{1}{Nh}\sum_{i=1}^{N} K\left(\frac{x_i - x_0}{h}\right).$$

where $k(z)$ is a kernel function that places greater weight on points x_i that are closer to x_0. The kernel function is symmetric around zero and integrates to one. Either $k(z) = 0$ if $|Z| \geq z_0$ for some z_0, or $k(z) \rightarrow 0$ as $z \rightarrow \infty$.

A kernel density plot requires the choice of a kernel function, $|z| < \sqrt{5}$ and a bandwidth h. Then a kernel density function will be evaluated at several values x_0, and then the estimates will be plotted against x_0. In Stata, the k density command produces the kernel density estimate. The default kernel function is the Epanechnikov kernel, which sets $k(z) = \frac{\left(\frac{3}{4}\right)*\left(1 - \frac{z^2}{5}\right)}{\sqrt{5}}$ for $|z| < \sqrt{5}$ and zero otherwise. This kernel function is said to be the most efficient in minimizing the mean integrated squared error. After creating the KDE plot, another variable was created for the estimation in the data set then a random sample from the distribution was selected and added in the model.

The following steps were done in estimating the error: (1) The Deterministic Regression model was fitted in FIES and the residuals were assessed. (2) Since the residuals do not follow a normal distribution, the frequency distribution was created for each region. (3) After the creation of frequency distribution of residuals together with the cumulative relative frequencies, the Class Mark (CM) for each error class was obtained. (4) The errors used in the model were obtained by replacing the CM in each randomly generated uniform number. Just like the DRI, the SRI has also its advantages and disadvantages. SRI can also produce imputed values that are near to the nonresponse observation if the model has a high R^2. Negative and unrealistic values can also be produced like in DRI.

The total income attached in CPH was regarded as missing data (Missing Completely at Random) and predicted using both the Deterministic and Stochastic Regression Imputation. Both Deterministic and Stochastic Regressions are known procedures in imputation methods.

6. Validation of Estimates

Bootstrapping

Bootstrapping, according to Efron and Tibshirani (1993) as mentioned in the study conducted by NSCB, is the name given to a set of statistical procedures that use computer- generated random numbers to simulate the distribution of an estimator. Each complete set of bootstrap values will yield a set of small-area estimates. The mean and standard deviation of a particular small-area estimate yields a point estimate and standard error for that particular area. In the case of poverty incidence, a large number of alternative predicted values for each household will be constructed in such a way as to take account of variability of the predicted values.

$$Y_{ij}^{b} = X_{ij}\beta^{b} + h_{i}^{b} + e_{ij}^{b}, b = 1, \ldots B,$$

where X_{ij}'s are the independent variables, $\hat{\beta}$ is an unbiased estimator of β with variance V_{β}, so each β^{b} will be drawn independently from a multivariate normal distribution with mean $\hat{\beta}$ and variance matrix V_{β}. The cluster level effects h_{i}^{b} will be taken from the empirical distribution of h_i. To take account of heteroscedasticity in the household-level residuals, α^{b} will be drawn from a multivariate normal distribution with mean $\hat{\alpha}$ and variance matrix V_{α}, it will be combined with Z_{ij} to give a predicted variance and use this to adjust the household-level effect:

$$e_{ij}^{b} = e_{ij}^{*b} x \sigma_{e,ij}^{b}$$

where e_{ij}^{*b} represents a random draw from the empirical distribution of e_{ij}^{*b}, either for the whole data set or just within the cluster chosen for h_i.

For each complete set of bootstrap values Y_{ij}^{b}, for a fixed value of *b,* a set of small area estimates will be generated. The mean and standard deviation of a particular small-area estimates, across all *b* values, will yield a point estimate and its standard error for that area. The following are the basic steps of bootstrap: (1) From the empirical distribution function, F_n, draw a random sample of size *n* with replacement. (2) The statistic of interest will be calculated, which will be the standard error of the small-area estimates. (3) Step 2 will be repeated *B* times where *B* is a large number to create *B* resamples. (4) A relative frequency histogram will be constructed from the *B* number by placing a probability $\frac{1}{B}$ at each point. The distribution obtained is the bootstrapped estimates of the sampling distribution of poverty incidence. This distribution can now be used to make inferences of the parameter being estimated. The final estimates of poverty incidence are the average poverty incidence with standard error for 1000 bootstrapped values.

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.

Since some of the ELL methodology implementations have fitted separate models for each stratum, this study also used different models for each of the stratum defined by the survey. In this study, different models were fitted in each of the different regions which were built using the first three FIES replicates and the model built was validated in the last replicate. A total of four model building sets and four model validation sets were used for each region. The table below shows the model validation strategies.

Table 1. Model Validation Strategies

Model Building Set	Validation Set
Replicates 1, 2, and 3	Replicate 4
Replicates 1, 2, and 4	Replicate 3
Replicates 1, 4, and 3	Replicate 2
Replicates 4, 2, and 3	Replicate 1

This part is the modification made by the researcher from the methodology employed by NSCB to check the accuracy of the models using DRI technique in predicting income.

To determine whether imputation method using DRI was able to maintain the same distribution of the actual data, the K-S test was utilized. The K-S test is a goodness of fit test concerned with the degree of agreement between two sets of sampled observations and some specified theoretical distribution (Siegel, 1988) as mentioned in the paper of Cortes (2007). In addition, this test was used in comparing the actual data of income in FIES and the imputed values from the different models to determine if the imputation method using DRI was able to maintain the same distribution of the actual data.

7. Poverty Measures

One of the poverty measures is the household income. The minimum income required to meet basic food needs and satisfy the nutritional requirements set by the Food and Nutrition Research Institute (FNRI) to ensure that one remains economically and socially productive is the Food Threshold. According to PSA's report in 2018, it is used to measure extreme or subsistence poverty. Further, the poverty threshold is a similar concept, expanded to include basic non-food needs such as clothing, housing, transportation, education and health expenses. In 2018, a family of five members needed at least Php 7,337 every month, on the average to meet the family's basic food needs. While at least Php 10,481 every month on the average to meet both basic food and non-food needs. These amounts represent the monthly food and monthly poverty threshold, respectively. Thus, households with income below the threshold are considered poor. This poverty statistics is essential in the government's priority concern in alleviating poverty in the country. It is used in identifying the severity of the poverty of Filipinos so that appropriate interventions can be done. In this paper, the researcher adopted the same set of indicators to estimate the poverty incidence in the Philippines at a lower level of aggregation (municipality level).

8. Production of Poverty Incidence

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 within the region.

In producing the final estimates, the poverty incidence was computed as:

$$P_R^b = \frac{\sum_{n_{ij}} I(E_{ij}^b < z)}{\sum_{ij \in R} n_{ij}},$$

where n_{ij} is the size of household ij in R and $I\left(E_{ij}^b < z\right)$ is an indicator function (equal to 1, when the per capita income is below the poverty line/threshold and 0, if 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 were obtained by parametric bootstrap. In this paper, the bootstrap method was applied as follows: (1) The selected model was fitted in the census data to predict the income. (2) Once the income has been predicted, another column was added which will determine if an individual is poor not using the poverty threshold for 2009 per province. (3) The total number of the poor was computed for each municipality. (4) The process was repeated 1000 times independently. The mean and standard error for the 1000 bootstrapped values served as the estimates for mean and standard error for each municipality.

9. Calibration of Estimates

One problem arising from the Small Area Estimation is that the estimates may not conform to the higher level for which errors are actually small and stable. In example, some estimates of poverty in lower level such as municipalities are not actually corresponding to its provincial and regional estimates. To achieve internal consistency, calibration of estimates is necessary to match the sample estimates to its population estimates. According to Park and Kim (2009), calibration estimation, where the sampling weights are adjusted to make certain estimators match known population totals, is commonly used in survey sampling.

Consider the problem of estimating the population total $Y = \sum_{i=1}^{N} y_i$ for a finite population of size *N*. Let *A* denote the index set of the sample obtained by a probability sampling scheme and let *y* be observed in the sample. The Horvitz-Thompson (HT) estimator of the form

$$\widehat{Y_d} = \sum_{i \in A} d_i y_i$$

is unbiased for *Y*, where $d_i = \frac{1}{\pi_I}$ is the inverse of the first order inclusion probability of unit *i* in the population. The weight d_i is often called the design weight since it directly obtained from the sampling design.

If, in addition to y_i, an auxiliary variable vector x_i, is available from the sample and the population total $X = \sum_{i=1}^{N} x_i$ is known, it is possible that

$$\Sigma_{i \in A} d_i x_i \neq X.$$

The class of calibration estimators, calibrated to *X*, is the class of estimators of the form

$$\widehat{Y_w} = \Sigma_{i \in A} w_i y_i ,$$

where w_i satisfies

$$\Sigma_{i \in A} w_i x_i \neq X.$$

Thus, we allow the final weight, w_i to be a function of x_i but not y_i.

In this study, the bootstrapped total number of poor was calibrated using the formula:

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

where $\widehat{Y_m^*}$ = calibrated municipal estimates;

$\widehat{Y_m}$ = municipal estimates from CPH

$\widehat{Y_R}$ = regional estimates from FIES;

$\widehat{Y_R^*}$ = total municipal estimate per region from CPH.

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

10. External Validation

Wealth index

The estimates that were produced were compared externally to wealth index. According to the Demographics and Health Survey (DHS) Program, wealth index is a composite measure of household's cumulative living standard. The wealth index is calculated using easy-to-collect data on a household's ownership of selected assets, such as televisions and bicycles; materials used for housing construction; and the types of water access and sanitation facilities. These variables are also available in CPH and FIES, which made the estimates comparable. The wealth index was computed using the CPH data and was compared to the estimates produced.

Computation of wealth index

Computation of wealth index was done by the researcher and was compared to the computed estimates to ensure the accuracy of the estimates of poverty. The computation was done using CPH form 3 (long form). The indicators were from the response in questions H14 and H15. In addition to this set of questions, the construction material of the roof and wall were also considered.

If the answer to the questions mentioned above is no, then it was considered as an indicator that the household is poor, otherwise non-poor. If the response for the construction material of the roof are either the codes 4, 5, 6, 7 and 8, then the household is considered poor, otherwise non-poor. For the construction material of the wall, the response from 4-10 also is an indicator that the household is poor, otherwise non-poor. The variable weight was calculated using the formula

$$W_i = 1 - \left(\frac{\sum h_i}{n} \right)$$

where: W_i = variable weight for the i^{th} non-poor indicator

h_{aa} = total number of households with i^{th} non-poor indicator

n = total number of households

The household score is computed as the sum of variable weights. Zero weight is assigned to the non-poor indicator that is not present in the household.

EA level Wealth Index

The Wealth Index for Enumeration Area level was computed as follows:

$$EA_{wealthindex} = \text{average of Household score within EA.}$$

Using CPH form 3

The Wealth Index was computed using 2 options: (1) National level (Weight/Score); and (2) Domain level (Province/HUCs)-Weight/Score. In this paper, the domain level wealth index was computed as the comparison was also done by region. The result of wealth index was compared to the poverty incidence computed in the proposed procedure. It is expected that the regions with high poverty incidence will have low wealth index so that the produced estimates are accurate. The wealth index per region were also correlated to the computed poor households to ensure the accuracy of the estimates.

11. Illustration: Small Area Estimation of Poverty Uncidence

Data sources

Family Income and Expenditure Survey (FIES), 2009

This study utilized the Family Income and Expenditure Survey (FIES 2009). In the Philippines, it is conducted by the PSA (formerly NSO) every three years. According to the report prepared by the World Bank in cooperation with the NSCB in 2005, the FIES contains information on household income, expenditure and consumption in addition to socio-demographic characteristics. Selected households are interviewed in two separate operations, each covering a half-year period, in order to allow for seasonal patterns in income and expenditure and improve accuracy of responses by shortening recall period.

The FIES 2009 employed the new Master Sample (MS) designed by NSO in 2003. An MS is defined as a sample from which subsamples are drawn to serve the needs of several surveys. The use of MS promote efficiency on the use of limited resources and it also allows the linking of the different survey variables in creating a richer database for a more meaningful and useful analyses. Usually, an MS is an area sample of clusters of households referred to as Primary Sampling Units (PSUs). The 2003 MS design covers all households in the Philippines excluding institutional households as well as households in the Least Accessible Barangays

(LABS). Barangays were considered as PSU. However, more than half of the barangays does not satisfy the size requirement. Thus, for the 2003 MS, a PSU is defined as a barangay or group of barangays containing at least 500 households. Further, a PSU is composed of enumeration areas (EAs). An EA is defined as compact segment of the PSU comprising on the average about 265 households. Thus, on the average, each PSU formed is composed of 3.5 EAs.

According to the NSO report on FIES in 2017, the 2003 MS consists of a sample of 2,835 PSUs. The whole data set was divided into four sub-samples or independent replicates wherein one replicate contains one fourth of the total PSUs; a half sample contains one-half of the four sub-samples or equivalent to all PSUs in two replicates. The final number of sample PSUs for each region was determined by first classifying PSUs as either self-representing (SR) or non-self-representing (NSR). Additionally, to facilitate the selection of sub-samples, the total number of NSR PSUs in each region was adjusted to make it a multiple of 4. These replicates were then used in the validation of estimates.

Census of Population and Housing (CPH), 2010

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. The NSO conducted the 2010 CPH in May 2010. This is the 13th census of population and the 6th census of housing undertaken in the country since 1903 (CPH 2010). It is designed to take an inventory of the total population and housing units in the Philippines and to collect information about their characteristics. The CPH commonly collected data from all households in CPH Form 2 – Common Household Questionnaire and CPH Form 3 – Sample Household Questionnaire.

A combination of complete enumeration and sampling of households was done to obtain population count and data on basic characteristics of the household population and housing units. For institutional population, a complete enumeration was done. The non-sample households were interviewed using the Common Household Questionnaire (CPH Form 2) while the sample households were interviewed using the Sample Household Questionnaire (CPH Form 3). On the other hand, institutional population in institutional living quarters were enumerated using the Institutional Population Questionnaire (CPH Form 4). These questionnaires were used to gather information on the demographic and socio-economic characteristics of the population, as well as the characteristics of households and housing units.

The CPH provides data that can be used for estimating reliable small area statistics such as the municipality level indicators for poverty. It is one valuable source of small area estimation. However, the CPH easily gets out of date and has a very limited indicators that can be used. It is important to note that the CPH data set used in this paper are the information coming from the short form (form2) which is free of sampling error because of 100% coverage.

Common variables and association of auxiliary variables with income

After careful checking on the questions between FIES and CPH, the common variables were found. 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, and highest grade completed by the household head. All variables except the age of the household head and family size were recoded to ensure that the variables were measured in the same way between the two data sets for the modeling purposes.

After identifying the common variables, a first step regression using dummy variables 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 since the probability of F is 0.00, which is also less than 0.05. 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.

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. Kolmogorov-Smirnov test was done in order to assure that the variables are statistically matched. The K-S tests were done for each of the common variables between FIES and CPH. 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.

As mentioned, 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. The person level characteristics encoded in rt1 are the following: sex of the household head, age of the household head, marital status of the household head, and highest grade completed of the household head. The household level characteristics encoded in rt2 are the following: building type, construction material of the walls, construction material of the roof, floor area, family size, province, municipality and barangay. The rt1 data was collapsed by geographical characteristics and the housing serial number with codes greater than or equal to 777777 in rt2 were dropped in the data so that the two data sets have equal number of observations. When the observations are already equal between rt1 and rt2, the two data sets were merged and was considered as one data set where the income was attached per region. The code 777777 refers to non-usual residence, 888888 refers to foreign diplomats, 888888 refers to vacation/ rest house and 999999 refers to vacant housing units.

Model building

The ELL methodology constructed different models for different areas. The methodology employed in this paper constructed also different models per region to allow the variation of income due to location. 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 were built in the first three replicates then validated in the last replicate. 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 test for normality was utilized. The result implies that the residuals are not normally distributed since it does not form a line that suggests a normal distribution.

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. After the transformation, the model was fitted in the first three replicates of FIES and validated in the last replicate. The summary of Kolmogorov-Smirnov tests done between the predicted and actual values of the total income in FIES using Deterministic Regression implies that most of the predicted values of income was not able to preserve the distribution of the actual total income in FIES using the model building set and the validation set. Most of the maximum differences obtained between the two distributions were greater than the K-S critical values that led to the rejection of the null hypothesis. Among the 68 models fitted, only 7 were able to preserve the distribution of the actual income. The FIES replicates 4 and 1 of Region 2, replicate 3 of Region 4b, replicate 3 of Region 9, replicate 3 of Region 10, replicates 2 and 4 of Region 15, and replicate 2 of Region 16. The rest of the models 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.

Other methods of assessing the performance of Deterministic Regression were also done such as Mean Deviation (MD), Mean Absolute Deviation (MAD) and Root Mean Square Deviation (RMSD). The mean deviation of all the replicates of 16 regions is P39,987.33 and a root mean square deviation of P274.87. The same were implied by the values of mean deviation and mean absolute deviation. It was also noticed that some of the mean deviations are negative. This means that the model was overfitted, implying that the predicted values are greater than the actual values of total income in FIES. Thus, the Deterministic Regression was not able to preserve the distribution of the true 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. The Stochastic Regression is basically the same as the Deterministic Regression, but with the addition of the error term. In this case, the estimation of the error term is very crucial. Before fitting the Stochastic Regression, the error terms in FIES were 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 such as getting the square and the cube root, even the standardization did not work for the error terms to be normally distributed. In this case, the non-parametric technique in estimating error was utilized. The non-parametric technique used is called Kernel Density Estimation (KDE).

The KDE is a non-parametric method of estimating error. A frequency distribution of error was first constructed after fitting Deterministic Regression in FIES. The errors were generated by replacing the Class Marks of the error classes for every uniform generated numbers. The procedure was done 1000 times and the errors were attached to the CPH data set.

After generating the error, the data is now ready for attaching income and generate poverty incidence by adding the error terms 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 1000 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. In this study, the first 1000 iterations were thrown away. Since 1000 errors were generated, 1000 different models were also produced as bootstrapped values.

Attaching income to CPH data

After attaching the errors in the CPH data set, the 1000 bootstrapped logarithmic incomes were produced for each observation in the data set. Another set of 1000 bootstrapped columns were produced for the exponential values since the income was transformed to logarithm at the start of the modeling. The next 1000 bootstrapped columns were produced for the per capita income. The per capita is the total income divided by the family size. Another 1000 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 5000 variables for the estimation of poor households, the data was then collapsed by municipality level to attain the municipal level of poor households in each region. The mean number of poor households out of the 1000 bootstrapped estimates together with the standard errors were used as estimates of the municipality level. The estimates were also calibrated so that it conforms to the regional level estimates.

Table 2 shows the summary of comparison of distribution between the true values in FIES and SR for each region using Kolmogorov-Smirnov (K-S) test. The test rejects the null hypothesis that the two populations have the same distribution if the maximum difference (D-stat) is greater than the K-S critical value with n–1 degrees of freedom, where n is the number of classes. As mentioned,14 out 17 regions or 82% of the total number of regions have the same distributions. The SR was able to preserve the distribution of most of the regions in the Philippines. Comparing the performance of the DR, SR still have the higher percentage of preserved distribution. Based on the result of DR, only 10.29% of the validation model of income was preserved. Hence, the SR is still better to use in attaching income in CPH.

Since SR was proven to be better in attaching the income in CPH in this paper, table 3 shows the estimated number of poor households and poverty incidence together with the standard errors. The estimates are the result of the 1000 bootstrapped estimates. The mean number of poor households and their standard errors were the estimates per municipality. Table 3 shows the aggregated values in regional level. Using the provincial poverty threshold, the poverty incidence (pi) was computed for each region. The table shows that the poorest region is Region 12 (SOCCSKSARGEN) with 53.13% pi, followed by Region 2 (Cagayan Valley), with 28.61% pi, and Region 16 (CARAGA) with 26.99% pi. While the region with the lowest poverty incidence is Region 13 (NCR) with only 2.48% pi.

Looking at the standard errors, it is noticeable that the values are too small especially in the poverty incidence except regions 4b and 8. The standard error

Table 2. Summary of Comparison of Distribution between the True Values in FIES and Fitted Values in CPH using Stochastic Regression

Region	Maximum Difference	K-S Critical Value	Degrees of Freedom	Decision
1	0.029	0.37543	12	Fail to reject
2	0.0179	0.37543	12	Fail to reject
3	0.1052	0.37543	12	Fail to reject
4a	0.1511	0.37543	12	Fail to reject
4b	0.4149	0.37543	12	Reject*
5	0.2361	0.32733	16	Fail to reject
6	0.0692	0.32733	16	Fail to reject
7	0.2621	0.32733	16	Fail to reject
8	0.2526	0.37543	12	Fail to reject
9	0.2368	0.32733	16	Fail to reject
10	0.1845	0.32733	16	Fail to reject
11	0.3291	0.37543	16	Reject*
12	0.1939	0.37543	12	Fail to reject
13	0.3131	0.37543	12	Fail to reject
14	0.2212	0.37543	12	Fail to reject
15	0.3901	0.37543	12	Reject
16	0.2658	0.37543	12	Fail to reject

depends on the number of factors. The poorer the fit of the model, the greater will be the standard error of the small area estimates. Also, the standard error of a particular small area estimate is intended to reflect the uncertainty in the estimate. Generally, the standard errors decrease proportionally as the square root of the population size. Standard errors are small at higher geographic levels but not at lower levels. The bootstrapping methodology includes the variability in the regression coefficients. Since the researcher used different models for each region, the sample size was affected by the explanatory variables included in the auxiliary information. There is a tendency that the model was over-fitted. Due to bootstrapping methodology which uses a finite number of bootstrap estimates (in this case is 1000) to approximate the distribution of the estimator, there was an uncertainty in the estimates and in the standard errors. Thus, the standard errors are small due to high number of bootstrap simulations.

Table 3. Regional Level Estimates using Stochastic Regression

Region	Number of poor households	standard error	Poverty incidence (%)	standard error
1	215779	2.5529	22.95	0.0472
2	205935	0.6872	28.61	0.0126
3	500073	5.4459	24.17	0.0423
4a	279288	23.5245	10.11	0.0815
4b	88329	0.6648	19.27	3.0786
5	232607	2.7049	21.45	0.0343
6	187896	3.4497	20.87	0.0422
7	186386	2.5252	16.76	0.0365
8	178817	1.6907	25.11	2.8479
9	152880	0.8762	21.44	0.0123
10	205163	3.2791	23.31	0.0431
11	117070	1.9687	12.75	0.0144
12	481003	11.9105	53.13	0.0729
13	75681	5.1815	2.48	0.0131
14	36896	0.4657	12.79	0.0182
15	137350	0.5832	26.01	0.0158
16	134390	0.6677	26.99	0.0129

11.5. Calibration

The ELL Methodology used by NSCB in 2005 produced small area estimates of poverty incidence. However, the municipal estimates do not conform with the regional estimates. As mentioned earlier, the researcher departed from the usual ELL Methodology by applying calibration in the estimates of the total poor, so that it conforms with the regional estimates. Table 4 shows the comparison of the number of poor households obtained using SR in CPH, with calibration and the FIES-based estimate. After applying the calibration techniques, the estimates are now almost the same except for some regions with slight difference probably due to the rounding off error in the process. The number of poor households in CPH conform with the FIES-based estimates.

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

Table 4. Regional Estimates on the Number of Poor Households Between FIES 2009 and CPH

Number of poor households from FIES 2009		Stochastic Regression with Calibration	
Region	**Number of poor households**	**Number of poor households**	**standard error**
1	179509	179509	2.1238
2	144773	144773	0.4831
3	237706	237706	2.5887
4a	247138	247138	20.8165
4b	159538	159536	1.2006
5	376044	376043	4.3729
6	348936	348936	6.4063
7	372498	372498	5.0467
8	290391	290387	2.7503
9	279138	279137	1.5998
10	298468	297579	7.7011
11	254551	254550	4.2805
12	269982	269982	6.6853
13	68163	68163	4.6668
14	66210	66209	0.8357
15	210759	210759	0.8949
16	226674	226673	1.1262

FIES estimates are subject to sampling error and may in some cases be further from the true values. FIES estimates was used to calibrate the produced estimates of SR in CPH.

Table 5 shows the comparison of estimates of the number of poor households using SR with and without calibration. It can be observed that the estimates are quite far between the two estimates. The estimates increase when calibration is applied. Since the objective of applying calibration is to get consistent estimates, the changes can be considered an improvement in the estimate. This implies further that the calibrated estimates are higher because the SR model over fit the income imputed.

Table 5. Regional Estimates on the Number of poor households with and without Calibration using Stochastic Regression

Region	Stochastic Regression without calibration		Stochastic Regression with calibration	
	Number of poor households	standard error	Number of poor households	standard error
1	215779	2.5529	179509	2.1238
2	205935	0.6872	144773	0.4831
3	500073	5.4459	237706	2.5887
4a	279288	23.5245	247138	20.8165
4b	88329	0.6648	159536	1.2006
5	232607	2.7049	376043	4.3729
6	187896	3.4497	348936	6.4063
7	186386	2.5252	372498	5.0467
8	178817	1.6907	290387	2.7503
9	152880	0.8762	279137	1.5998
10	205163	3.2791	297579	7.7011
11	117070	1.9687	254550	4.2805
12	481003	11.9105	269982	6.6853
13	75681	5.1815	68163	4.6668
14	36896	0.4657	66209	0.8357
15	137350	0.5832	210759	0.8949
16	134390	0.6677	226673	1.1262

In table 6, the estimates of poverty incidence with and without calibration were displayed using SR. The table reveals that the poverty incidence also increased when calibration was applied as well as the standard errors. As in the case of the poor households, when the income was over fitted, the poverty incidence was also affected. Thus, the calibrated values are more reliable as it conforms to the regional estimates of poverty incidence using FIES.

11.6 External validation

The wealth index is a proxy measure of the long-term standard of living of the household. It is based on household ownership of durable goods; dwelling

Table 6. Regional Estimates on Poverty Incidence With and Without Calibration using Stochastic Regression

Region	Stochastic Regression without calibration		Stochastic Regression with calibration	
	Poverty incidence (%)	standard error	Poverty incidence (%)	standard error
1	22.95	0.0472	19.09	0.0393
2	28.61	0.0126	20.12	0.0089
3	24.17	0.0423	11.49	0.0201
4a	10.11	0.0815	8.95	0.0722
4b	19.27	3.0786	26.94	0.7176
5	21.45	0.0343	34.68	0.0554
6	20.87	0.0422	38.77	0.0784
7	16.76	0.0365	33.50	0.0730
8	25.11	2.8479	34.23	0.0507
9	21.44	0.0123	39.14	0.0224
10	23.31	0.0431	33.92	0.0626
11	12.75	0.0144	27.73	0.0313
12	53.13	0.0729	29.82	0.0409
13	2.48	0.0131	2.2329	0.0118
14	12.79	0.0182	22.96	0.0327
15	26.01	0.0158	39.91	0.0243
16	26.99	0.0129	45.52	0.0218

characteristics; source of drinking water; type of sanitation facilities; and other characteristics related to the household's socioeconomic status. This is being conducted by the National Demographic and Health Survey (NDHS) every 3 years. The wealth index was constructed by assigning a weight or factor score to each household asset. These scores were summed by household, and individuals were ranked according to the total score of the household in which they reside. The sample was then divided into quintiles—five groups, each with the same number of individuals.

To be able to validate the estimates externally, the Wealth Index per region was computed by the researcher using the available data in CPH form 3. After

Table 7. Correlation Coefficient (p-values) between Wealth Index and Number of Poor Households per Region

Region	10%	20%	30%
1	0.8737 (0.000)	0.9352 (0.000)	0.9733 (0.000)
2	0.7876 (0.000)	0.8768 (0.000)	0.9431 (0.000)
3	0.4922 (0.000)	0.5477 (0.000)	0.5661 (0.000)
4a	0.7930 (0.000)	0.9332 (0.000)	0.9757 (0.000)
4b	0.6097(0.000)	0.8198 (0.000)	0.9371(0.000)
5	0.5593 (0.000)	0.7954 (0.000)	0.9303 (0.000)
6	0.1539 (0.1244)	0.5276 (0.000)	0.5905 (0.000)
7	0.5274 (0.000)	0.9067 (0.000)	0.9683 (0.000)
8	-0.0819 (0.4942)	-0.0714 (0.5510)	-0.0584 (0.6473)
9	0.8937 (0.000)	0.9723 (0.000)	0.9944 (0.000)
10	0.0647 (0.5492)	0.0755 (0.4842)	0.0617 (0.5680)
11	0.7556 (0.000)	0.9501 (0.000)	0.9762 (0.000)
12	0.4830 (0.004)	0.7255 (0.000)	0.8455 (0.000)
13	0.9689 (0.000)	0.9837 (0.000)	0.9897 (0.000)
14	0.2717 (0.0618)	0.6570 (0.000)	0.9160 (0.000)
15	0.2289 (0.0386)	0.2556 (0.0205)	0.3433 (0.0016)
16	0.5780 (0.000)	0.6449 (0.000)	0.6741 (0.000)

obtaining the wealth Index per region, the estimates were divided in quintiles. The bottom 10%, 20% and 30% of the poor from the wealth index were correlated to the estimated number of poor households in CPH. The correlation coefficients together with the significance values were shown in table 7.

The results show that the bottom 30% of the poor estimates from the wealth index are highly correlated with the poor household estimates in CPH generally except for Regions 8 and 10 wherein the p-values are greater than the significance value of 0.05. The rest of the regions have 0.00 significance values which validated the significance of the correlation between the two measures. The null hypothesis that there is no significant correlation between the wealth index and poverty incidence was rejected. Concluding a significant relationship. The same results

Table 8. Proportion of Municipalities that Fall into Each Poverty Incidence Class per Region

Region	Without Calibration				With Calibration			
	N (%)				N (%)			
	<20%	20%-25%	25%-30%	>30%	<20%	20%-25%	25%-30%	>30%
1	0 (0%)	113 (97%)	3 (3%)	1 (0.85%)	107 (91%)	9 (8%)	1 (0.85%)	0 (%)
2	0 (0%)	2 (2%)	76 (81%)	16 (17%)	43 (46%)	50 (54%)	0 (0%)	0 (0%)
3	5 (4%)	78 (61%)	44 (34%)	1 (0.78%)	128 (100%)	0 (0%)	0 (0%)	0 (0%)
4a	142 (100%)	0 (0%)	0 (0%)	0 (0%)	142 (100%)	0 (0%)	0 (0%)	0 (0%)
4b	52 (71%)	18 (25%)	2 (3%)	1 (1%)	0 (0%)	10 (14%)	61 (84%)	2 (3%)
5	12 (11%)	102 (89%)	0 (0%)	12 (11%)	0 (0%)	0 (0%)	2 (2%)	112 (98%)
6	16 (16%)	85 (85%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	101 (100%)
7	107 (100%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	9 (8%)	98 (92%)
8	0 (0%)	93 (65)	44 (31%)	6 (4%)	0 (0%)	0 (0%)	2 (1%)	141 (99%)
9	4 (6%)	68 (94%)	1 (1%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	72 (100%)
10	1 (1%)	77 (88%)	10 (11%)	0 (0%)	0 (0%)	0 (0%)	1 (1%)	87 (99%)
11	49 (100%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	7 (14%)	34 (69%)	8 (16%)
12	0 (0%)	0 (0%)	0 (0%)	49 (100%)	0 (0%)	0 (0%)	31 (63%)	18 (37%)
13	30 (100%)	0 (0%)	0 (0%)	0 (0%)	30 (100%)	0 (0%)	0 (0%)	0 (0%)
14	77 (100%)	0 (0%)	0 (0%)	0 (0%)	21 (27%)	30 (39%)	18 (23%)	8 (10%)
15	1 (0.85%)	31 (26%)	78 (67%)	7 (6%)	0 (0%)	0 (0%)	0 (0%)	117 (100%)
16	0 (0%)	8 (11%)	62 (87%)	1 (1%)	0 (0%)	0 (0%)	0 (0%)	71 (100%)

were shown in the bottom 10% and 20%. However, the correlation coefficients are higher in the bottom 30%. This implies that the CPH estimates of poor households reflects the bottom 30% of poor using wealth index. This somehow validates the estimates produced in this paper.

Table 8 shows the proportion of municipalities falling into each class of poverty incidence. The classes were categorized as <20% pi, between 20%-25%, between 25%-30% and >30% per region. It can be noticed that the pi using calibration technique concentrated in one class except for some regions as compared to the pi without calibration. Moreover, if the values will be compared to the Per Capita Poverty Threshold and Poverty Incidence among Families for 2009. The values reflect the pi computed using calibration. In example, the pi in Region 1 in 2009 is 16.8%. This reflects the pi computed using calibration technique in table 8 wherein 91% of the municipalities falls below 20% pi. Same conditions can be observed for the rest of the regions.

12. Conclusion

Small area estimates of poverty in the Philippines at municipal level were produced by combining survey data with auxiliary data derived from the census. Two models were used to predict the income in the census and the results were compared. The Deterministic Regression (DR) and Stochastic Regression (SR) were both used. In assessing the performance of the two imputation-like procedures, the SR is the model that was able to preserve the distribution of income in most (not all) of the regions in the country. A total of 14 out of 17 regions were preserved.

As mentioned, the procedure used departed from the ELL methodology applied by the NSCB in few important ways: first, the models were constructed in the first 3 replicates of FIES and validated in the last replicate. Second, the error attached to SR was estimated using non-parametric estimation (KDE). Third, the estimates of poor households were calibrated to conform with the regional estimates. Finally, the estimates were compared to the wealth index as a form of external validation.

Different models were used for each region to cater the variation of income by geographical location. The mean number of poor households derived from 1000 bootstrap estimates and their standard deviations were used as the estimates in the municipal level. The regional estimates are the aggregated values of the poor household estimates in municipalities. Standard errors produced were quite small because of the large number of bootstrap values and the number of samples per municipality since the researcher used the CPH wherein the sampling is 100%. The bottom 30% of the wealth index is highly correlated with the estimates of poor households.

Considering the results of this paper, the following conclusions were derived: The Stochastic Regression Imputation (SRI) is better to use as compared to Deterministic Regression Imputation (DRI) in attaching income to CPH. The SRI was able to preserve the distribution of 82% of the total number of regions. The DRI was able to preserve only 10.22% of the validation sets. Since the error in fitting the DRI in CPH does not follow a well-known distribution (such as the Normal distribution), the non-parametric way of estimating error was used to generate the errors attached in SRI. The technique is called Kernel Density Estimation (KDE) or the histogram method, which was found to be effective in using the SR. Using the calibration technique achieved municipal estimates that conforms to the regional estimates. The estimates of the poor households in CPH reflects the bottom 30% of the wealth index.

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. However, there is still a need in sustaining and enhancing the poverty alleviation programs in the country by targeting the poor efficiently through the use of small area estimation, especially the procedure utilized in this study.

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.

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, Dr. Rechel G. Arcilla, Mr. Frumencio Co, Mr. Marcus Jude San Pedro, De La Salle University, De La Salle University-Dasmariñas (DLSU-D), and Philippine Statistical Research and Training Institute (PSRTI).

References

Asian Development Bank (ADB). 2020. *Key Indicators for Asia and the Pacific 2020*. Manila.

Asian Development Bank (ADB). 2021. *Mapping the Spatial Distribution of Poverty Using Satellite Imagery in the Philippines*. Manila.

Elbers,C., Lanjouw, J. and Lanjouw, P. 2003. “Micro-level Estimation of Poverty and Inequality.” *Econometrica* Vol. 71 No. 1 (January, 2003), 355-364.

Cortes, D. B., & Pangan, E. T. 2007. “Imputation Procedures for partial nonresponse: The Case of 1997 family income and expenditure survey (FIES).” Undergraduate Thesis. De La Salle University.

Data 2x. 2017. “Big Data and the Well-Being of Women and Girls Application of Social Scientific Frontier.” https://www.data2x.org/wp-content/ uploads/2017/03/Big-Data-and-the-WellBeing-of-Women-and-Girls.pdf.

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

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.

Kalton, G. 1983. *Compensating for Missing Survey Data.* Inst for Social Research, Michigan.

Kish, L. 1980. *Design and Estimation for Domains*, 29(4), 209-222. Retrieved February 27, 2019, from https://www.jstor.org/stable/2987728?seq=1#page_scan_tab_contents.

Marchetti, S., Pratesi, M., Caterina, G., and Salvati, N. 2015. "Small Area Model-Based Estimators Using Big Data Sources." *Journal of Official Statistics* 31(2): 263-281.

Meinfelder, F.K. 2009. "Analysis of Incomplete Survey Data-Multiple Imputation via Bayesian Bootstrap Predictive Mean Matching.: Dissertation. Retrieved 2018 from:

https://scholar.google.com.ph/scholar?q=predictive+mean+matching&hl=en&as_sdt=0&as_vis=1&oi=scholart&sa=X&ved=0ahUKEwjgxL6h183RAhWBxpQKHU_sBDgQgQMIGDAA

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." Retrieved 2017, from https://psa.gov.ph/sites/default/files/NSCB_LocalPovertyPhilippines_0.pdf

N. Eagle, M. Macy, and R. Claxton. 2010. "Network Diversity and Economic Development." *Science* 328 (5891): 1029-1031.

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.

Philippine Statistics Authority (PSA). n.d. "Poverty statistics." Retrieved 2017, from http://nap.psa.gov.ph/products/poverty.asp

Philippine Statistics Authority (PSA). 2012. "On Poverty Threshold and Income." Retrieved 2017, from http://nap.psa.gov.ph/announce/ForTheRecord/04Apr07_se_povertygap.asp

Philippine Statistics Authority (PSA). 2015. "2010 Census of Population and Housing." Retrieved 2018, from http://www.deped.gov.ph/orders/do-21-s-2010

Philippine Statistics Authority (PSA). 2016. "Small Area Poverty Estimation Project Outline of Presentation." *Combining Survey and Census Data*. Retrieved 2018, from http://www.unescap.org/sites/default/files/Session_3Bb_Combining_survey_and_census_data_for_poverty_maps_Philippines.pdf

Philippine Statistics Authority (PSA). 2016. "2015 Full year Official Poverty Statistics of the Philippines." Retrieved 2019, from https://psa.gov.ph/sites/default/files/2015%20Full%20Year%20Official%20Poverty%20Statistics%20of%20the%20Philippines%20Publication.pdf

Philippine Statistics Authority (PSA). 2013. *National Demographics and Health Survey*. Rockfield, Maryland, USA: ICF International.

Rao, JNK. 2003. *Small Area Estimation*. Wiley Series in Survey Methodology.

Tingzon, I., Orden, A., Go, K.T., and Sy, S. 2019. Mapping Poverty in the Philippines Using Satellite Imagery, and Crowd-Sourced Geospatial Information. The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences, XLII-4/W19. https://www.researchgate.net/ publication/338131416_MAPPING_POVERTY_ IN_THE_PHILIPPINES_USING_MACHINE_ LEARNING_SATELLITE_IMAGERY_AND_CROWD-SOURCED_GEOSPATIAL_INFORMATION

U. Dorji. 2019. “Exploring Night Light as Proxy for Poverty and Income Inequality Approximation in Thailand.” 10.1109/TENCON.2019.8929247

Wealth Index. 2008. “Demographics and Health Survey Program.” Retrieved 2018, from https://www.dhsprogram.com/topics/wealth-index/Wealth-Index-Construction.cfm.

World Bank. n.d. “Poverty Overview.” Retrieved 2018, from http://www.worldbank.org/en/topic/poverty/overview