Synthesis of Small Area Poverty Models: A MICE Approach

Lara Paul A. Ebal and Zita VJ. Albacea

Institute of Statistics,
University of the Philippines Los Baños

This paper applied a methodology to synthesize the regression coefficient estimates of the regional small area poverty models across time periods by using independently collected data sets. The synthesis was linked to missing data analysis and adopted the MICE approach to consider flexibility on the nature of data sets which involve categorical variables. The MICE approach was applied using the 2009 and 2012 poverty models of Region I developed by the PSA SAE Team. Results showed that the proportion of members in the household who have at least college education (ALL_ATCOLL) has the highest positive impact followed by whether or not a household has a non-relative member who is a domestic helper (DOMESTIC_HELPER), and whether or not a barangay where the household resides is accessible to national highway (BGY_HIGHWAY) while whether or not the marital status of household head is married (HMS_MARRIED), whether or not the roof of the housing unit where the household resides is made of light materials (cogon, nipa, anahaw) without wood (ROOF_LIGHT_OLD), average family size in the barangay where the household resides (FAMSIZE), and whether or not a barangay where the household resides has at least 50 proportion of the 10 years old and overpopulation are farmers, farm laborers, fishermen, loggers and forest product gatherers (BGY_AGRI) were found to have high negative impacts on per capita household income. Moreover, the time dummy variable was found significant indicating that there was a difference between the 2009 and 2012 models in terms of their intercepts. Results of the study showed only that some predictors were consistently significant and have relatively high impacts on poverty status. It was also shown that poverty status varies across time periods.

Keywords: Small Area Poverty Models, Multiple Imputation (MI), Missing at Random (MAR), Multiple Imputation by Chained Equations (MICE)

1. Introduction

Poverty models are obtained using a small area estimation technique with the first model at the national level developed in 2000. Every three years and thereafter, when FIES are conducted, regional models were also developed. Several regional poverty models in a certain region are constructed across time periods, thus, a possibility of synthesizing these poverty models can be explored to have a deeper understanding of poverty in a region (Ebal, 2021). Regional poverty models from different periods of time may also have different sets of predictors, thus, making it difficult for researchers to identify which of the included predictors have a high or low impact on poverty through time. Synthesizing coefficients of predictors from models with different predictors is difficult to achieve since the interpretation of a coefficient depends on other predictors included in the model (Becker and Wu, 2007).

Earlier studies were conducted to synthesize models with different sets of variables by linking them to missing data analysis (Wu and Pigott, n.d.). Existing methodologies in linking the synthesis of models to missing data analysis, however, requires another assumption on the pattern of missingness, which is either monotone or arbitrary. In the same manner, existing methodologies have focused on continuous variables with the assumption of multivariate normal distribution, thus, making it difficult to synthesize model coefficients of categorical predictors.

The methodology on synthesizing model coefficients that links to missing data analysis was successful in attaining the objective of having complete data. Even with the complete data, however, the distributions of poverty across time periods may also differ. This is done by checking if there is a change in pattern of poverty across time periods where in such case has not yet been fully explored.

Given the above situations about poverty models, an appropriate approach must be performed to properly synthesize them. This paper then, aims to obtain a synthesized poverty model of Region 1 using the 2009 and 2012 small area poverty models. The final poverty model of a region will eventually give an overall picture of its poverty status through time with the estimated coefficients of the final model representing the significant increase or decrease in the per capita income based on the set predictors.

Part 2 outlines the methodology; and Part 3 discusses the findings. The last part states the conclusion.

2. Methodology

2.1. Data Structure

Since poverty models have different sets of predictors for a given time period, data structure must be investigated. Figure 1 shows the data structure of pooled data sets based on two models.

	Y	X_1	X_2	X_3	T
Year 1 $Y_{1i} = b_{11}X_{11i} + b_{12}X_{12i} + \varepsilon_{1i}$	Y_{11} Y_{12} $\vdots$ Y_{1n_1}	X_{111} X_{112} $\vdots$ X_{11n_1}	X_{121} X_{122} $\vdots$ X_{12n_1}		1 1 $\vdots$ 1
Year 2 $Y_{2i} = b_{21}X_{21i} + b_{23}X_{23i} + \varepsilon_{2i}$	Y_{21} Y_{22} $\vdots$ Y_{2n_2}	X_{211} X_{212} $\vdots$ X_{21n_2}		X_{231} X_{232} $\vdots$ X_{23n_1}	0 0 $\vdots$ 0

Figure 1. Data structure of pooled data sets with time dummy variable and following an arbitrary pattern of missingness.

The figure considers poverty models with the dependent variable as the logarithm of household per capita income Y_{ki}, where k denotes the year level and *i* represents household. The ***X****s* are predictors in the poverty model which can vary across time period. The corresponding data sets of independent sets or samples of household from these variables are then pooled. This independently pooled cross-section data gives more precise estimates since there are more observations used in the estimation (Wooldridge, 2013). Once the different set of households are pooled from different time periods given predictors that are constant, time dummy variable must also be included to consider that the population with which these data come from may have different distributions as affected by time when the data collection was implemented.

Moreover, the missing blocks in Figure 1 were also addressed in concatenated data sets to satisfy assumption of having same set of predictors to satisfy the assumption in synthesizing regression model coefficients (Becker and Wu, 2007). One approach is to use the Multiple Imputation of Chained Equations (MICE).

2.2 Multiple Imputation by Chained Equations (MICE)

MICE was applied to generate several complete data sets, and at the same time allow the creation of flexible multivariate models since normality assumption is not considered in the analysis of poverty models which includes categorical variables.

MICE involves different phases, namely: 0) Initial Phase, 1) Imputation Phase, 2) Modeling Phase, and 3) Pooling Phase. The formulas presented for each phase are adapted from Rubin (1987). The concatenated data sets shown in Figure 1 is transformed into an $n \times (p+2)$ matrix $\boldsymbol{V}$ given in Equation (1) with p as the number predictors in the concatenated data sets.

$$\boldsymbol{V} = \begin{bmatrix} v_{11}^{obs} & v_{12}^{obs} & v_{13}^{mis} & v_{14}^{obs} & \cdots & v_{1k}^{obs} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ v_{n_1 1}^{obs} & v_{n_1 2}^{obs} & v_{n_1 3}^{mis} & v_{n_1 4}^{obs} & \cdots & v_{n_1 k}^{obs} \\ \hdashline v_{\backslash 21}^{obs} & v_{22}^{obs} & v_{23}^{obs} & v_{24}^{mis} & \cdots & v_{1k}^{obs} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ v_{n_2 1}^{obs} & v_{n_2 2}^{obs} & v_{n_2 3}^{obs} & v_{n_2 4}^{mis} & \cdots & v_{n_1 k}^{obs} \end{bmatrix} = \begin{bmatrix} \boldsymbol{v_1} & \boldsymbol{v_2} & \boldsymbol{v_3} & \cdots & \boldsymbol{v_k} \end{bmatrix} \quad (1)$$

where $\boldsymbol{v_j}$ be the j^{th} column ($j=1,2,\ldots, p+2$) of the two appended vectors. The following are the steps of the different stages of the MICE approach which is from initial phase to pooling phase:

Stage 0 (Initial Phase of MICE):

Step 1. As input, use the concatenated data sets with missing observations as shown in Figure 1 which is transformed.

Step 2. Do a preliminary imputation on the missing values using the mean imputation technique. With poverty models, do this for all predictors of the models with missing observations.

Step 3. As an output is the data set with initial imputed values which will be referred as 'initial complete data set'

Stage 1 (Imputation Phase of MICE):

Step 1. As input, use 'initial complete data set' which is produced in *Stage 0.*

Step 2. Identify a predictor of the poverty model with formerly missing observation and now with initial imputed values. Call this variable $\boldsymbol{v}_j$. Based on the characteristics of $\boldsymbol{v}_j$, identify the form of the imputation model.

Step 3. Set the components (dependent and independent variables) of the imputation model for variable $\boldsymbol{v}_j$. Note that $\boldsymbol{v}_j$ is the dependent variable of the imputation model and the rest of the variables in the data set are possible independent variables including auxiliary variables and time variables. These auxiliary variables are interaction terms of the independent variables while time variables are dummy variables indicating the time period when the data were collected or observed. Auxiliary variables can be included in the imputation model to gain efficiency in the imputation analyses (Madley-Dowd et al. 2019).

Step 4. Using the identified form of the imputation model in Step 2, fit the model with the identified components of the model in Step 3 using the initial complete data set. The resulting 'best' model is the predicting imputation model for variable $\boldsymbol{v}_j$.

Step 5. Using the 'best' predicting model for $\boldsymbol{v}_j$, predict the missing values of $\boldsymbol{v}_j$ and use the predicted values to replace the imputed values of $\boldsymbol{v}_j$ in the initial complete data set. The resulting data set with predicted values of serving as its imputed values is now referred to as complete data set with $\boldsymbol{v}_j^{imp}$.

Step 6. For another predictor of the poverty model with formerly missing observation and now with initial imputed values, repeat *Steps 2* to *5* with complete data set with $\boldsymbol{v}_j^{imp}$ as input data set. This step is repeated until all predictors of the poverty model with formerly missing observations will now contain the imputed values using the identified imputation model. The resulting data set imputed values is now referred to as 'first iteration output data set'.

Step 7. The 'first iteration output data set' is now considered as input data set in *Step 1*. Repeat *Steps 2* to *6* for several iterations ($i = 2$ to k) where k is the number of iterations until convergence conditions are met. The resulting data set in the k^{th} iteration is the 'first complete data set' in this stage.

Note that the steps in this stage are repeated m times resulting m complete data sets where m is the number of times *Stage 1* is implemented indexed as h so that h goes from 1 to m.

Stage 2 (Modeling Phase of MICE):

Step 1. As input data set, use the resulting '1st complete data set' obtained in *Stage 1*.

Step 2. Fit the following full regression model with the natural logarithm of household per capita as the dependent variable and the rest of the variables in the '1st complete data set' as the predictors. The poverty model with time dummy variable is expressed as

$$\begin{aligned} \boldsymbol{Y}_1 &= \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{e}_1 \\ &= \beta_{10} + \beta_{11}\boldsymbol{x}_{11} + \beta_{12}\boldsymbol{x}_{12} + \ldots + \tau_1\boldsymbol{T}_1 + \boldsymbol{e}_1 \end{aligned} \qquad (2)$$

where $\boldsymbol{Y}_1$ is the vector of the natural logarithm of the per capita of households in the '*1^{st}* complete data set', $\boldsymbol{X}_1$ is the matrix of predictors in the '*1^{st}* complete data set', $\boldsymbol{\beta}_1$ is the vector of the regression coefficients with β_{1j} ($j=1,2,\ldots,p$) be the regression coefficient of the j^{th} predictor in the '*1^{st}* complete data set' and τ_1 is the coefficient of true dummy variable in the '*1^{st}* complete data set'. The estimated equation is then expressed as

$$\hat{\boldsymbol{Y}}_1 = \hat{\beta}_{10} + \hat{\beta}_{11}\boldsymbol{x}_{11} + \hat{\beta}_{12}\boldsymbol{x}_{12} + \ldots + \hat{\tau}_1\boldsymbol{T}_1 \qquad (3)$$

Step 3. Repeat *Steps 1* and *2* for $h = 2$ to m resulting to m sets of regression coefficient estimates and also m sets of variances of the regression coefficient estimates for each predictor. These m sets of regression coefficient estimates for p predictors, regression constant and time variable coefficient comprise an $m \times (p+2)$ matrix is referred to as matrix of preliminary estimates. Another output is $m \times (p+2)$ matrix composed of the variance estimates and will be referred to as matrix of preliminary variance estimated $V(\hat{\beta})$. The matrices $\hat{\beta}$ and $V(\hat{\beta})$ are expressed as

$$\boldsymbol{\hat{\beta}} = \begin{bmatrix} \hat{\beta}_{10} & \hat{\beta}_{11} & \cdots & \hat{\beta}_{0p} & \hat{\tau}_{1} \\ \hat{\beta}_{20} & \hat{\beta}_{21} & \cdots & \hat{\beta}_{2p} & \hat{\tau}_{2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \hat{\beta}_{m0} & \hat{\beta}_{21} & \cdots & \hat{\beta}_{mp} & \hat{\tau}_{m} \end{bmatrix} \tag{4}$$

$$\boldsymbol{V}\left(\boldsymbol{\hat{\beta}}\right) = \begin{bmatrix} v\left(\hat{\beta}_{10}\right) & v\left(\hat{\beta}_{11}\right) & \cdots & v\left(\hat{\beta}_{0p}\right) & v(\hat{\tau}_{1}) \\ v\left(\hat{\beta}_{20}\right) & v\left(\hat{\beta}_{21}\right) & \cdots & v\left(\hat{\beta}_{2p}\right) & v(\hat{\tau}_{2}) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ v\left(\hat{\beta}_{m0}\right) & v\left(\hat{\beta}_{21}\right) & \cdots & v\left(\hat{\beta}_{mp}\right) & v(\hat{\tau}_{m}) \end{bmatrix} \tag{5}$$

Stage 3 (Pooling Phase of MICE):

Step 1. As input data set, use the two matrices (4) and (5) obtained in Stage 2.

Step 2. For each column of the matrix (6), obtain the mean using the expressions below:

$$\begin{aligned} \hat{\beta}^{*} &= \begin{bmatrix} \dfrac{\sum_{h=1}^{m} \hat{\beta}_{h0}}{m} & \dfrac{\sum_{h=1}^{m} \hat{\beta}_{h1}}{m} & \cdots & \dfrac{\sum_{h=1}^{m} \hat{\beta}_{hp}}{m} & \dfrac{\sum_{h=1}^{m} \hat{\tau}_{h}}{m} \end{bmatrix} \\ &= \begin{bmatrix} \hat{\beta}_{0}^{*} & \hat{\beta}_{1}^{*} & \cdots & \hat{\beta}_{1p}^{*} & \hat{\tau}^{*} \end{bmatrix} \end{aligned} \tag{6}$$

Step 3. Formulate the synthesized poverty model using the final estimates ($\hat{\beta}_{j}^{*}$and $\hat{\tau}^{*}$) obtained in *Step 2* as stated in the following expression:

$$\begin{aligned} \boldsymbol{\hat{Y}} &= \boldsymbol{\hat{\beta}}^{*}\boldsymbol{X} + \boldsymbol{\hat{\tau}}^{*}\boldsymbol{T} \\ &= \hat{\beta}_{0}^{*} + \hat{\beta}_{1}^{*}\boldsymbol{x}_{1} + \hat{\beta}_{2}^{*}\boldsymbol{x}_{2} + \ldots + \hat{\tau}^{*}\boldsymbol{T} \end{aligned} \tag{7}$$

Step 4. Obtain the components (within-imputation variance ($\hat{U}^{*}$), between-imputation variance ($\hat{V}^{*}$) and additional source of sampling variance $\left(\dfrac{\hat{V}^{*}}{m}\right)$) of the total variance ($\hat{S}^{*}$) of the final estimates of regression constant ($\hat{\beta}_{\boldsymbol{0}}^{*}$) and regression coefficient of each predictor *j* ($\hat{\beta}_{\boldsymbol{j}}^{*}$) and time variable coefficient estimate ($\hat{\tau}^{*}$) obtained in *Step 2* as expressed by the following:

$$\begin{aligned} \hat{U}^{*} &= \begin{bmatrix} \dfrac{\sum_{h=1}^{m} v(\hat{\beta}_{h0})}{m} & \dfrac{\sum_{h=1}^{m} v(\hat{\beta}_{h1})}{m} & \cdots & \dfrac{\sum_{h=1}^{m} v(\hat{\beta}_{hp})}{m} & \dfrac{\sum_{h=1}^{m} v(\hat{\tau}_{h})}{m} \end{bmatrix} \\ &= \begin{bmatrix} \hat{U}_{0}^{*} & \hat{U}_{1}^{*} & \cdots & \hat{U}_{1p}^{*} & \hat{U}_{p+1}^{*} \end{bmatrix} \end{aligned} \tag{8}$$

$$\begin{aligned} \hat{V}^{*} &= \begin{bmatrix} \dfrac{\sum_{h=1}^{m} (\hat{\beta}_{h0} - \hat{\beta}_{0}^{*})^{2}}{m\text{ - }1} & \dfrac{\sum_{h=1}^{m} (\hat{\beta}_{h1} - \hat{\beta}_{1}^{*})^{2}}{m\text{ - }1} & \cdots & \dfrac{\sum_{h=1}^{m} (\hat{\beta}_{hp} - \hat{\beta}_{p}^{*})^{2}}{m\text{ - }1} & \dfrac{\sum_{h=1}^{m} (\hat{\tau}_{h} - \hat{\tau}^{*})^{2}}{m\text{ - }1} \end{bmatrix} \\ &= \begin{bmatrix} \hat{V}_{0}^{*} & \hat{V}_{1}^{*} & \cdots & \hat{V}_{1p}^{*} & \hat{V}_{p+1}^{*} \end{bmatrix} \end{aligned} \tag{9}$$

$$\hat{S}^* = \left[\hat{U}_0^* + \hat{V}_0^* + \frac{\hat{V}_0^*}{m} \quad \hat{U}_1^* + \hat{V}_1^* + \frac{\hat{V}_1^*}{m} \quad \ldots \quad \hat{U}_p^* + \hat{V}_p^* + \frac{\hat{V}_p^*}{m} \quad \hat{U}_{p+1}^* + \hat{V}_{p+1}^* + \frac{\hat{V}_{p+1}^*}{m} \right] \tag{10}$$

$$= \left[\hat{S}_0^* \quad \hat{S}_1^* \quad \ldots \quad \hat{S}_p^* \quad \hat{S}_{p+1}^* \right]$$

Step 5. Using the results of *Step 4*, compute the following measures to assess the synthesized poverty model obtained in *Step 3*.

$$RIV = \left[\frac{\left(\hat{V}_0^* + \frac{\hat{V}_{0j}^*}{m} \right)}{\hat{U}_0^*} \quad \frac{\left(\hat{V}_1^* + \frac{\hat{V}_1^*}{m} \right)}{\hat{U}_1^*} \quad \ldots \quad \frac{\left(\hat{V}_p^* + \frac{\hat{V}_p^*}{m} \right)}{\hat{U}_p^*} \quad \frac{\left(\hat{V}_{P+1}^* + \frac{\hat{V}_{p+1}^*}{m} \right)}{\hat{U}_{p+1}^*} \right] \tag{11}$$

$$= \left[\hat{R}_0^* \quad \hat{R}_1^* \quad \ldots \quad \hat{R}_p^* \quad \hat{R}_{p+1}^* \right]$$

$$FMI = \left[\frac{\left(\hat{V}_0^* + \frac{\hat{V}_{0j}^*}{m} \right)}{\hat{S}_0{}^*} \quad \frac{\left(\hat{V}_1^* + \frac{\hat{V}_1^*}{m} \right)}{\hat{S}_1{}^*} \quad \ldots \quad \frac{\left(\hat{V}_p^* + \frac{\hat{V}_p^*}{m} \right)}{\hat{S}_p{}^*} \quad \frac{\left(\hat{V}_{p+1}^* + \frac{\hat{V}_{p+1}^*}{m} \right)}{\hat{S}_{P+1}{}^*} \right] \tag{12}$$

$$= \left[\hat{F}_0^* \quad \hat{F}_1^* \quad \ldots \quad \hat{F}_p^* \quad \hat{F}_{p+1}^* \right]$$

$$RE = \left[\frac{1}{\left(1 + \frac{FMI_0}{m} \right)} \quad \frac{1}{\left(1 + \frac{FMI_1}{m} \right)} \quad \ldots \quad \frac{1}{\left(1 + \frac{FMI_p}{m} \right)} \quad \frac{1}{\left(1 + \frac{FMI_{p+1}}{m} \right)} \right] \tag{13}$$

$$= \left[\hat{E}_0^* \quad \hat{E}_1^* \quad \ldots \quad \hat{E}_p^* \quad \hat{E}_{p+1}^* \right]$$

It is noted that different number of imputations can be set to help improve these diagnostic measures, thus, improve the quality of the synthesized poverty model.

3. Results

3.1. Region I Small Area Poverty Models

The proposed methodology using the MICE approach was empirically applied using the Region 1 small area poverty models in 2009 and 2012 since they had at least one common predictor, that both followed the design of same master sample and have the same dependent variable – natural logarithm of household per capita income.

Table 2 shows the coefficients of the nine predictors in the 2009 Region I poverty model which are defined in Table 3. This model was obtained using 2,277 observations and has nine predictors three of which are observed at the household level and six are observed at the barangay level. These variables are said to have significant relationship with the natural logarithm of household per capita income in 2009. The estimated coefficient of -0.0579 may be interpreted as a decrease of 0.0579 in the natural logarithm of household per capita income when the household is in a barangay which is classified as urban. In terms of the adequacy of the model, it has an adjusted R2 equal to 35.16%. This implies that around 35% of the total variation in logarithmic form of per capita income is explained by the identified predictors.

Table 2. Identified predictors in the 2009 Region I poverty model.

PREDICTOR	ESTIMATED COEFFICIENT
ALL_ATCOLL	1.2851
BGY_AGRI	-0.0579
BGY_FAMSIZE	-0.1380
BGY_HIGHWAY	0.1386
BGY_PER_HHALL1524	0.6956
HMS_MARRIED	-0.2081
LAUNION	-0.0704
URBAN	0.0518
ROOF_LIGHT_OLD	-0.2821

Table 3 shows the definitions of the predictors included in the 2009 Region I poverty model with the corresponding levels of disaggregation.

Table 3. Description of identified predictors in the 2009 Region I poverty model.

PREDICTOR	DEFINITION	LEVEL OF DISAGGREGATION
ALL_ATCOLL (X_1)	Proportion of members in the household who have at least college education	Household
HMS_MARRIED (X_2)	Takes the value of one (1) if the marital status of household head is married and zero (0), otherwise	Household
ROOF_LIGHT_OLD (X_3)	Takes the value of one (1) if roof of the housing unit where the household resides is made of light materials (cogon, nipa, anahaw) without wood, and zero (0), otherwise	Household
URBAN (X_4)	Takes the value of one (1) if the barangay has a population size of 5,000, has at least one establishment with a minimum of 100 employees, or has five or more establishments with 10 to 99 employees, and five or more facilities within the two-kilometer radius from the barangay hall and zero (0), otherwise	Barangay
BGY_AGRI (X_5)	Takes the value of one (1) if the barangay where the household resides has at least 50 percent of the 10 years old and overpopulation are farmers, farm laborers, fishermen, loggers and forest product gatherers, and zero (0), otherwise	Barangay
BGY_FAMSIZE (X_6)	Average family size in the barangay where the household resides	Barangay
BGY_HIGHWAY (X_7)	Takes the value of one (1) if the barangay where the household resides is accessible to national highway, and zero (0), otherwise	Barangay
BGY_PER_HHALL1524 (X_8)	Average proportion of persons residing in the barangay aged between 15 to 24	Barangay
LAUNION (X_9)	Takes the value of one (1) if the barangay where the household resides is in La Union, and zero (0), otherwise	Barangay

Table 4 shows the 2012 Region 1 poverty model using 2,270 observations. The estimated coefficient predictor ALL_ATCOLL which was 0.0505 may be interpreted as an increase of 0.0505 in the natural logarithm of household per capita income for every unit increase in the proportion of members in the household who have at least college education. In terms of the adequacy of the model, it has an adjusted R2 equal to 43.30%. This can be interpreted that there is around 43.30% of the total variation in natural logarithmic form of per capita income that is being explained by the identified predictors.

Table 4. Identified predictors in the 2012 Region I poverty model.

PREDICTOR	ESTIMATED COEFFICIENT
INTERCEPT_	10.6057
ALL_ATCOLL	1.4392
BGY_FAMSIZE	-0.0748
BUILDING_COMMERCIAL	0.9894
DOMESTIC_HELPER	0.7994

It is noted that the predictors are of different types and of different levels of disaggregation. The predictor ALL_ATCOLL is a continuous type since its values are in proportion. Specifically, it is defined as the proportion of members in the household who have at least a college education. Each value of the proportion is measured from each household, thus, the resulting level of disaggregation which is at a household level. On the other hand, the predictor URBAN is coded to be one when a barangay was classified as urban and zero, otherwise. This implies that such predictor is a dummy or a categorical variable and at the same time at a barangay level of disaggregation since it is a barangay characteristic.

Table 5 shows the definitions of the predictors included in the 2012 Region I poverty model and the corresponding levels of disaggregation. Similar to the 2009 model, the predictors are a combination of continuous and categorical variables and a combination of household and barangay levels of disaggregation. This poverty model has four predictors which are believed to have significant relationship with the natural logarithm of household per capita income in 2012.

Table 5. Description of identified predictors in the 2012 Region I poverty model.

PREDICTOR	DESCRIPTION	LEVEL OF DISAGGREGATION
ALL_ATCOLL (X_1)	Proportion of members in the household who have at least college education	Household
BUILDING_ COMMERCIAL(X_{10})	Takes the value of one (1) if the household resides in a housing unit whose type of housing unit is classified as commercial building, or zero (0), otherwise	Household
DOMESTIC_ HELPER (X_{11})	Takes the value of one (1) if household has a non-relative member who is a domestic helper, or zero (0), otherwise	Household
BGY_FAMSIZE (X_6)	Average family size in the barangay where the household resides	Barangay

3.2 Concatenation of Data Sets

Table 6 shows the concatenated data set for Region 1 poverty models. It shows that there are variables with complete and missing observations. Predictors at household level which have missing values are HMS_MARRIED (X2), ROOF_LIGHT_OLD (X3), BUILDING_COMMERCIAL (X10), and DOMESTIC_HELPER (X11). Predictors observed at the household level in a given year are declared missing in the other year because the sampled households in 2009 may not be the same sample households in 2012. For example, the values of predictor HMS_MARRIED (X2) in 2012 were not available for sampled households in 2009 because the sampled households in 2012 may be different from those sampled in 2009. Moreover, some predictors observed at the barangay level have missing observations in either of the survey years, but values of these predictors can be generated from the census data or administrative records of the given year. For example, the values of predictor BGY_AGRI (X5) in 2012 can be generated from the census data of 2010 for sampled barangays in 2012 although the sampled barangays in 2012 were different from those sampled in 2009. This is because the statistics for all barangays are available in the census data set. Based on the above statements the data structure of the combined data set is shown in Table 6 where Y represents the natural logarithm of household per capita income.

Table 6. Variables with complete and missing observations.

Year	Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}
2009	▨	▨			▨	▨	▨	▨	▨	▨	▨	▨
2012	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	░	░

*unshaded region represents missing block of observations

There were around 50% missing values from the four predictors with incomplete observations. Specifically, variables HMS_MARRIED and ROOF_LIGHT_OLD have 49.92% missing values in 2012, and variables BUILDING_COMMERCIAL and DOMESTIC_HELPER have 50.08% missing values in 2009.

Table 7 shows the descriptive statistics of the four variables with missing values. Specifically, it shows the number of existing observations of each variable, the corresponding mean, and standard deviation. The variables HMS_MARRIED and ROOF_LIGHT_OLD which have 2,277 observations have means of 0.7536 and 0.0558, respectively. These average values imply that there were around 75 in every hundred households with married household heads and around six in every hundred housing units with light and old roofs, respectively. On the other hand, BUILDING_COMMERCIAL and DOMESTIC_HELPER have only 2,270 observations with average values of 0.0018 and 0.0150, respectively. These average values can be interpreted as in every thousand households; two were residing in commercial buildings while 15 households were with domestic helpers.

Table 7. Descriptive statistics of observed values of variables before applying MICE.

VARIABLE	OBS	MEAN	STD. DEV
HMS_MARRIED	2277	0.7536	0.4310
ROOF_LIGHT_OLD	2277	0.0558	0.2295
BUILDING_COMMERCIAL	2270	0.0018	0.0419
DOMESTIC_HELPER	2270	0.0150	0.1215

Since there were variables with missing values, then the MICE approach was implemented to impute the missing values of each variable given the appropriate imputation model. The imputation models used were binary logistic models since all variables with missing observations were dummy variables (Royston and White, 2011). The imputed data sets were generated and were compared with observed data sets using some diagnostics. The percentage distribution of the observed, and the imputed data sets of variables HMS_MARRIED, ROOF_LIGHT_OLD, BUILDING_COMMERCIAL and DOMESTIC_HELPER are shown in Figures 8, 10, 12, and 14, respectively. The imputed data sets from m = 1 up to m = 4 almost have the same percentage of zeros and ones with those in the observed values of each variable, indicating a similar distribution between the observed and imputed data sets. Moreover, the descriptive statistics of the observed and imputed values (m = 1 up to m = 4) of HMS_MARRIED, ROOF_LIGHT_OLD, BUILDING_COMMERCIAL and DOMESTIC_HELPER are presented in Figures 13, 15,17, and 19, respectively.

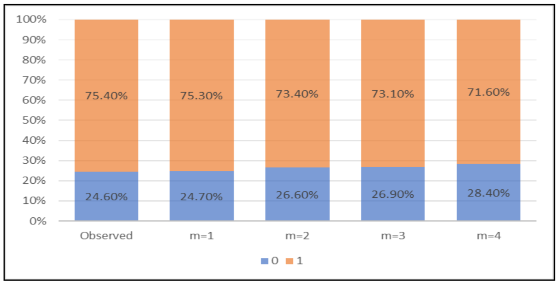

Figure 8. Percentage distribution of the observed and imputed data sets of HMS_MARRIED.

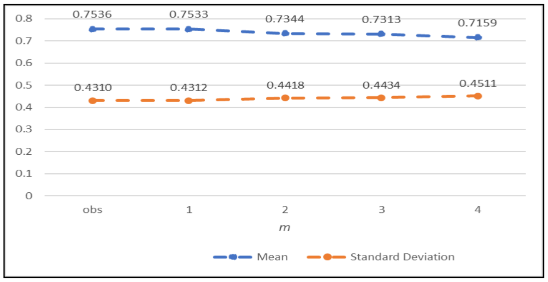

Figure 9. Descriptive statistics of imputed values by number of imputation and of observed values of MS_MARRIED

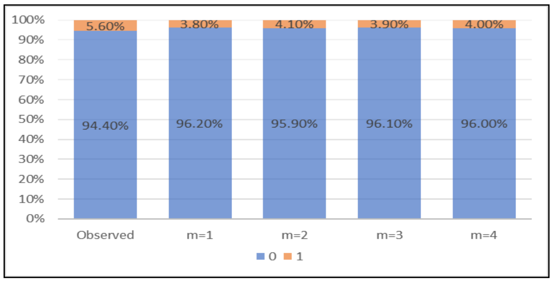

Figure 10. Percentage distribution of the observed and imputed data sets of ROOF_LIGHT_OLD.

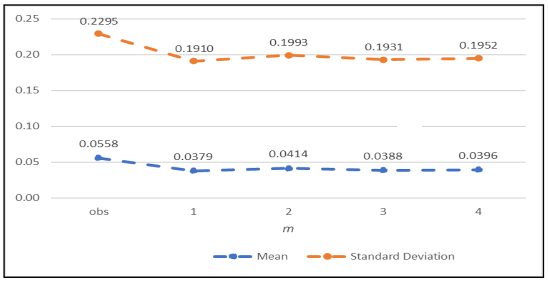

Figure 11. Descriptive statistics of imputed values by number of imputation and of observed values of ROOF_LIGHT_OLD.

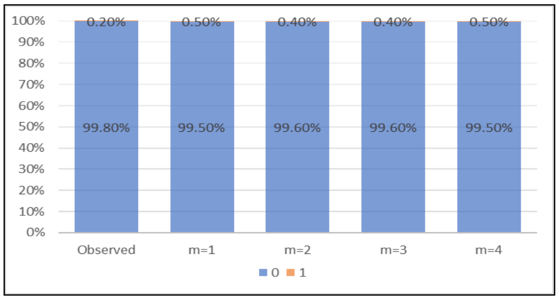

Figure 12. Percentage distribution of the observed and imputed data sets of BUILDING_COMMERCIAL

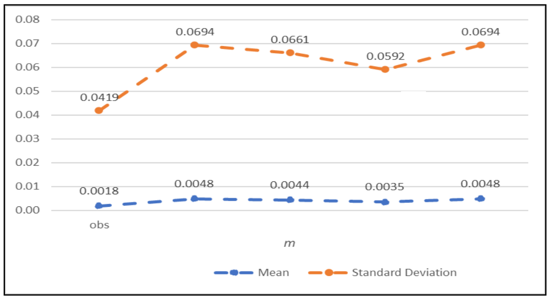

Figure 13. Descriptive statistics of imputed values by number of imputation and of observed values of BUILDING_COMMERCIAL

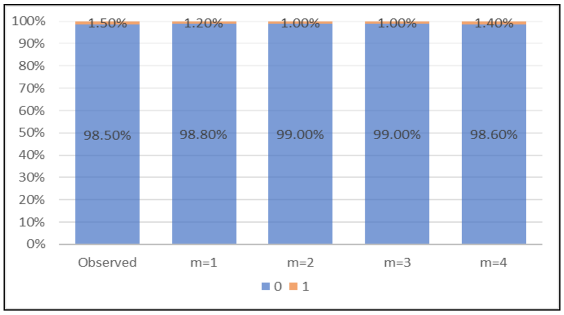

Figure 14. Percentage distribution of the observed and imputed data sets of DOMESTIC_HELPER.

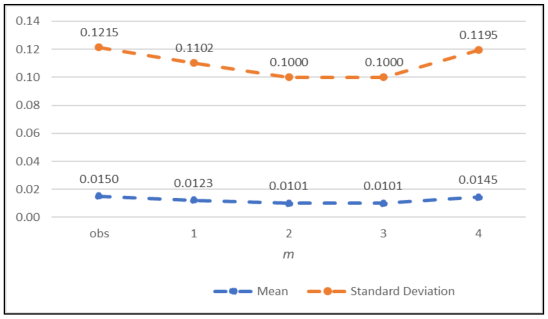

Figure 15. Descriptive statistics of imputed values by number of imputation and of observed values of DOMESTIC_HELPER.

The imputed data sets have means and standard deviations close to the mean and standard deviation of the observed values. These indicate that the behavior of the observed data set is similar with the imputed data sets. Moreover, even if there are discrepancies in the statistics between the observed and imputed data sets, it does not indicate a problem since it is also expected under MAR assumption (Nguyen et al., 2017).

3.4 Synthesized Poverty Model

Table 8 shows the synthesized poverty model after pooling all the estimates obtained from four complete data sets generated using the MICE approach. The table shows that the significant predictors at 5% level are ALL_ATCOLL, BGY_FAMSIZE, BGY_AGRI_1, BGY_HIGHWAY, HMS_MARRIED, ROOF_LIGHT_OLD, DOM_HELP and TIME. The time dummy variable is significant at 5% level which indicates that there was a difference between 2009 and 2012 models in terms of its intercept. On the other hand, the insignificant predictors at 5% level are BGY_PER_HHALL1524, LAUNION, URBAN and BLDG_COMMERCIAL.

Table 8. Synthesized poverty model using the proposed method.

PREDICTOR	STD. COEFF.	COEFF.	S.E.	95% C.I. LL	 UL
INTERCEPT	-	11.4065	0.1550	11.1025	11.7104
ALL_ATCOLL	0.2849	0.9107**	0.0497	0.8133	1.0082
BGY_FAMSIZE	-0.0855	-0.1666**	0.0287	-0.2229	-0.1102
BGY_AGRI_1	-0.0809	-0.1187**	0.0213	-0.1604	-0.0769
BGY_HIGHWAY	0.0655	0.1375**	0.0272	0.0843	0.1908
BGY_PER_HHALL1524	-0.0251	-1.1257	0.6662	-2.4551	0.2036
HMS_MARRIED	-0.1645	-0.2694**	0.0260	-0.3215	-0.2174
LAUNION	-0.0167	-0.0330	0.0294	-0.0907	0.0248
URB	0.0147	0.0321	0.0344	-0.0355	0.0998
ROOF_LIGHT_OLD	-0.1303	-0.4389**	0.0402	-0.5182	-0.3596
BUILDING_COMMERCIAL	0.0363	0.4723	0.3502	-0.2548	1.1995
DOM_HELP	0.1136	0.7073**	0.1182	0.4635	0.9511
TIME	0.0312	0.0446*	0.0210	0.0034	0.0858

* significant at α=0.05
** significant at α=0.01

Among the significant predictors, ALL_ATCOLL has the highest positive impact (0.2849) followed by DOMESTIC_HELPER (0.1136) and BGY_HIGHWAY (0.0655) while

HMS_MARRIED (-0.1645), ROOF_LIGHT_OLD (-0.1303), FAMSIZE (-0.0855), and BGY_AGRI (-0.0809) are found to have high negative impacts on per capita household income. On the other hand, these insignificant predictors turned out to have very low impacts on household per capita income.

Table 9 shows the results of the diagnostic tests done to assess the performance of the process. All predictors have very negligible values of the within, between and total variances except for predictor BGY_PER_HHALL1524. The negligible value (0.0000) of between variance indicates that the regression coefficients obtained from the four complete data sets for a certain predictor (say, BGY_FAMSIZE) did not really vary. The negligible value (0.0000) of within variance also indicates that the standard errors for a certain predictor were negligible for all four complete data sets giving an average of 0.0000. On the other hand, the predictor BGY_PER_HHALL1524 has the largest total variance (0.4592) even with complete observations. This was influenced by the high value of within variance (0.4409) caused by varying values for each complete data set.

Table 9. Model diagnostics of synthesized poverty model

	IMPUTATION VARIANCE			RVI	FMI	RE
PREDICTOR	WITHIN	BETWEEN	TOTAL			
ALL_ATCOLL	0.0025	0.0000	0.0025	0.0071	0.0071	0.9982
BGY_FAMSIZE	0.0008	0.0000	0.0008	0.0236	0.0234	0.9942
BGY_AGRI	0.0004	0.0000	0.0005	0.0282	0.0279	0.9931
BGY_HIGHWAY	0.0007	0.0000	0.0007	0.0052	0.0051	0.9987
BGY_PER_HHALL1524	0.4409	0.0147	0.4592	0.0415	0.0409	0.9899
HMS_MARRIED	0.0005	0.0001	0.0007	0.3004	0.2570	0.9396
LAUNION	0.0008	0.0000	0.0009	0.0525	0.0515	0.9873
URB	0.0011	0.0001	0.0012	0.0937	0.0901	0.9780
ROOF_LIGHT_OLD	0.0014	0.0001	0.0016	0.1227	0.1163	0.9717
BUILDING_COMMERCIAL	0.0770	0.0365	0.1226	0.5925	0.4230	0.9044
DOM_HELPER	0.0091	0.0039	0.0140	0.5401	0.3982	0.9095
TIME	0.0004	0.0000	0.0004	0.0277	0.0275	0.9932
_cons	0.0234	0.0005	0.0240	0.0271	0.0269	0.9933

4. Conclusion

The regional small area poverty models of Region in 2009 and 2012 were synthesized using a methodology which is based on MICE. The missing values in the concatenated data sets corresponding to the predictors of 2009 and 2012 models were imputed with the use of binary logistic models as appropriate imputation models. Auxiliary variables were also considered together with the number of imputations to improve the imputation process. Consequently, the imputed data sets of variables have almost the same percentage of zeros and ones with those in the observed values of each variable, indicating a similar distribution between the observed and imputed data sets. In addition, the imputed data sets have means and standard deviations close to the mean and standard deviation of the observed values, which again indicates a similar behavior the observed and imputed data sets although discrepancies do not indicate a problem since it is also expected under the MAR assumption (Nguyen et al., 2017).

The synthesized poverty model showed that ALL_ATCOLL has the highest positive impact followed by DOMESTIC_HELPER and BGY_HIGHWAY while HMS_MARRIED, ROOF_LIGHT_OLD, FAMSIZE, and BGY_AGRI were found to have high negative impacts on per capita household income. Moreover, the time dummy variable was found significant indicating that there was a difference between the 2009 and 2012 models in terms of their intercepts.

References

BECKER, B. J. and WU, M. J. 2007. Synthesis of Regression Slopes in Meta-Analysis. Statistical Science. 22(3): 414-429.ANDREWS, F. and WITHEY, S., 1976, Social Indicators of Well-being, New York: Plenum Press.

EBAL, L. P.A. 2021. Synthesis of Philippine Small Area Poverty Models. Unpublished PhD Dissertation. University of the Philippines Los Baños.

MADLEY-DOWD P., HUGHES, R., TILLING, K. and HERON, J. 2019. The proportion of missing data should not be used to guide decisions on multiple imputation. Journal of Clinical Epidemiology. 110:63-73

NGUYEN C.D., CARLIN, J.B. and LEE, K.J. 2017. Model Checking in Multiple Imputation: An Overview and Case Study. Emerging Themes in Epidemiology. 14(8): 1-12.

ROYSTON P. and WHITE, I. R. 2011. Multiple Imputation by Chained Equations (MICE): Implementation in STATA. Journal of Statistical Software. 45(4):1-20

RUBIN D.B. 1987. Multiple Imputation for Nonresponse in Surveys. John Wiley & Sons Inc., New York.STATACORP. 2019. STATA Base Reference Manual Release 16. Retrieved on September 20, 2020 from https://www.stata.com/manuals/r.pdf

WOOLDRIDGE, J. M. 2013. Introductory Econometrics: A Modern Approach. 5th Edition. Cengage Learning. Mason, Ohio, USA.

WU and PIGOTT, T. n.d. Methods for Synthesizing the Results of Regression: An Empirical Example of Applying Multiple Imputation. Loyola University Chicago. Retrieved on July 21, 2014 from http://www.campbellcollaboration.org/ artman2/ uploads/1/Methods_Pigott.pdf.