Barrios, Erniel B.; Mina, Christian D.

Working Paper

Profiling Poverty with Multivariate Adaptive Regression Splines

PIDS Discussion Paper Series, No. 2009-29

Provided in Cooperation with:
Philippine Institute for Development Studies (PIDS), Philippines

Suggested Citation: Barrios, Erniel B.; Mina, Christian D. (2009) : Profiling Poverty with Multivariate Adaptive Regression Splines, PIDS Discussion Paper Series, No. 2009-29, Philippine Institute for Development Studies (PIDS), Makati City

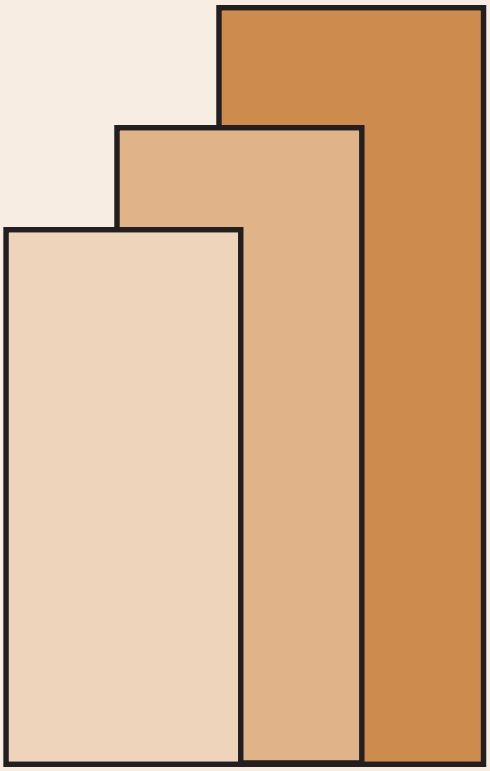
This Version is available at:
https://hdl.handle.net/10419/126787

Philippine Institute for Development Studies
Surian sa mga Pag-aaral Pangkaunlaran ng Pilipinas

Profiling Poverty with Multivariate Adaptive Regression Splines

Christian D. Mina and Erniel B. Barrios

DISCUSSION PAPER SERIES NO. 2009-29



September 2009

For comments, suggestions or further inquiries please contact:
The Research Information Staff, Philippine Institute for Development Studies
5th Floor, NEDA sa Makati Building, 106 Amorsolo Street, Legaspi Village, Makati City, Philippines
Tel Nos: (63-2) 8942584 and 8935705; Fax No: (63-2) 8939589; E-mail: publications@pids.gov.ph
Or visit our website at http://www.pids.gov.ph

Profiling Poverty with Multivariate Adaptive Regression Splines

Christian D. Mina
Research Specialist
Philippine Institute for Development Studies
NEDA sa Makati Building
106 Amorsolo St., Legaspi Village
Makati City, Philippines

Erniel B. Barrios (correspondence)
Professor, School of Statistics
University of the Philippines
Diliman, Quezon City 1101
Philippines
(ernielb@yahoo.com; ebbarrios@up.edu.ph)

Abstract

Using data from the 2003 Family Income and Expenditure Survey and 2005 Community-Based Monitoring System for a City, Multivariate Adaptive Regression Splines (MARS) is used in identifying household poverty correlates in the Philippines. Models produced by MARS are more parsimonious yet contains theoretically and empirically sound set of household poverty correlates and have high accuracy in identifying a poor household. MARS provides a better alternative to logistic regression for a more efficient and effective implementation of a proxy means test in the identification of potential beneficiaries of poverty alleviation programs.

Keywords: multivariate adaptive regression splines, logistic regression, poverty correlates, proxy means test, community-based monitoring system

1. Introduction

In many developing countries like the Philippines, government funds allocated to poverty alleviation programs are very limited. This necessitates the correct identification of program beneficiaries to ensure that funds are not leaked to those who least needed it. Various targeting mechanisms, see for example (Manasan and Cuenca, 2007), are available to help identify the program beneficiaries. The proxy means test is one of the popular tools in program beneficiaries' identification. Many studies also found this as a better alternative to means test because it is cheaper to implement and less prone to manipulation, see (World Bank, 2000) and (Balisacan, 1997). When the proxy means test was used to target the poor for a social program in Chile, it was reported to have the highest targeting rate as compared to the twenty-nine other targeted welfare programs in Latin America, (Grosh, 1994), as cited by (Ahmed and Bouis, 2002) and (Grosh and Baker, 1995).

Proxy means test generates a composite score from a set of proxy indicators of poverty commonly derived from socioeconomic surveys of households. Given the candidate socioeconomic variables that are determined by economic theory and empirical evidence, a statistical method is used to come up with a good set of poverty correlates as well as weights that are then used in generating the composite score. (Houssou et al., 2007) noted that the most commonly used statistical method for this purpose is the Multiple Linear Regression (MLR) estimated by Ordinary Least-Squares (OLS) because of its computational simplicity. MLR, however, requires many distributional assumptions that are not always satisfied in real-life data. This is particularly true when the set of predictors[1] have more categorical variables. In this case, Logistic Regression (LR) provides an alternative and was used in

[1] used interchangeably with 'independent variables' in this study

searching for non-income-based poverty determinants, e.g., (Tabunda, 2000), (Reyes, 2003), and (Reyes, 2006).

LR also have its own limitations since like any other classical regression methods, it develops the so-called 'global' model, wherein each independent variable enters the model as a whole to explain its contribution to the dependent variable, (Muñoz and Fellicisimo, 2004). This means that all of the values of the independent variables appeared to be relevant in explaining the variation of the dependent variable, which is not necessarily true in reality. Also in LR, missing data are either dropped or replaced by the mean values of the variable resulting in the deterioration of the model performance as the problem of missing data aggravates. Although there are now more efficient methods developed to address missing data like expectation maximization algorithm, computing support is not always available (Francis, 2007).

Multivariate Adaptive Regression Splines (MARS), first introduced as a data mining tool, is able to address the above limitations of LR and other classical regression methods. MARS is a nonparametric method hence, it is expected to perform as well as, or even better than, the classical regression techniques when distributional assumptions are not satisfied and it allows local and thus more accurate function approximation. MARS is not affected by any volume of missing data since it automatically introduces indicator functions for every variable that contains missing values. Furthermore, this method is designed to capture higher-order interactions, even in high-dimensional settings. But unlike other available methods that can capture complex relationships among the variables such as Classification and Regression Tree (CART) or Artificial Neural Networks (ANNs), MARS produces very simple and easy-to-interpret models, just like those of the simple regression.

This study assesses the usefulness of MARS in identifying a set of household poverty correlates in the Philippines hence, it provides a more flexible alternative to predictive modeling method in poverty analysis that are usually constrained by the parametric form of the model. Effective targeting of beneficiaries is essential in every poverty alleviation program of the government to minimize leakage and hence, optimize its alleviation impact. This is particularly crucial in the case of the Philippines where budget for such programs are limited and a large proportion of the population are poor. This paper attempts to find a tool to minimize the leakage problem in targeting through MARS. The paper further illustrates the analytical algorithm for large data sets that are high dimensional and is dominated by missing data and possibly, measurement errors.

2. Multivariate Adaptive Regression Splines

MARS is a hybrid of nonparametric methods like recursive partitioning regression and additive model. The data is left to reveal the variable knot locations while the user need not input any specification into the model. The basis functions in MARS (which serve as independent variables) are truncated linear functions, which address the problem of discontinuity of recursive partitioning algorithms. In contrast to additive models, MARS allows interactions up to an order specified by the user, and trades off the interaction order and complexity of the additive functions and interactions, see (Frank, 1995) and (De Veaux et al., 1993a) for details.

The performance of MARS varies depending on data structure, (Ture et al., 2005) but is generally known for predictive accuracy, computational speed, simplicity of interpretation, among others. (Leathwick et al., 2006) compared GAM and MARS models and noted the advantages of MARS in

cases involving large data sets. MARS models are also parsimonious and provide more extensive predictions.

(Muñoz and Fellicisimo, 2004) used two different ecological data sets to compare MARS and other modeling techniques such as LR, principal component regression and CART. One data contains 1,285 cases with lower spatial resolution at a continental scale and suffered from biased sampling. The other data have 103,181 cases with higher spatial resolution at a regional scale and was obtained through random sampling. MARS performed consistently well using the two very different data sets. Furthermore, MARS performed best in the first data set followed closely by LR models. Using the second data set, CART performed best but not significantly better than MARS, while LR models performed poorly.

Using a motor vehicle injury data consisting of 59 cases and 689 controls and with up to 3% missing values for some of the variables, (Kuhnert et al., 2000) had shown that MARS have outperformed CART and LR, in terms of accuracy and flexibility as a modeling tool. In an experimental study, (Mukkamala et al., 2004) used 5 classes of intrusion detection data sets with large differences in size and MARS is found to be superior to Support Vector Machines (SVMs), the widely used intrusion detection system, in classifying the two most important classes of intrusion detection data. Although it was not the best in other classes, accuracy of MARS was consistently high for all classes compared to existing and other alternative methods, i.e., SVMs and ANNs.

(Haughton and Loan, 2004) compared the different statistical techniques in modeling vulnerability of a panel of 4,272 households from a socioeconomic survey. MARS, together with CART and another method called Tetrad algorithm, resulted in parsimonious models and were able to capture non-

linearity and interaction effects. (Foster and Stine, 2004) argued that predictive models have this tendency of overfitting the data when interaction effects are ignored. Thus, MLR and LR models in (Haughton and Loan, 2004) included variables which lost their significance once transformed or interacted with other variables.

Certainly, MARS is a highly commended predictive modeling technique in almost all applications. Nonetheless, it also has its shortcomings. In fact, (Jin et al., 2000) argued that sample size has the largest impact on MARS, i.e., its predictive performance deteriorates when the sample size becomes small. (Briand et al., 2007) has noted that many of the existing studies suggest the use of at least 200 observations to bring the predictive advantage of MARS. Also, simulations showed that MARS is quite sensitive to outliers and strong collinearities among predictors. Similar findings were obtained by (De Veaux et al., 1993b) using chemical engineering data. MARS was found 4 times as accurate for larger data sets containing 250 and 1,000 data points than neural networks and linear regression. For smaller data set (containing 50 points), linear regression was found to perform best. The study also argued that MARS is not robust in the presence of noise and correlated predictors, which is particularly true in time-series applications.

3. Correlates of Poverty

Many studies on poverty attempted to search for a set of non-income-based poverty correlates, they are not only limited to household characteristics but extended up to barangay or community characteristics. (Balisacan, 1997) noted that location, dwelling, and family characteristics as well as ownership of durable goods can predict income well and thus, are considered ‘promising’ proxy indicators of income. Furthermore, many studies contributed in the identification of proximate

indicators of poverty, e.g., (Lipton and Ravallion, 1995), (Balisacan, 1992, 1994), (World Bank, 1990), (Herrin and Racelis, 1994), (Marquez and Virola, 1997) had identified certain demographic and occupational attributes of household members, such as sex of household head and household composition. For example, male-headed households tend to have higher standards of living than female-headed ones, holding other things constant. All other things remaining the same, households with higher number of young children are more likely to have lower standards of living. Meanwhile, households engaged in agriculture are known to have lower standards of living than those in other sectors like industry and services.

(ILO, 1974), (Balisacan, 1993), (Bautista and Lamberte, 1996), (HDN and UNDP, 1997) argued that households located in major urban centers have generally higher standards of living than those located in rural areas because of the large differences between their access to infrastructure and basic social services. Aside from urbanity, other location attributes such as regions and provinces also have correlation with income or living standards of households.

In attempt to identify policy variables that can be considered in the formulation of poverty reduction strategies for the Philippines, (Reyes, 2003) estimated six LR models for the different rounds of the FIES (i.e., 1985, 1988, 1991, 1994, 1997, and 2000). With poverty status as the dependent variable, only three factors in all six models were included, namely: (1) highest educational attainment of household head; (2) family size, and; (3) proportion of income derived from agriculture. The results indicate that the probability of being non-poor increases as the highest educational attainment of household head increases. However, for the same level of educational attainment of household head, the probability of being non-poor of a larger household tend to be relatively lower compared to those with fewer members. Meanwhile, the probability of being non-poor decreases as the share of

agricultural income increases. On the other hand, for the same proportion of agricultural income to total household income, the probability of being non-poor of a household decreases with family size.

Other studies are not mainly focused on identification of poverty correlates but were able to establish a link between income, or measure of household welfare, and a possible proxy indicator. For instance, (Orbeta, 2006) attempted to document the relationship between family size, poverty and vulnerability to poverty. There are strong associations between larger family size, poverty incidence and vulnerability to poverty.

4. Analytical Framework

Two different data sets were used to be able to assess the effects of various factors like data collection method, geographical scale, volume of missing and quality of data on model performance. The first data is the 2003 Family Income and Expenditures Survey, a nationwide survey of households (with lesser volume of missing data and higher level of data quality) and is the basis for the computation of official poverty statistics in the Philippines. The 2005 Community-Based Monitoring System (CBMS) data for Pasay City was also used. This is a citywide complete enumeration of households hence expected to include higher volume of missing and inferior quality data.

Using the 2003 Master Sample (MS) which is based on data from the 2000 Census of Population and Housing (CPH), a multi-stage sampling was employed in selecting the sample households with region as the sampling domain. The primary sampling units (PSUs) were selected within a set of strata using probability proportional to estimated size (PPES) sampling, where the measure of size was the number of households in the PSU and based on the 2000 CPH. The primary groupings or sampling

domains[2] were the 17 regions of the country in which the sample was allocated. Within each region, further stratification was performed using geographic groupings such as provinces and highly urbanized independent cities. Within each of these groups formed in a region, further stratification was done using proportions of strong houses and of households in agriculture in the PSUs and a measure of per capita income as stratification factors. Housing units were considered in the final stage of sampling instead of household size since the former are fixed and no resources are available to track down moving households. All households within the selected housing units were enumerated, except in few cases when a housing unit has more than 3 households. In such a case, a sample of 3 households was selected with equal probability. 2003 FIES contains a total of 42,094 sample households.

CBMS is an organized process of data collection and processing at the local level and of integration of data in local planning, program implementation and impact-monitoring. Its intended users are the local government units (LGUs), national government agencies, non-government organizations (NGOs), and civil society. It also serves as a local poverty monitoring system, (Reyes, 2006). Census of households is conducted regularly by each of the cities, to make sure that their databanks can provide regular source of baseline information for various activities, which include the following: preparation of development profiles; design, targeting and impact monitoring of social services and development programs, and; poverty mapping. The original plan for its implementation is every year. But since data encoding, validation and processing may take more than a year, census will now be conducted every three years. In addition, it will also now be term-based. CBMS can generate a wide range of LGU-specific indicators, but at the minimum, the database contains the core set of 14 local indicators, which: (i) capture the multidimensional aspects of poverty; (ii) have been confined to

[2] Domain is a part of the population for which separate estimates are planned in the sample design (NSO, 2003).

output and impact indicators, and; (iii) are being measured to determine the welfare status of the population. (See Table 1 for the list of core indicators) (PEP-CBMS Network Coordinating Team, 2008) The census had been conducted from March to December 2005 and the City Government was able to collect all the required data from all the 201 barangays in the city, which comprised of a total of 65,108 households. (Diaz and Cariño, 2007)

Table 1. Core indicators of CBMS

Basic Needs	Indicator	
Health	1	Proportion of child deaths aged 0-5 years old
	2	Proportion of women deaths due to pregnancy-related causes
Nutrition	3	Proportion of malnourished children aged 0-5 years old
Shelter	4	Proportion of households living in makeshift housing
	5	Proportion of households classified as squatters/informal settlers
Water and Sanitation	6	Proportion of households without access to safe water supply
	7	Proportion of households without access to sanitary toilet facilities
Basic Education	8	Proportion of children 6-12 years old not in elementary school
	9	Proportion of children 13-16 years old not in secondary school
Income	10	Proportion of households with income below poverty threshold
	11	Proportion of households with income below subsistence threshold
	12	Proportion of households which experienced food shortage
Employment	13	Proportion of persons who are unemployed
Peace and Order	14	Proportion of persons who were victims of crime

Different ‘versions’ of the data were used. The ‘original’ versions are those without any modifications. ‘Non-missing data’ versions, where all observations with missing values for any of the variables were deleted, were also considered. To determine if the performance of the model can be affected by size of data, 10% of the data set were also extracted and used as a different set. Furthermore, both the full and 10% data sets were divided into two sub-samples, namely: 60% training sample (which is used to identify the model) and 40% validation sample (where the model is validated for its predictive accuracy). Table 2 shows these different data sets.

Table 2. Data sets used and their number of observations

Data set	No. of observations
2003 FIES	
'Original'	
Full (100%)	42,094
Training (60%) / Testing (40%)	25,256 / 16,838
Subset (10%)	4,209
Training (6%) / Testing (4%)	2,525 / 1,684
'Non-missing'	
Full (100%)	36,578
Training (60%) / Testing (40%)	21,947 / 14,631
Subset (10%)	3,658
Training (6%) / Testing (4%)	2,195 / 1,463
2005 CBMS Pasay City	
'Original'	
Full (100%)	65,108
Training (60%) / Testing (40%)	39,066 / 26,042
Subset (10%)	6,511
Training (6%) / Testing (4%)	3,907 / 2,604
'Non-missing'	
Full (100%)	64,027
Training (60%) / Testing (40%)	38,416 / 25,611
Subset (10%)	6,403
Training (6%) / Testing (4%)	3,842 / 2,561

The dependent variable in this study is poverty status based on income. It is derived by comparing the per capita income of a household with the official poverty threshold estimated by the National Statistical Coordination Board (NSCB). This is the per capita income necessary to meet basic food and non-food needs of each household for the province where the household resides. Per capita income is computed as the total income reported by a household divided by the total number of

members in that household. If per capita income is below this threshold, a household is considered poor. Otherwise, a household is tagged as non-poor, (Reyes, 2003).

Many studies have already established the correlation between income and other household socioeconomic indicators such as household head profile, household composition, ownership of assets, access to basic amenities, and housing structure and tenure, see for example (Reyes, 2003, 2006). From these literatures, Table 3 summarizes the list of variables and their definitions.

Table 3. Descriptions of the variables

Variable	Description
Dependent:	
hpov_p	household poverty status based on income (poor = 0; non-poor = 1)
Independent:	
age	age of household head (in years)
sex	sex of household head (female = 0; male = 1)
educ	highest educational attainment of household head (0 = no grade completed; 1 = elementary undergraduate; 2 = elementary graduate; 3 = high school undergraduate; 4 = high school graduate; 5 = college undergraduate; 6 = at least college graduate; 7 = post-graduate)
hnagri	kind of business of household head (0 = engaged in agriculture; 1 = not engaged in agriculture)
hsize	household size (or, total number of members in a household)*
ofw	presence of an overseas Filipino worker (OFW) in a household (0 = no; 1 = yes)**
hmkshft	whether or not a household is living in makeshift housing (0 = living; 1 = not living)
hsquat	whether or not a household is living in squatter's area (0 = living; 1 = not living)
helec	access to electricity (0 = no access; 1 = with access)
hwater_o	main source of water supply (1 = rain; 2 = spring, river, stream, etc.; 3 = peddler; 4 = dug well; 5 = shared, tubed/piped well; 6 = shared, faucet, community water system; 7 = own use, tubed/piped well; 8 = own use, faucet, community water system)
htoilet_o	type of toilet facility (0 = none; 1 = others (pail system, etc.); 2 = open pit; 3 = closed pit; 4 = water-sealed)
hwtv	ownership of television set (0 = no; 1 = yes)
hwvhs	ownership of VHS/VTR/VCD/DVD player (0 = no; 1 = yes)
hwref	ownership of refrigerator/freezer (0 = no; 1 = yes)
hwwash	ownership of washing machine (0 = no; 1 = yes)
hwaircon	ownership of airconditioner (0 = no; 1 = yes)
hwcar	ownership of vehicle (0 = no; 1 = yes)
hwphone	ownership of telephone/cellphone (0 = no; 1 = yes)
hwcomputer	ownership of computer (0 = no; 1 = yes)
hwoven	ownership of microwave oven (0 = no; 1 = yes)
hurb***	area of residence of household (0 = rural; 1 = urban)

* average for the two rounds of survey

** in FIES, an OFW is assumed to be present in a household if that household receives cash from household members who are contract workers or working abroad

*** not included in 2005 CBMS Pasay City models

Table 4 summarizes the hypothesized relationships of household poverty status with the set of predictors based on the literature.

Table 4. Hypothesized relationships of poverty status with independent variables

Hypothesized relationship	Variable
positive (+)	Age of household head Highest educational attainment of household head Kind of business or work of household head OFW indicator Living/not living in makeshift housing Living/not living as informal settler/squatter Access to electricity Type of toilet facility Main source of water supply Ownership of assets: television set, VHS/VTR/VCD/DVD player, refrigerator/freezer, washing machine, airconditioner, vehicle, telephone/cellphone, computer, microwave oven Urbanity, or area of residence
negative (-)	Household size
either positive or negative (+/-)	Sex of household head

Two classes of models are estimated in this study, namely: LR and MARS. The LR model is given by the following. Let Y be the random binary variable, the probability $P(Y=1)$ is given by:

$$P(Y=1) = p = \frac{e^{\beta X}}{1+e^{\beta X}}, \qquad (1)$$

where: β = vector of coefficients, and;

X = vector of independent variables.

The above equation represents what is known as the (cumulative) *logistic distribution function*. If *P(Y=1)*, the probability that an event occurs, is given by (1), then, *1–P(Y=1)*, the probability that an event does not occur, is:

$$1 - P(Y = 1) = 1 - p = \frac{1}{1 + e^{\beta X}} \tag{2}$$

Therefore, equations (1) and (2) can be written as:

$$\frac{P(Y = 1)}{1 - P(Y = 1)} = \frac{p}{1 - p} = \frac{e^{\beta X}}{1 + e^{\beta X}} \bullet \frac{1 + e^{\beta X}}{1} = e^{\beta X} \tag{3}$$

Equation (3) is simply the *odds ratio* in favor of an event occurring – the ratio of the probability that an event will occur to the probability that it does not occur.

MARS, on the other hand, is a nonparametric method for fitting adaptive regression that uses piecewise basis functions to define relationships between a dependent variable and a set of predictors. The following is the function approximation of MARS:

$$\hat{f}(\mathbf{x}) = a_0 + \sum_{m=1}^{M} a_m B_m^{(q)}(\mathbf{x}), \tag{4}$$

such that:

a_0 = coefficient of the constant basis function, or the constant term;

$\{a_m\}_1^M$ = vector of coefficients of the non-constant basis functions, $m = 1, 2, \ldots, M$;

$B_m^{(q)}(\mathbf{x}) = \prod_{k=1}^{K_m} \left[s_{km} \cdot \left(x_{v(k,m)} - t_{km} \right) \right]_+^q$ = vector of non-constant (truncated) basis functions, or the tensor product spline basis;

where:

m = number of non-constant basis functions (1, 2, …, M);

q = the power to which the spline is raised in order to control the degree of smoothness of the resultant function estimate, which in this case is equal to 1;

+ = denotes that only positive results of the right-hand side of the equation are considered; otherwise, the functions evaluate to 0. Thus, the term truncated;

s_{km} = indicates the (left/right) sense of truncation, which assumes only 2 values, i.e., ±1, representing the standard basis function and its mirror image. For s_{km} equal to +1, the basis function will have a value x-t if $x>t$ and 0 if $x \leq t$. If it is -1, the basis function will have a value t-x when $x<t$, while 0 if $x \geq t$;

$x_{v(k,m)}$ = value of the predictor;

$v(k,m)$ = label of the predictor ($1 \leq v(k,m) \leq n$);

n = number of predictors;

t_{km} = "knot" location on the corresponding predictor space or region, or value that defines an inflection point along the range of the predictor;

k = maximum level or order of interaction, or the number of factors, in the mth basis function ($1, 2, \ldots, K_m$).

Estimation of the parameters in MARS is done by the Penalized Least Squares (PLS) with the following objective function:

$$Obj.\ P(x) = \min \sum \left(y_i - \hat{f}(x_i) \right)^2 + \lambda \int f''(x_i) dx_i \qquad (5)$$

where: first term = residual sum of squares;

second term = roughness penalty term, which is weighted by λ (which is the smoothing constant).

The penalty term is large when the integrated second derivative of the regression function $f''(x)$ is large – that is, when *f (x)* is 'rough' (with rapidly changing slope). At one extreme, when the λ is set to zero (and if all the values of *x* are distinct), the objective function simply interpolates the data. At the other extreme, if λ is very large, then the objective function will be selected so that its second derivative is everywhere zero, implying a globally linear least-squares fit to the data, (Fox, 2002).

One of the most useful applications of variable nesting in MARS is in dealing with missing values among the independent variables. MARS creates two basis functions for any variable with missing data, one for the presence of missing values and one for the absence, (Francis, 2007). However, MARS does not consider interactions with missing value indicators to be genuine interactions. Thus, if MARS is directed to generate only an additive model, it may still contain interactions involving these missing value indicators, (Salford Systems, 2001).

Knot selection proceeds until some maximum model size is reached, which is usually user-specified. After overfitting the model with so many basis functions, a backwards-pruning procedure is applied in which those basis functions that contribute least to model fit are progressively removed. At this stage, a predictor variable can be dropped from the model completely if none of its basis functions contribute meaningfully to predictive performance. The sequence of models generated from this process is then evaluated using the so-called Generalized Cross-Validation (GCV), and the model with the best predictive fit is finally selected. (Friedman, 1991) proposed an approximation to the cross-validation criterion that requires only one evaluation of the model, which is a modification of the GCV criterion proposed by (Craven and Wahba, 1979), for use in conjunction with linear fitting methods.

$$GCV(M) = \frac{\frac{1}{N}\sum_{i=1}^{N}\left[y_i - \hat{f}_M(x_i)\right]^2}{\left[1 - \frac{C(M)}{N}\right]^2}, \quad (6)$$

where: numerator = lack-of-fit on the training data (bias)

denominator = (inverse) penalty for increasing model complexity *C(M)* (sort of "variance")

MARS minimizes *GCV(M)*, which reduces the bias of the model estimates but at the same time increases the variance due to additional parameters included to improve the fit of the model.

Accuracy Measures

(Sharma, 1996) stressed that classification of the same data used in model estimation is biased since it only reflects model fit and not necessarily its predictive ability. The optimal strategy is to develop models using the training samples then applied models to testing samples to evaluate the predictive performance of the models.

(Houssou et al., 2007) used a number of measures proposed by (IRIS, 2005) in assessing the predictive performance of poverty models developed using different estimation methods. The measures are the following:

(1) *total accuracy* or the percentage of the total households whose poverty status is correctly predicted by the model;

(2) *specificity*, which is defined as the percentage of $y = 0$ observations (which in this case are poor households) that are correctly predicted, or the *poverty accuracy*;

(3) *sensitivity*, which is defined as the percentage of $y = 1$ observations (which in this case are non-poor households) that are correctly predicted, or the *non-poverty accuracy*;

(4) *undercoverage rate* is defined as the ratio of the number of poor households that are erroneously predicted as non-poor to the total number of poor households, and is related to *exclusion error* (also called *false negative rate*);

(5) *leakage rate* is defined as the ratio of the number of non-poor households that are erroneously predicted as poor to the total number of households that are predicted as poor, or the *inclusion error* (also called *false positive rate*), and;

(6) *Balanced Poverty Accuracy Criterion (BPAC)*, which is defined as poverty accuracy less the absolute difference between undercoverage and leakage, each expressed as a percentage of the total number of poor households.

Note that the leakage rate used in this study is based on the definition used by (Reyes, 2006) and (Manasan and Cuenca, 2007), and is different from the one used in the computation of BPAC in the study of (Houssou et al., 2007). Instead of using total number of households that are predicted as poor, the latter used total number of actual poor households as the denominator.

5. Results and Discussion

The FIES data illustrates some of the theory and empirical insights on the profile of poverty in the literature. Figure 1 exhibits the distribution of poor and non-poor households by size. Non-poor households are generally smaller compared to poor households. In Figure 2, the head in a poor household tends to have lower educational attainment than those of the non-poor. Poor households have relatively higher percentage of male heads while non-poor households have relatively higher

percentage of female heads. Some 70 percent of poor household heads are engaged in agriculture. In contrast, there are only 30 percent of non-poor households whose heads are in the agriculture sector.

Figure 1. Proportion of households by household size, FIES data

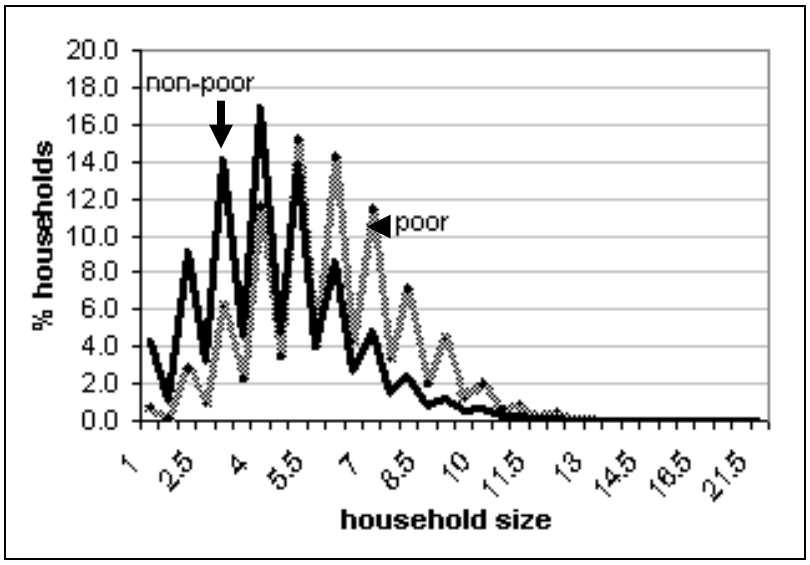
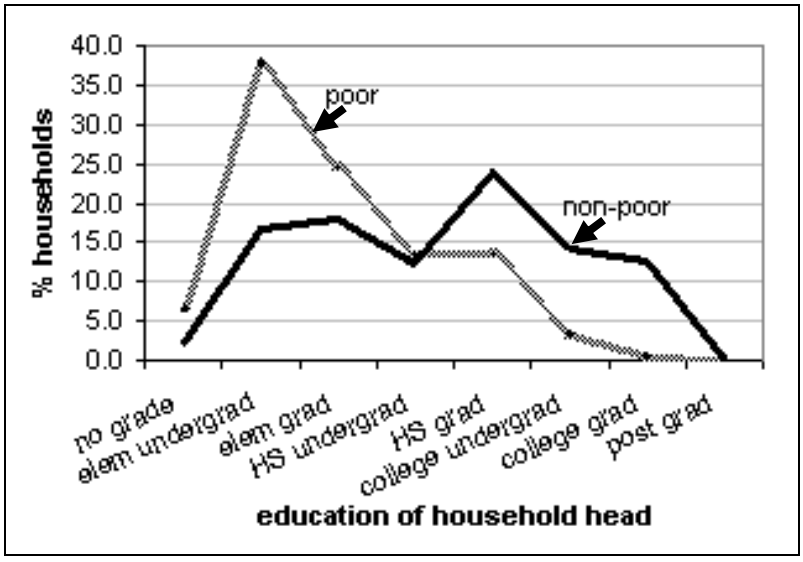


Figure 2. Proportion of households by education of household head, FIES data



CBMS data, featuring an urban center, exhibit a different scenario from the one reflected in FIES. In Figure 3, although there are still poor households from the third up to the tenth decile, almost 93 percent belong only to the first two deciles. In terms of magnitude, poor households only dominate in the first decile. The non-poor households are present only from the second up to the tenth decile and

none belong to the first decile. In Figure 4, the distributions of poor and non-poor households by household size are also skewed to the right, just like in FIES. The average household size among poor households is 5 while 4 for non-poor households. Unlike in FIES, Figure 5 reveals that for education of household head, the non-poor group starts to outweigh the poor only at the college undergraduate level. More than 75 percent of heads in poor households did not reach tertiary level of education.

Figure 3. Proportion of households by income decile, CBMS data

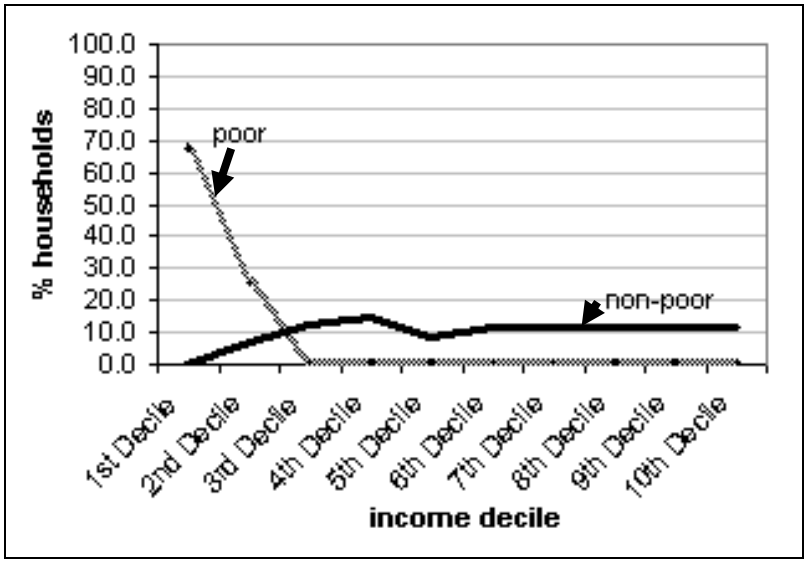


Figure 4. Proportion of households by household size, CBMS data

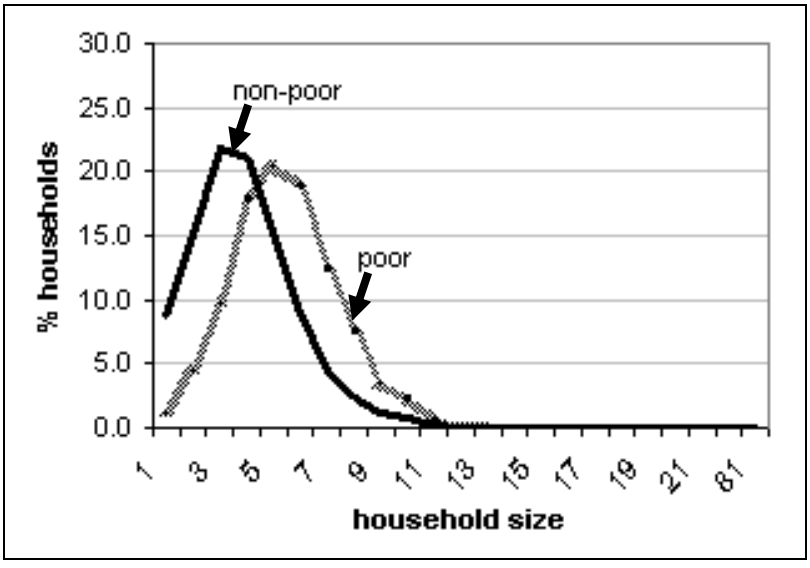
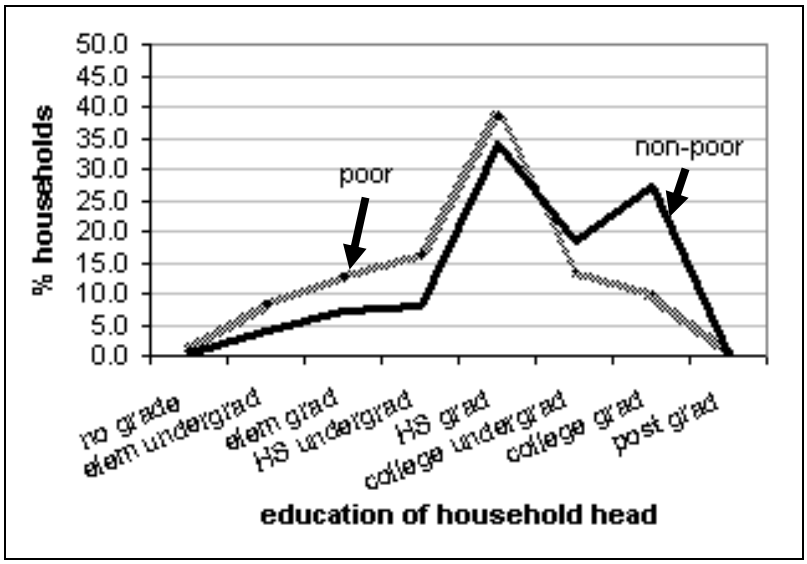


Figure 5. Proportion of households by education of household head, CBMS data



5.1. Logistic Regression Models

After different versions of the FIES dataset were explored, the final results are summarized in Table 5. The variables from columns 2 to 8 are those that are not dropped in any of the models under different versions of the data set. Column 9 presents the total number of times the variable appeared in all of the LR models.

Table 5. Significant variables in LR models for 2003 FIES data

Variable	'Original'	'Non-missing'	Full (100%)	Training (60%)	Subset (10%)	Subset - Training (6%)	Overall	No. of times appeared in the model
age	x	x	x	x	x	x	x	8
sex	x	x	x	x	x	x	x	8
educ	x	x	x	x	x	x	x	8
hnagri	x			x	x	x		7
hsize						x		3
ofw	x	x	x	x	x	x	x	8
hmkshft				x	x			4
hsquat								1
helec		x		x	x			6
hwater_o		x	x	x		x		7
htoilet_o	x			x	x	x		7
hwtv					x			4
hwvhs	x	x	x	x	x	x	x	8
hwref	x	x	x	x	x	x	x	8
hwwash	x	x	x	x	x	x	x	8
hwaircon		x	x	x				6
hwcar	x	x	x	x	x	x	x	8
hwphone	x	x	x	x	x	x	x	8
hwcomputer			x	x				4
hwoven			x	x				4
hurb			x					4
No. of variables	11	12	14	17	14	13	9	

There are more predictors in the LR models using the non-missing versions of the FIES data set. All of the variables included in the non-missing versions of the data set are not included in the original versions. Similarly, those present in the original versions are no longer included in the non-missing versions. As the size of the data becomes larger, more complicated structure is expected as a result of possible increase in the extent of its heterogeneity. Thus, a more parsimonious model is generally produced from smaller data sets. Since FIES is from a probability sample of households, inclusion of few variables in the model may be enough to explain the variation in poverty status of households. This is particularly true when models are estimated using smaller data sets.

A total of nine predictors were not able to lose their discriminatory power in any of the LR model estimations. The hypothesized relationships with income of significant predictors are confirmed in the parameter estimates in Table 6.

Table 6. Most 'present' predictors in all LR models for 2003 FIES

Variable	Sign	Significance
age	+	always significant
educ	+	always significant
sex	-/+	mostly insignificant
ofw	+	sometimes insignificant
hwvhs	+	always significant
hwref	+	always significant
hwwash	+	always significant
hwcar	+	mostly insignificant
hwphone	+	always significant

Households with heads belonging to the higher age group and who have higher educational attainment tend to have higher income and thus, more likely to be non-poor. The result is consistent with the findings of (Reyes, 2003) that education of household head has positive correlation with income. On the other hand, sex of household head has either negative or positive correlation with income. The more common relationship of 'sex' with income and poverty status is negative, although it is mostly not significant in all models. Households who own assets such as VHS/VTR/VCD/DVD player, refrigerator/freezer, washing machine, vehicle, and telephone/cellphone tend to have higher income and more likely to be non-poor. This further agrees with previous studies on poverty correlates, e.g., (Reyes, 2006), (Tabunda, 2000), and (Balisacan, 1997). While having a member who is an OFW usually facilitates households' acquisition of assets, but since there are only few households who have OFW members, this turned to be not sufficient predictor of poverty.

On the contrary, predictors that seldom enter the LR models are the following: ‘hsize’, ‘hmkshft’, ‘hsquat’, ‘hwtv’, ‘hwcomputer’, ‘hwoven’, and ‘hurb’. Among these variables, the most common ones that result to the rejection of the specification test for the estimated models are ‘hsize’ and ‘hwtv’. One possible explanation for this is that the effects of these variables become irrelevant when added with the combined contribution of a number of variables in the model. Apparently, their effects can already be explained by those of the other variables. On the other hand, ‘hmkshft’, ‘hsquat’, and ‘hurb’ are mostly insignificant and have reversed signs. Almost same percentage of poor and non-poor households live in makeshift housing and as informal settlers. The two ownership variables, i.e., ‘hwcomputer’ and ‘hwoven’, are often not significant because the number of poor households who own these assets substantially decreased when a smaller subset of data are randomly used for LR estimation.

The results of LR estimations for CBMS data are summarized in Table 7 where significant variables in all of the estimated models using different versions of the data set are presented. Column 9 presents the total number of times the variable appeared in all of the models. The non-missing versions of the data included relatively more predictors than the original versions. After dropping all missing data, variables like ‘age’, ‘hsquat’, and ‘helec’ turned out to be significant predictors of poverty. Smaller data sets tend to include more predictors to be able to predict well the variation in poverty status. The effect of decreasing sample size may have resulted in inclusion of some of the variables such as ‘educ’, ‘hsize’, ‘hwtv’, and ‘hwwash’, although their contributions are not significant.

Table 7. Significant variables in LR models for 2005 CBMS data

Variable	'Original'	'Non-missing'	Full (100%)	Training (60%)	Subset (10%)	Subset - Training (6%)	Overall	No. of times appeared in the model
age		x	x	x	x			7
sex	x	x	x	x	x	x	x	8
educ						x		3
hnagri								0
hsize					x	x		4
ofw	x	x	x	x	x	x	x	8
hmkshft	x	x	x	x	x	x	x	8
hsquat		x	x	x	x			7
helec		x	x	x		x		7
hwater_o								2
htoilet_o	x	x	x	x	x	x	x	8
hwtv					x			3
hwvhs	x	x	x	x	x	x	x	8
hwref	x	x	x	x	x	x	x	8
hwwash					x	x		4
hwaircon	x	x	x	x	x	x	x	8
hwcar	x	x	x	x	x	x	x	8
hwphone	x	x	x	x	x	x	x	8
hwcomputer	x	x	x	x	x	x	x	8
hwoven	x	x	x	x	x	x	x	8
No. of variables	11	14	14	14	16	15	11	

The variables that led to rejection of the specification test are ‘educ’ and hsize’. It seems that their contributions in discriminating between poor and non-poor households become irrelevant when combined with all other variables in the model. The variables ‘hwtv’, ‘hwwash’, and ‘hwater_o’ yield coefficients whose signs are reversed from the hypothesized relationship with income and poverty status. It is possible that the percentages of poor and non-poor households that own television set are almost equal. Thus, when different data sets are used, non-poor households are outnumbered by the poor. This is not surprising for urban setting, where ownership of television is not really uncommon even for poor households. The variable ‘hwwash’ is not significant because almost all households own a washing machine. Meanwhile, incorrect sign of ‘hwater_o’ is due to a very high percentage of non-poor households having ‘peddler’ as their main source of water supply.

The variables that are present in all LR models for CBMS data set are shown in Table 8. All predictors have positive relationship with income and poverty status, particularly the ownership variables, housing materials (i.e., makeshift or not), and access to basic amenities (i.e., type of toilet facility), which are all consistent with the findings of (Reyes, 2006). However, ownership of refrigerator/freezer, airconditioner, vehicle, and microwave oven are not always significant, particularly when smaller data sets are used. For small data sets, ownership of these assets becomes rare hence, empirical relationship is difficult to establish in the model.

On the indicator 'ofw', this is consistent with the findings of (Tabuga, 2007) that OFW remittances are usually received by non-poor households. It implies that there are relatively more non-poor households which have at least one OFW among their members. Meanwhile, the percentage of non-poor households with female heads is relatively higher than that of poor households.

Table 8. Common predictors in LR models for CBMS data

Variable	Sign	Significance
sex	-	sometimes insignificant
ofw	+	always significant
hmkshft	+	always significant
htoilet_o	+	always significant
hwvhs	+	always significant
hwref	+	sometimes insignificant
hwaircon	+	sometimes insignificant
hwcar	+	sometimes insignificant
hwphone	+	always significant
hwcomputer	+	always significant
hwoven	+	sometimes insignificant

5.2. Multivariate Adaptive Regression Splines

The variables that are always significant in the MARS models are summarized in Table 9. The number of predictors included in the MARS models estimated using the non-missing versions of the data is relatively higher than that in models estimated using the original versions. Similarly, models estimated using smaller data sets have more variables than those estimated using the larger data sets. There is only one variable that is significant for different interaction orders when penalty is incorporated. Without penalty, six variables are present in all models.

Regardless of penalty values in MARS modeling, household size is the only predictor that is present in all models. This is consistent with (Orbeta, 2006) who also concluded that bigger household size is associated with higher poverty and vulnerability to poverty. As noted by (Orbeta, 2006), this strong link between household size and poverty has been continuously observed over twenty-five years for which FIES data is available. There are other significant variables present in most of the models, e.g., ‘hwtv’, ‘hwref’, ‘hnagri’, and ‘hwphone’. Although ownership of television set as well as refrigerator/freezer and telephone/cellphone has been increasing in recent times in urban areas, the percentage of poor households owning these assets especially in rural areas are still far below the percentage of non-poor households.

Table 9. Significant variables in all MARS models for different versions of 2003 FIES

Variable	'Original'	'Non-missing'	Full (100%)	Training (60%)	Subset (10%)	Subset - Training (6%)	k = 1	k = 2	k = 3	p = 0	p = 0.05	p = 0.10	Overall	No. of times appeared in the model
age										x				26
sex														8
educ														23
hnagri										x				37
hsize	x	x	x	x	x	x	x	x	x	x	x	x	x	72
ofw														17
hmkshft														0
hsquat														0
helec														27
hwater_o														23
htoilet_o														25
hwtv		x	x	x	x									63
hwvhs										x				24
hwref					x	x				x				54
hwwash														21
hwaircon														0
hwcar														3
hwphone										x				36
hwcomputer														0
hwoven														2
hurb														1
No. of variables	1	2	2	2	3	2	1	1	1	6	1	1	1	

The best MARS models using all versions of CBMS data are summarized in Table 10. The number of significant variables for the original version of CBMS data set is the same as that for the non-missing version. The use of smaller data sets and allowing higher-order interactions led to inclusion of telephone/cellphone in the model, in addition to household size. Meanwhile, compared to FIES models, CBMS models tend to include relatively higher number of predictors under no-penalty and moderate-penalty scenarios.

Household size and telephone/cellphone ownership appeared to be the most significant variables that can discriminate between poor and non-poor households in CBMS data. These are followed by ownership of refrigerator/freezer and having a makeshift housing.

Table 10. Significant variables in MARS models for CBMS data

Variable	'Original'	'Non-missing'	Full (100%)	Training (60%)	Subset (10%)	Subset - Training (6%)	$k = 1$	$k = 2$	$k = 3$	$p = 0$	$p = 0.05$	$p = 0.10$	Overall	No. of times appeared in the model
age										x				24
sex														1
educ										x				32
hnagri														0
hsize	x	x	x	x	x	x	x	x	x	x	x	x	x	72
ofw														10
hmkshft	x									x				50
hsquat														1
helec														4
hwater_o										x				32
htoilet_o										x				33
hwtv														12
hwvhs														25
hwref														55
hwwash														22
hwaircon														0
hwcar														7
hwphone		x			x	x		x	x	x	x			70
hwcomputer														17
hwoven														8
No. of variables	2	2	1	1	2	2	1	2	2	7	2	1	1	

The following variables are among the most insignificant in MARS models for CBMS data: ‘hnagri’, ‘hwaircon’, ‘sex’, ‘hsquat’, ‘helec’, ‘hwcar’, ‘hwoven’, ‘ofw’, ‘hwtv’, and ‘hwcomputer’. The occurrence of categories of these variables among poor and non-poor households is similar, leading to model’s failure to recognize the dichotomy of poor and non-poor households. When smaller data sets are used, the discriminatory power of these variables decreased further, if not lost at all.

5.3. Comparison of Different Models

This paper ought to predict with higher accuracy the poor or potential beneficiaries of various poverty alleviation programs. The most ‘relevant’ measures to monitor here are the following: undercoverage rate, poverty accuracy, and BPAC, which is based mostly on the frequency of actual and predicted poor households. The correct identification of the poor is very important for program implementation and policy formulation. Although leakage increases as undercoverage decreases, there are some second-stage screening procedures that can lower the former. (Reyes, 2006) used a higher probability

cut-off in classifying the households which led to a higher leakage rate but lower undercoverage rate, and then applied electricity consumption as a second-stage screening variable to cut down the leakage rate. Various measures of predictive accuracy of the models for different versions of the data set are summarized in Appendix A.

To check the statistical significance of the differences among the models estimated in terms of the computed accuracy measures as well as the factors that contribute to these results, a series of parametric *t*-tests and their nonparametric counterparts are employed. The assumptions of normality and constancy in variance (for independent samples) were verified prior to the application of such tests. Tests for Normality on all these accuracy measures reject the normality assumption when all observations are considered and the reverse is true when the data is divided into sub-groups.

FIES and CBMS

Models estimated from FIES have better predictive performance than those estimated from CBMS. Specifically, FIES models have significantly higher accuracy in predicting poverty, lower undercoverage, and thus, higher BPAC. On the other hand, CBMS models performed better in predicting non-poverty. Similar set of results are observed when assessment was done separately for LR and MARS models.

The two data sets are different in a number of ways, which somehow explain the differences in the results. FIES is a nationwide survey wherein sample households are selected through a multi-stage stratified sampling procedure. This ensures that household-respondents are distributed among different combinations of socioeconomic attributes. In effect, it is relatively easier to find a set of

variables that could discriminate well between poor and non-poor households. On the other hand, CBMS is a complete enumeration of households within the entire city. Because it covers only a very small segment of the population (i.e., all urban), a large portion of households here share some common characteristics. For instance, because city is located in National Capital Region, households residing here have relatively higher employment opportunities and greater access to infrastructure and basic social services than those located in rural areas. Hence, higher living standard is expected and ownership of household appliances is more of a necessity than a luxury. Accuracy of the fitted model is enhanced when there is a large variation among the independent variables included since this is a more reasonable reflection of reality.

Furthermore, FIES has relatively lower volume of missing data and less susceptible to measurement errors. The presence of missing observations adversely affects predictive ability of the models due to the lost information associated with the estimation algorithm. Similarly, the presence of measurement errors tends to lower the discriminatory power of the affected variables because of the massive misclassification. While there are very few missing data in FIES, most of the variables in CBMS have missing observations. Further, in FIES, 11 observations were found to be extremely different from the rest of the sample, while a total of 88 outliers were found in CBMS. Most of these outlying observations, particularly in CBMS, are poor households, but own many (at least 4) assets, have access to basic amenities (i.e., electricity, safe water supply, sanitary toilet facility) and possess some characteristics of non-poor households. Another set of outlying observations are households who reported assets yet no access to electricity, a possible case of increasing incidence of illegal access to electricity in the area.

Original Data and Use of Non-missing Data Alone

Models estimated using 'non-missing' FIES data performed significantly better than those estimated using the 'original' FIES data sets, true for LR models, in terms of poverty accuracy and undercoverage rate. This is explained by the presence of missing values that contributes to the deterioration of the predictive performance of the models. Some of the important information carried by these observations that would be useful in building a better model is lost. Table 11 shows the total number of valid observations for all versions of the FIES data set used in LR model estimation.

Table 11. Number of valid observations in LR estimation, FIES data

Data set	Original	Non-missing
Full (100%)	36,578	36,578
Training (60%)	21,896	21,947
Subset (10%)	3,349	3,444
Training (6%)	2,025	2,195

The original versions of the data set performed better in terms of predicting non-poverty. (Hosmer and Lemeshow, 2000) noted that LR always favors classification into larger group. In this case, larger group is the non-poor group. Because percentage of non-poor households is relatively higher in the original version of FIES data than that in the non-missing version, as shown in Table 12, non-poverty accuracy of the models estimated using the original version is expected to be higher than that of the models estimated using the non-missing version.

Table 12. Percentage of poor and non-poor households in testing samples, FIES data

Data set	% poor	% non-poor
Original	25.59	74.41
non-missing	26.06	73.94

In MARS models, 'original' data sets performed better than the 'non-missing' data sets in terms only of total accuracy and non-poverty accuracy. For other measures, the difference between the two data sets is not significant. Because of automatic generation of missing value indicators in MARS, missing data in the original data set are not dropped in the analysis. These missing values are allowed to interact with any subset of other variables that may contribute in identifying the best model. Table 13 shows the total number of observations of all versions of FIES data set used in MARS estimation. In CBMS data as well, 'non-missing' data sets have higher predictive accuracy than 'original' data sets as measured by poverty accuracy, undercoverage rate and BPAC. There is also a lower accuracy in predicting non-poverty.

Table 13. Number of valid observations in MARS estimation, FIES data

Data set	Original	Non-missing
Full (100%)	42,094	36,578
Training (60%)	25,256	21,947
Subset (10%)	4,209	3,444
Training (6%)	2,525	2,195

Table 14 summarizes the total number of valid observations used in the estimation of LR models for all versions of CBMS data set. The difference between the sample sizes of the original and non-missing versions is not large. Thus, there is insignificant amount of information lost when missing observations were excluded in the analysis, which resulted in insignificant difference between the predictive accuracy of the models estimated from these two data versions.

Table 14. Number of valid observations in LR estimation, CBMS data

Data set	Original	Non-missing
Full (100%)	64,112	64,027
Training (60%)	38,446	38,416
Subset (10%)	6,411	6,403
Training (6%)	3,845	3,842

MARS models estimated using CBMS data set were found to perform well in the non-missing versions through the prediction of the poor and getting low exclusion error. Although original versions have relatively larger samples and thus larger amount of information that would contribute in finding better-fitting models compared to the non-missing versions, as shown in Table 15, resulting models led to lower predictive accuracy. Unfortunately, the models were able to predict only non-poor households and poor households were erroneously classified as non-poor. However, when missing data were excluded from these original data sets, MARS models were able to identify the poor households properly.

Table 15. Number of valid observations in MARS estimation, CBMS data

Data set	Original	Non-missing
Full (100%)	65,108	64,027
Training (60%)	39,066	38,416
Subset (10%)	6,511	6,403
Training (6%)	3,907	3,842

The MARS-fitted models using the original versions of the full and training samples of CBMS data set contain several missing value indicators of the different variables, which interacted with the non-missing regions of the other variables. It should be noted that the interaction of variable X_1 with a missing value indicator can only have non-missing values if that missing value indicator is for that particular variable, say X_1_missing. If the missing value indicator is for another variable, say X_2_missing, the values of variable X_1 can only be used when X_2 is missing. Thus, the presence of several interaction terms of missing value indicators of the different variables with the non-missing regions of other variables adversely affected the fit and thus, predictive accuracy of the model suffered. In the case of the original versions of CBMS data set, there is a possibility that values of the non-constant basis functions were offset due to such type of interaction. Since the value is greater

than 0.50, all poor households (which comprised of only 15% of the total population) are predicted as non-poor.

Large (40%-sample) and Small (4%-sample)

Data size, in general, does not significantly influence model performance for FIES data. For LR models, the results favor the smaller subsets as indexed by poverty accuracy and undercoverage rate. The performance of LR models, as measured through the goodness-of-fit, increases with sample size, (Sharma, 1996). But this is true only if all other things are being held constant. It should also be noted that LR fits a logistic function (nonlinear) and that increasing sample size does not always assure good model fit. If the population is heterogeneous like in FIES, there is a higher probability that the resulting fit of the model will be poor when larger sample size is used because more observations may contradict the logistic pattern of all other points. If sample size is smaller, fewer observations will be required by LR to fit into the logistic distribution. In fact, (Perlich et al., 2003) noted that the accuracy of an LR model is very good for small- to moderate-sized data sets but it levels off as the sample size increases. Moreover, FIES models estimated using larger samples have significantly higher non-poverty accuracy than those estimated using smaller subsets because percentage of non-poor in the former is higher than that in the latter, as shown in Table 16.

Table 16. Percentage of poor and non-poor households in testing samples, FIES data

Data set	% poor	% non-poor
large	25.44	74.56
small	26.20	73.80

For large and small data sets, MARS models performed similarly using FIES data (all accuracy measures, except non-poverty accuracy). In general, MARS estimates models involving lower-order

interactions when the sample size is small, while it entertains higher-order interactions as potential candidates when sample size is large. However, when models estimated using small-sized data set contains many interaction terms (especially if they are of higher-order), there is a tendency for the model to be overfitted, resulting in relatively lower predictive performance. MARS models estimated using larger samples have relatively higher non-poverty accuracy than those estimated using smaller samples because the latter models are almost as complex as the former.

For CBMS data set, sample size matters for both LR and MARS models, particularly in terms of the three accuracy measures. Classification accuracy of LR models are higher in smaller data sets because only few observations are forced to fit into the logistic function. Similarly, MARS models performed better in smaller data sets mainly because of the presence of several missing value indicators interacting with non-missing values of other variables in models estimated using larger subset of the original version of CBMS.

LR vs. MARS

Considering the different nuisance conditions of the data in the course of the analysis, MARS models performed better than LR models. In particular, MARS models led to higher poverty accuracy and BPAC as well as lower classification errors (as measured by both undercoverage and leakage rates) than LR models. This is true for FIES data sets, particularly when larger data sets are used. As a nonparametric method, MARS does not assume any specific form of distribution from the data and the models estimated by this method are expected to follow the distribution of the data, resulting in better function approximation and thus, better predictive performance. This holds true if the data set is

considerably large yet do not have very high fluctuations in their values. LR, on the other hand, is constrained by the functional form and requires the data to fit into a logistic distribution.

For CBMS data set, LR and MARS models are, in general, not significantly different from each another. However, in terms of total accuracy, MARS models performed better than LR models. Poor performance of the models estimated using the original full versions of the data due to presence of many missing values has lesser negative impact on the overall performance of MARS compared to poor performance of LR models estimated from large and original versions of the data sets.

For FIES data, absence of interaction in MARS models yields better performance as indicated by the 3 accuracy measures, followed by 3-way interaction models. Simpler models, or those consist only of main variables, are found to be the best in analyzing FIES data. For CBMS data, 2- and 3-way interaction models performed equally well in terms of the 3 accuracy measures while both performed better than the no-interaction models. This implies that CBMS data sets need more complex model to be able to explain better the variation in income and thus, poverty status of the households, particularly for larger data sets. When there is a limited number of indicators that clearly discriminate population dichotomies, interaction terms can help improve the predictive ability of MARS models.

For both FIES and CBMS data sets, MARS models performed best under a no-penalty scenario. No penalty is equivalent to data interpolation with complete absence of smoothness. Models performed equally well when penalty is set to 0.05 (moderate) and 0.10 (heavy). This implies that imposing penalty is not necessarily useful in prediction, particularly when the number of candidate predictors is small. Penalty limits the inclusion of certain predictors and thus, the predictive performance of the model. This is particularly true for CBMS data, which actually needs a relatively more complex

model to be able to predict poverty status of households more accurately. For FIES data, it seems that the original set of predictors and their interactions is good enough to predict the poverty status of households accurately.

6. Conclusions

Efficient and effective implementation of proxy means test lies on the ability of the regression technique to identify smaller number of socioeconomic variables that would predict household income and thus identify potential program beneficiaries as accurately as possible. This is particularly crucial in developing countries like the Philippines, where budget allocation for poverty alleviation programs are limited. Selection of the most theoretically and empirically sound set of indicators can somehow balance the minimization of undercoverage and leakage thereby leading to optimal impact of a poverty alleviation program.

MARS is a useful tool and a better alternative to LR in identifying a more parsimonious yet theoretically and empirically sound set of household poverty correlates and consequently, predicting income-based household poverty status (particularly the poor) with higher accuracy. Under a no-penalty scenario, MARS produced relatively smaller models than LR, wherein only few variables are needed to explain the variation in poverty status of households. This is important for a cost-effective proxy means test. MARS was able to include variables that can discriminate well between poor and non-poor households, which are as follows: household size, ownership of television set, and kind of business/work of household head in FIES data, and; household size only in CBMS data.

MARS models have higher predictive accuracy than LR models, particularly in terms of the three accuracy measures, namely: poverty accuracy, undercoverage and BPAC. Specifically, MARS performs significantly well under the following conditions: (i) large variations among the independent variables, leading to good spread of observations; (ii) low proportion of missing observations; (iii) large number of observations; (iv) lower interaction order (when the number of candidate predictors is relatively small), and; (v) lower penalty values (also when the number of candidate predictors is relatively small).

References

Ahmed, A. U. and Bouis, H. E. (2002) Weighing What's Practical: Proxy Means Tests for Targeting Food Subsidies in Egypt. *FCND Discussion Paper* No. 132, Washington, D.C.: International Food Policy Research Institute.

Balisacan, A. M. (1992) Rural Poverty in the Philippines: Incidence, Determinants and Policies. *Asian Dev. Rev.,* 10, 125-163.

Balisacan, A. M. (1993) Agricultural Growth, Landlessness, Off-Farm Employment ad Rural Poverty in the Philippines. *Econ. Dev. and Cult. Change*, 41, 533-562.

Balisacan, A. M. (1994) *Poverty, Urbanization, and Development Policy: A Philippine Perspective*, Quezon City: University of the Philippine Press.

Balisacan, A. M. (1997) In Search of Proxy Indicators for Poverty Targeting: Toward a Framework for a Poverty Indicator and Monitoring System. Paper Prepared for the National Statistics Office's Component in the UNDP-Assisted Project: *Strengthening Institutional Mechanisms for the Convergence of Poverty Alleviation Efforts*.

Bautista, R. M. and M. M. Lamberte. (1996) The Philippines. *Asian-Pacific Econ. Lit.*, 10 , 16-32.

Briand, L. C., B. Freimut, and F. Vollei. (2007) Using Multiple Adaptive Regression Splines to Understand Trends in Inspection Data and Identify Optimal Inspection Rates. ISERN TR 00-07. (Available at: http://www.salfordsystems.com).

Craven, P. and G. Wahba. (1979) Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerical Mathematics*, 31, 317-403.

De Veaux, R. D., A L. Gordon, J. C. Comiso, and N. E. Bacherer. (1993a) Modeling of Topographic Effects on Antarctic Sea Ice Using Multivariate Adaptive Regression Splines. *J. of Geophysical Res.,* 98, 20,307-20,319.

De Veaux, R. D., D. C. Psichogios, and L. H. Ungar. (1993b) A Comparison of Two Nonparametric Estimation Schemes: MARS and Neural Networks. *Computers and Chem. Engineering*, 17(8), 819-837.

Diaz, P. H. and R. M. M. Cariño. (2007) Pasay City: Eager to Learn from CBMS. In *Fighting Poverty: Lessons Learned from Community-Based Monitoring System Implementation Highlights of Case Studies,* V. A. Bautista (ed.). (Available at: http://www.pep-net.org).

Foster, D. and R. Stine. (2004) Variable Selection in Data Mining: Building a Predictive Model for Bankruptcy. *J. of the Amer. Stat. Assoc.*, 99, 303-313.

Fox, J. (2002) Nonparametric Regression. Appendix to: *An R and S-PLUS Companion to Applied Regression*.

Francis, Louise. (2007) Martian Chronicles: Is MARS Better Than Neural Networks?. (Available at: http://www.salfordsystems.com).

Frank, I. E. (1995) Tutorial: Modern Non-linear Regression Methods. *Chemometrics and Intelligent Lab. Sys.,* 27, 1-19.

Friedman, J. H. (1991) Multivariate Additive Regression Splines. *Ann. of Stat.,* 19(1), 1-67.

Grosh, M. E. (1994) *Administering Targeted Social Programs in Latin America: From Platitudes to Practice*. Washington, D.C.: World Bank.

Grosh, M. E. and Baker, J. L. (1995) Proxy Means Tests for Targeting Social Programs: Simulations and Speculation. *LSMS Working Paper* No. 118, Washington, D.C.: The World Bank.

Haughton, D. and Loan, L. T. T. (2004) Vulnerability of Vietnamese Households, 1992-1998. Working Paper (June 2004). Ford Foundation/General Statistics Office Project, Vietnam.

Herrin, A. N. and R. H. Racelis. (1994) *Measuring the Coverage of Public Programs on Low-Income Families: Philippines, 1992*. Pasig: NEDA Integrated Population and Planning Project.

Hosmer, D. W. and S. Lemeshow. (2000). *Applied Logistic Regression*, 2nd ed., Canada: John Wiley & Sons, Inc.

Houssou, N., M. Zeller, G. V. Alcaraz, S. Schwarze, and J. Johannsen. (2007) Proxy Means Tests for Targeting the Poorest Households: Applications to Uganda. Paper Prepared for Presentation at the 106th Seminar of the EAAE "Pro-Poor Development in Low Income Countries: Food, Agriculture, Trade, and Environment", 25-27 October 2007 in Montpellier, France.

International Labour Organization (ILO). (1974). *Sharing in Development: A Programme of Employment, Equity and Growth in the Philippines*. Geneva: ILO.

IRIS. (2005) Note on assessment and improvement of tool accuracy. Mimeograph (Revised version from June 2, 2005). IRIS Center, University of Maryland.

Jin, R., W. Chen, and T. W. Simpson. (2000) Comparative Studies of Metamodeling Techniques Under Multiple Modeling Criteria. *AIAA-2000-4801.* American Institute of Aeronautics and Astronautics, Inc.

Kuhnert, P. M., K. Do, and R. McClure. (2000) Combining Non-Parametric Models with Logistic Regression: An Application to Motor Vehicle Injury Data. *Comp. Stat. and Data Anal.*, 34, 371-386.

Leathwick, J. R., J. Elith, and T. Hastie. (2006) Comparative Performance of Generalized Additive Models and Multivariate Adaptive Regression Splines for Statistical Modeling of Species Distributions. *Ecol. Mod.*, 199, 188-196.

Lipton, M. and M. Ravallion. (1995) Poverty and Policy. In *Handbook of Development Economics*, Volume III, ed. J. Behrman and T. N. Srinivasan. Amsterdam: North Holland.

Manasan, R. G. and J. S. Cuenca. (2007) Who Benefits from the Food-for-School Program and *Tindahan Natin* Program: Lessons in Targeting. *Phil. J. of Dev.,* 34(1), 1-33.

Marquez, N. R. and R. A. Virola. (1997) Monitoring Changes in the Characteristics of the Philippine Poor: 1985 to 1994. Proceedings of the Sixth National Convention on Statistics, Manila: National Statistical Coordination Board.

Mukkamala, S. A., H. Sung, A. Abraham, and V. Ramos. (2004). *Intrusion Detection Systems Using Adaptive Regression Splines*.

Muñoz, J. and A. M. Fellicisimo. (2004) Comparison of Statistical Methods Commonly Used in Predictive Modeling. *J. of Vegetation Sci.*, 15, 285-292.

National Statistics Office (NSO). (2003) 2003 Master Sample (MS). *Draft* (as of October 14, 2003).

Orbeta, A. C., Jr. (2006) Poverty, Vulnerability and Family Size: Evidence from the Philippines. *ADB Institute Discussion Paper No. 29*, Asian Development Bank Institute.

PEP-CBMS Network Coordinating Team. (2008) Community-Based Monitoring System (Primer). (Available at: http://www.pep-net.org).

Perlich, C., F. Provost, and J. Simonoff. (2003) Tree Induction vs. Logistic Regression: A Learning-Curve Analysis. *J. of Machine Learning Res.*, 4, 211-255.

Reyes, C. M. (2003) The Poverty Fight: Has It Made an Impact?. *Perspective Paper Series* No. 2. Makati City: Philippine Institute for Development Studies.

Reyes, C. M. (2006) Alternative Means Testing Options Using CBMS. *PIDS Discussion Paper Series* No. 2006-22. Makati City: Philippine Institute for Development Studies.

Salford Systems, Inc. (2001) *MARS™ User Guide*. (Available at: http://www.salfordsystems.com).

Sharma, S. (1996). *Applied Multivariate Techniques*. New York: John Wiley & Sons, Inc.

Tabuga, A. D. (2007) International Remittances and Household Expenditures: The Philippines Case. *PIDS Discussion Paper Series* No. 2007-18. Makati City: Philippine Institute for Development Studies.

Tabunda, A. M. L. (2000) Towards the Construction of a Classification Rule for Poverty Status. National Statistics Office and United Nations Development Programme, Manila, Philippines.

Ture, M., I. Kurt, A. T. Kurum, and K. Ozdamar. (2005) Comparing Classification Techniques for Predicting Essential Hypertension. *Expert Systems with App.*, 29, 583-588. (Available at: http://www.elsevier.com/locate/eswa).

World Bank. (1990). *World Development Report*. Washington, D. C.: World Bank.

World Bank, International Monetary Fund, Asian Development Bank and Inter-American Development Bank Core Team for APEC Financial Ministers. (2000) *Social Safety Nets in Response to Crisis: Lessons and Guidelines from Asia and Latin America.* Washington, D.C., (mimeographed).

Appendix A. Predictive performance of the best set of LR and MARS models

(1) LR models

Data set	Total accuracy	Poverty accuracy/ Specificity	Non-poverty accuracy/ Sensitivity	Undercoverage rate	Leakage rate	Balanced Poverty Accuracy Criterion
2003 FIES						
'Original'						
Testing (40%)	80.29	48.70	90.53	51.30	37.52	26.64
Subset						
Testing (4%)	83.99	60.42	92.58	39.58	25.21	41.19
'Non-missing'						
Testing (40%)	79.15	54.90	87.86	45.10	38.12	43.61
Subset						
Testing (4%)	82.78	63.68	89.38	36.32	32.53	58.07
2005 CBMS Pasay City						
'Original'						
Testing (40%)	85.27	1.04	99.79	98.96	53.49	-96.71
Subset						
Testing (4%)	85.83	25.45	96.56	74.55	43.18	-29.77
'Non-missing'						
Testing (40%)	86.22	0.88	99.87	99.12	47.46	-97.45
Subset						
Testing (4%)	88.21	22.38	98.42	77.62	31.25	-45.06

Note: All figures are in percentage (%).

Appendix A. (cont'd.)

(2) MARS models with k=1 and p=0

Data set	Total accuracy	Poverty accuracy/ Specificity	Non-poverty accuracy/ Sensitivity	Undercoverage rate	Leakage rate	Balanced Poverty Accuracy Criterion
2003 FIES						
'Original'						
Testing (40%)	82.72	61.11	90.91	38.89	28.18	46.21
Subset						
Testing (4%)	82.96	64.15	90.70	35.85	26.06	50.92
'Non-missing'						
Testing (40%)	82.46	64.30	90.10	35.70	26.78	52.11
Subset						
Testing (4%)	81.20	63.79	88.41	36.21	30.53	55.61
2005 CBMS Pasay City						
'Original'						
Testing (40%)	85.29	0.00	100.00	100.00	-	-100.00
Subset						
Testing (4%)	87.67	32.06	97.56	67.94	30.00	-22.14
'Non-missing'						
Testing (40%)	87.83	26.13	97.70	73.87	35.45	-33.40
Subset						
Testing (4%)	88.13	23.84	98.11	76.16	33.87	-40.12

Notes: All figures are in percentage (%).
k = interaction order; p = penalty

Appendix A. (cont'd.)

(3) MARS models with k=2 and p=0

Data set	Total accuracy	Poverty accuracy/ Specificity	Non-poverty accuracy/ Sensitivity	Undercoverage rate	Leakage rate	Balanced Poverty Accuracy Criterion
2003 FIES						
'Original'						
Testing (40%)	83.01	60.90	91.39	39.10	27.16	44.50
Subset						
Testing (4%)	83.25	65.38	90.61	34.62	25.87	53.56
'Non-missing'						
Testing (40%)	82.71	63.40	90.83	36.60	25.58	48.58
Subset						
Testing (4%)	81.89	61.45	90.34	38.55	27.55	46.26
2005 CBMS Pasay City						
'Original'						
Testing (40%)	85.29	0.00	100.00	100.00	-	-100.00
Subset						
Testing (4%)	88.21	34.86	97.69	65.14	27.13	-17.30
'Non-missing'						
Testing (40%)	87.82	27.20	97.52	72.80	36.27	-30.12
Subset						
Testing (4%)	88.32	24.13	98.29	75.87	31.40	-40.70

Notes: All figures are in percentage (%).
k = interaction order; p = penalty

Appendix A. (cont'd.)

(4) MARS models with k=3 and p=0

Data set	Total accuracy	Poverty accuracy/ Specificity	Non-poverty accuracy/ Sensitivity	Undercoverage rate	Leakage rate	Balanced Poverty Accuracy Criterion
2003 FIES						
'Original'						
Testing (40%)	83.20	61.07	91.59	38.93	26.65	44.33
Subset						
Testing (4%)	83.19	63.34	91.37	36.66	24.88	47.66
'Non-missing'						
Testing (40%)	82.80	64.39	90.55	35.61	25.86	51.23
Subset						
Testing (4%)	81.61	60.28	90.43	39.72	27.73	43.69
2005 CBMS Pasay City						
'Original'						
Testing (40%)	85.29	0.00	100.00	100.00	-	-100.00
Subset						
Testing (4%)	87.86	30.53	98.06	69.47	26.38	-27.99
'Non-missing'						
Testing (40%)	87.89	26.86	97.65	73.14	35.31	-31.62
Subset						
Testing (4%)	87.78	23.55	97.74	76.45	38.17	-38.37

Notes: All figures are in percentage (%).
k = interaction order; p = penalty

Appendix A. (cont'd.)

(5) MARS models with k=1 and p=0.05

Data set	Total accuracy	Poverty accuracy/ Specificity	Non-poverty accuracy/ Sensitivity	Undercoverage rate	Leakage rate	Balanced Poverty Accuracy Criterion
2003 FIES						
'Original'						
Testing (40%)	81.22	55.28	91.06	44.72	29.91	34.15
Subset						
Testing (4%)	80.64	57.23	90.28	42.77	29.22	38.09
'Non-missing'						
Testing (40%)	80.22	56.38	90.25	43.62	29.13	35.93
Subset						
Testing (4%)	80.93	61.45	88.99	38.55	30.24	49.53
2005 CBMS Pasay City						
'Original'						
Testing (40%)	85.29	0.00	100.00	100.00	-	-100.00
Subset						
Testing (4%)	87.48	24.68	98.64	75.32	23.62	-43.00
'Non-missing'						
Testing (40%)	87.42	22.87	97.74	77.13	38.13	-40.16
Subset						
Testing (4%)	87.50	19.48	98.06	80.52	39.09	-48.55

Notes: All figures are in percentage (%).
k = interaction order; p = penalty

Appendix A. (cont'd.)

(6) MARS models with k=2 and p=0.05

Data set	Total accuracy	Poverty accuracy/ Specificity	Non-poverty accuracy/ Sensitivity	Undercoverage rate	Leakage rate	Balanced Poverty Accuracy Criterion
2003 FIES						
'Original'						
Testing (40%)	81.26	56.28	90.74	43.72	30.27	36.98
Subset						
Testing (4%)	80.17	52.14	91.70	47.86	27.89	24.44
'Non-missing'						
Testing (40%)	80.65	58.83	89.83	41.17	29.12	41.82
Subset						
Testing (4%)	79.36	53.50	90.05	46.50	31.02	31.07
2005 CBMS Pasay City						
'Original'						
Testing (40%)	85.29	0.00	100.00	100.00	-	-100.00
Subset						
Testing (4%)	86.98	29.52	97.20	70.48	34.83	-25.19
'Non-missing'						
Testing (40%)	87.24	21.20	97.81	78.80	39.25	-43.90
Subset						
Testing (4%)	87.39	20.64	97.74	79.36	41.32	-44.19

Notes: All figures are in percentage (%).
k = interaction order; p = penalty

Appendix A. (cont'd.)

(7) MARS models with k=3 and p=0.05

Data set	Total accuracy	Poverty accuracy/ Specificity	Non-poverty accuracy/ Sensitivity	Undercoverage rate	Leakage rate	Balanced Poverty Accuracy Criterion
2003 FIES						
'Original'						
Testing (40%)	81.26	56.30	90.73	43.70	30.28	37.05
Subset						
Testing (4%)	80.94	57.23	90.70	42.77	28.32	37.07
'Non-missing'						
Testing (40%)	80.51	58.64	89.72	41.36	29.42	41.73
Subset						
Testing (4%)	79.84	55.61	89.86	44.39	30.61	35.75
2005 CBMS Pasay City						
'Original'						
Testing (40%)	85.29	0.00	100.00	100.00	-	-100.00
Subset						
Testing (4%)	86.98	29.77	97.15	70.23	35.00	-24.43
'Non-missing'						
Testing (40%)	87.24	21.17	97.81	78.83	39.24	-43.99
Subset						
Testing (4%)	86.88	24.71	96.53	75.29	47.53	-28.20

Notes: All figures are in percentage (%).
k = interaction order; p = penalty

Appendix A. (cont'd.)

(8) MARS models with k=1 and p=0.10

Data set	Total accuracy	Poverty accuracy/ Specificity	Non-poverty accuracy/ Sensitivity	Undercoverage rate	Leakage rate	Balanced Poverty Accuracy Criterion
2003 FIES						
'Original'						
Testing (40%)	80.87	54.89	90.72	45.11	30.84	34.26
Subset						
Testing (4%)	80.64	57.23	90.28	42.77	29.22	38.09
'Non-missing'						
Testing (40%)	80.22	56.38	90.25	43.62	29.13	35.93
Subset						
Testing (4%)	80.93	61.45	88.99	38.55	30.24	49.53
2005 CBMS Pasay City						
'Original'						
Testing (40%)	85.29	0.00	100.00	100.00	-	-100.00
Subset						
Testing (4%)	87.48	24.68	98.64	75.32	23.62	-43.00
'Non-missing'						
Testing (40%)	87.42	22.87	97.74	77.13	38.13	-40.16
Subset						
Testing (4%)	87.50	19.48	98.06	80.52	39.09	-48.55

Notes: All figures are in percentage (%).
k = interaction order; p = penalty

Appendix A. (cont'd.)

(9) MARS models with k=2 and p=0.10

Data set	Total accuracy	Poverty accuracy/ Specificity	Non-poverty accuracy/ Sensitivity	Undercoverage rate	Leakage rate	Balanced Poverty Accuracy Criterion
2003 FIES						
'Original'						
Testing (40%)	81.27	56.41	90.70	43.59	30.32	37.35
Subset						
Testing (4%)	78.15	53.97	88.10	46.03	34.89	36.86
'Non-missing'						
Testing (40%)	80.15	53.61	91.32	46.39	27.79	27.86
Subset						
Testing (4%)	79.90	53.50	90.82	46.50	29.32	29.21
2005 CBMS Pasay City						
'Original'						
Testing (40%)	85.29	0.00	100.00	100.00	-	-100.00
Subset						
Testing (4%)	86.98	30.28	97.06	69.72	35.33	-22.90
'Non-missing'						
Testing (40%)	87.18	21.45	97.69	78.55	40.17	-42.68
Subset						
Testing (4%)	87.27	20.93	97.56	79.07	42.86	-42.44

Notes: All figures are in percentage (%).
k = interaction order; p = penalty

Appendix A. (cont'd.)

(10) MARS models with k=3 and p=0.10

Data set	Total accuracy	Poverty accuracy/ Specificity	Non-poverty accuracy/ Sensitivity	Undercoverage rate	Leakage rate	Balanced Poverty Accuracy Criterion
2003 FIES						
'Original'						
Testing (40%)	81.74	58.85	90.42	41.15	30.05	42.97
Subset						
Testing (4%)	81.77	58.45	91.37	41.55	26.41	37.88
'Non-missing'						
Testing (40%)	80.51	58.64	89.72	41.36	29.42	41.73
Subset						
Testing (4%)	79.36	53.04	90.24	46.96	30.79	29.67
2005 CBMS Pasay City						
'Original'						
Testing (40%)	85.29	0.00	100.00	100.00	-	-100.00
Subset						
Testing (4%)	86.98	30.28	97.06	69.72	35.33	-22.90
'Non-missing'						
Testing (40%)	87.38	22.95	97.69	77.05	38.61	-39.65
Subset						
Testing (4%)	87.43	19.19	98.02	80.81	40.00	-48.84

Notes: All figures are in percentage (%).
k = interaction order; p = penalty