# Classification Trees for Poverty Mapping

A thesis presented in partial
fulfilment of the requirements
for the degree of
Master
in Applied Statistics

at Massey University,
Palmerston North, New Zealand.

Tian Mao
2010

**Abstract**

Measuring differences in poverty levels within a country is important for aid allocation. Small area estimates of poverty incidence can be found by combining census and survey data. The usual method uses multiple regression, but an intuitive alternative is to build a classification tree for classifying households as poor or non-poor. This research presents some preliminary results using this method, and compares them to the traditional regression method.

# Acknowledgements

To fully acknowledge everyone who has helped me throughout the year. Many people have supported me in many ways over this year, to all of them I am very appreciative. A few who deserve particular mention are listed below.

Firstly to my supervisors Geoff Jones, Siva Ganesh and Stephen Haslett. They were so encouraging and been ongoing support, guidance over the period of writing the thesis as well as ensured my Master year has been successful, educational and enjoyable. Throughout the year they have been more than a supervisor to me, they have showed me the way of having academic and independent thinking. From them I learnt to be an organized person for knowing every step of your way. This has helped me often.

To the members of the Department of Statistics and those now in the statistics group in the Institute of Fundamental Sciences. Thank you for been patient and help me in times of need.

Finally, I dedicate my work to my parents. I cannot thank them enough for their support through the years I am studying in New Zealand. It has had its ups and downs, hopefully will make it all worthwhile in the future.

# Contents

# Chapter 1

# Introduction

## 1.1 Poverty

Literally poverty means being unable to afford the basic human needs, such as clean water, health care, education, nutrition, or shelter, etc. Poverty is defined as a shortfall in the level of income or expenditure from a poverty line, representing the amount needed to meet the basic needs for a person or household. Measuring poverty helps to identify which countries, or regions within a country, are in greatest need of remedial action. Different countries take different approaches to remedial action regarding the details of cost for the basic needs, such action could be food supply or monetary allowance. Another purpose could be for measuring the performance of a change in economic policy, to give feedback on the effectiveness of the policy.

Poverty lines are calculated to represent the monetary resources required to meet the basic human needs for the members of a household. The methodology using to calculate poverty line concerns the choice of income or expenditure as both indicate welfare, the measure of expenditure is rather difficult and an under-estimate of income will be collected as people tend to report less than their actual earnings. If earning for a person is below the poverty line then it will classified as poor otherwise its non-poor. On household basis, if the average earning per person of a household is below the poverty line it will define as a 'poor' household and 'non-poor' otherwise.

The official poverty lines are established following a cost-of-basic-needs approach. After defining the poverty line, property of households are known which then gives the proportion of household in a certain region is below the poverty line. It defines the poverty incidences in household level, alternative, the poverty incidence can also be defined in individual level. There brings the concept of poverty incidence, which

is defined as the proportion of people whose living condition is below the poverty line in a given area. The incidence is calculated as proportion of individuals in households living in a region on average per capita income or expenditure below the poverty line. The original data has larger poverty incidence in person-level than household-level, as the poor families tend to be larger than non-poor. The poverty incidence is measured by statisticians from governmental or international agencies. The government statistical agency conducts a family income and expenditure survey every three years in order to collect information on household income and expenditure with respect to socio-demographic characteristics. In this survey data the basic unit for measuring income or expenditure is the household, even though poverty incidence is usually calculated based on per person. The survey gives a reliable estimate of poverty for the whole country, and for large administrative regions within the country.

Our data has been collected in the Philippines by its National Statistical Coordination Board; it conducts the family income and expenditure survey (FIES) to calculate poverty lines and estimate poverty. The FIES for 2000 officially has poverty incidence of 27%, if the measurement is taken based on the household level. The person-level of poverty incidence is 34.15%. The higher poverty incidence in person-level is due to households which been defined as poor tend to have larger number of family number than the non-poor family.

## 1.2 Survey Design

Probability based sampling means to select part of population using known probabilities so that one may estimate something about the whole population. Thus, to estimate the poverty in a country statisticians collect relevant variables from a group of households that have been randomly selected within the study area. Some important questions for such study are how best to obtain the sample and make the observations and, once the sample data are in hand, how best to use them to estimate the characteristic of the whole population.

Random sampling is part of statistical practice concerned with the selection of an unbiased or random subset of individual observations within a population of individuals intended to yield some knowledge about the population of concern. Using

this method it is like to be less biased, and the estimation provides the standard errors. The random sampling procedure selects sampling units for all elements of the sampling frame (Lohr, 1998).

Simple random sampling selects units out of a population and gives each possible sample an equal chance of selection. The disadvantage of using this method is it is a time consuming and expensive procedure, data set large as our data which the census include the whole population in a country, practically it is not possible to implement this method. One other method need to be mentioned is stratified random sampling, every population element is grouped uniquely into a group of stratum, in our data would be stratified according to region. The stratified random sampling has a property that important sub-populations of interest can but need not be represented in the sample relative to their proportion in the population. Comparing to simple random sampling a balanced sample been selected with respect to target characteristics across all stratified regions. One other method is cluster sampling; the method is conducted in close areas due to the cost constraints. The systematic sampling relies on arranging the target population according to some ordering scheme and then selecting elements at regular intervals through that ordered list. Systematic sampling involves a random start and then proceeds with the selection of ever $k^{th}$ element from then onwards.

A typical survey design for national household surveys would be to stratify data according to geographical districts or rural and urban methods, and contains sub-populations with respect to all of target characteristics. Cluster sampling has been done after the stratification, for the economical purposes the cost of applying this method is to bring up the error; the clustering could be operated according to the smallest geographical unit such as postcode area, meshblocks or villages (Lohr, 1998). Clusters can be selected at random or using probability proportional to size method, in which it means the selection probability for each cluster is set to be proportion to its size measure.

Not every household or person is included in the survey sample, only a certain amount of households participate the survey, it considers as survey weights. To be specific as if a household (person) been selected out of hundred households (people) then this household (person) representing for this hundred household (people). The survey weight is the inverse of sampling probabilities. Survey weights vary from one municipal to another; the reason for survey weights varying are because of different

sampling proportions in different strata, and non-response.

The design could affect the analysis in various ways; only two aspects show up in our design. First is the survey weights we have conducted due to the data collection method. To calculate the populations mean equation form up as shown below, and for survey sample since weights been involved here hence the weights need to counted, again the equation shown below (where $w_i$ is the survey weight):

$$Population\ Mean = \frac{1}{N}\sum_{i \in c} X_i$$

$$Survey\ Mean = \sum_{i \in s} w_i X_i$$

Second, the design of using clusters will affect the calculating of standard errors, as for usual cases the equation for getting standard error is $\sigma/\sqrt{n}$, since cluster sampling selects respondents from certain areas therefore it is more likely that each sample may have correlation between one and another. The calculation of standard errors in this case would involve the design effects (Chambers & Skinner, 2003). There are couple of methods for calculating the standard error such as Balanced Repeated Replication, Jackknife and Linearization.

The data we have been using collects information relevant to family income and expenditure, the data we applying is from the Philippine and called FIES 2000 (FIES = Family Income and Expenditure Survey), it is conducted in July 2000 for the period January 1 to June 30 and July 1 to December 31 at 2001. The technique of multi-stage stratified random sampling been used for the sample design of FIES 2000. Barangays are the primary sampling units and these are stratified into urban or rural within each province and use the systematic sampling selection method with probability proportional to size.

## 1.3 Small Area Estimation with ELL method

Small area estimation is a technique to estimate parameters for small sub-populations, generally its been used when the sub-population of interest is included in a larger survey. The term 'Small Area' usually refers to a small geographical area, in our data the small area could refer as municipality or barangay. It could also refer to a

'small domain'. Since the survey has been carried out for the population as a whole (here is a nation wide survey), the sample size within some particular small areas (say municipality) is too small to generate accurate estimates from the data. In order to overcome this problem, it is possible to use additional data that exists for these small areas to obtain estimates, which sometimes refer as 'borrow strength' (Rao, 2003).

Small area methods 'borrow strength' from related or similar areas through explicit or implicit statistical models, it connect the small areas poverty estimates by using available supplementary data in both survey and census. A common small area model was first discovered in 1988 (Battese, Harter & Fuller, 1988) and been widely used today. The target variable denoted as Y, in our data its a poverty measure, the estimations for a small population are required corresponding to small geographical areas. For each subpopulation estimations of Y be obtained from sample survey data directly, however these estimates are not sufficiently accurate because within subpopulation the sample size is typically small (or zero), therefore these direct estimates have large standard errors.

For poverty measure a specific method been designed in small area estimation. The method was first discovered by Elbers, Lanjouw & Lanjouw in 2003, also known as ELL method. The linear regression model is applied on the log-transformed income values, using random cluster effects, where the clusters may be differ from the small areas.

The model can be formed as a matrix relationship between Y and X:

$$Y_{ij} = X_{ij}\beta + \mu_i + \varepsilon_{ij} \tag{1.1}$$

Where Y is the response variable, log-transformed income value for group i household j. $\mu_i$ is the cluster effects and $\varepsilon$ is the household effects.

For conducting the ELL method there are a few steps need to follow, first a set of auxiliary variables been identified which available for the whole population as well as the survey. A two stage least squares procedure with an equal correlated covariance structure been involved in fitting a survey data for the original method. The algebraic adjustment been used in ELL method does not properly account for the sample survey weights (Elbers et al., 2003). The standard application of the

small area estimation applying the estimation model to known X values with entire population in order to produce Y values for every household. Average all the Y values over each small area to obtain point estimate and the standard error of which is inferred from appropriate asymptotic theory. The household-level predictions of Y are converted into poverty status using the poverty line. It is which this is averaged over each small area, not Y. The bootstrapping procedure applies in the ELL method for producing unbiased estimates and standard errors. Here, fact that perferring the tree-based method is not because of ELL is more difficult. The ELL method assumes a linear model; it is harder to accommodate interactions. Conversely, the tree-based method is easily accommodate nonlinearity and interactions into the model.

## 1.4 Contribution

In the next Chapter we will explore the use of classification tree to directly predict poverty at household level, and compare its performance with the standard (ELL) method using a real data set. To our knowledge, this is the first time that classification trees have been used for poverty estimation.

The survey weights been participated in data collection, hence we need to consider how to adapt classification trees to account for survey design. This is done in Chapter 3.

We try to find ways to produce standard errors for tree-based poverty estimates in Chapter 4.

Throughout here we will apply the method using real data to illustrate the processes that we have mentioned above.

# Chapter 2

# Classification

## 2.1 Choosing a split

A classification tree (Breiman & Friedman, 1984) is a type of technique to classify all the households into categories of poor or non-poor. The RPART program (which is a package in R statistic software) builds classification models of a very general structure using a two stage procedure; the resulting models can be represented as binary trees. The classification tree has a tree-like structure and it makes no assumptions about distributions of data. The tree is formed by nodes, each node is simply a question about one variable, for example 'Is value of variable all_coed > 19.09% ? '. Ultimately the branches finish at a leaf or terminal node, where all objects are of the same type, or there are too few objects to classify further. A binary split comes out from each node, for example to answer the question above there are only two answers which are either 'yes' or 'no'.

The poverty data concentrates on defining households as 'poor' and 'non-poor' by including all the variables in the model (such as family size, education level for each family, average age of a household, etc). The way to split the classification tree is operated according to a certain rule as one household falling into one specific category, apply to the poverty data would be defining as poor and non-poor.

In the actual data there are 72% of households have been defined as non-poor and 28% fall into poor. Each binary split has its own proportion of poor and non-poor, therefore the information obtained by using different variable would be different. There are two different criteria for measuring the information, one is Entropy and the other is Gini.

In general, if there are m categories with proportions $p_1$, $p_2$, $p_3$,..., $p_m$ then the criteria are: E(x) is the Entropy criterion defined in equation 2.1 and G(x) is the Gini criterion defined in equation 2.2.

$$E(x) = -\sum_{k=1}^{m} p_k \log_2(p_k) \tag{2.1}$$

$$G(x) = 1 - \sum_{m} p_k^2 \tag{2.2}$$

In order to help understand better for the method we will illustrate it by using the example of the first variable selection in the poverty data.

The first step is the '(unexplained) information' before the very first split. In the Poverty data, $p_1$ is the proportion of non-poor household and $p_2$ is the proportion of poor household out of the whole data set, so using the Entropy measure the initial information is equation 2.1:

$$I(\textit{'poverty data'}) = -p_1 \log(p_1) - p_2 \log(p_2)$$
$$= -0.72 \log(0.72) - 0.28 \log(0.28) = 0.2575$$

Step two, take one explanatory variable such as 'all_coed' (which means proportion of household members 10 year and over with college education), and choose a cutoff value of that variable eg. 0.19. This defines a binary split of the data into two sets $S_1$ and $S_2$, see Table 2.1. Construct two tables in order to understand better:

Table 2.1: General Split Table

| Binary Splits | $S_1$ | $S_2$ |
|---|---|---|
| Proportion.1 | $P_{11}$ | $P_{21}$ |
| Proportion.2 | $P_{12}$ | $P_{22}$ |
| Weights | $w_1$ | $w_2$ |

Table 2.2: All_coed Split Table

| Poverty | all_coed $< 19\%$ | all_coed $\geq 19\%$ |
|---|---|---|
| Poor | 1059 | 9875 |
| Non-Poor | 14206 | 14389 |
| Total | 15265 | 24264 |

Hereby the two table as the second Table 2.2 illustrates the binary splits are 15265 (which is $S_1$) and 24264 (which is $S_2$). The equation for information with this

particular binary split value would be:

$$\begin{aligned} I(split) &= w_1 I(S_1) + w_2 I(S_2) \\ &= w_1 I(all\_coed < 19\%) + w_2 I(all\_coed \geq 19\%) \\ &= w_1(-p_1 \log(p_1) - p_2 \log(p_2)) + w_2(-p_1 \log(p_1) - p_2 \log(p_2)) \\ &= (15265/39529)(-0.93 \log(0.93) - 0.069 \log(0.069)) \\ &+ (24264/39529)(-0.59 \log(0.59) - 0.41 \log(0.41)) \\ &= 0.22 \end{aligned}$$

The equation above is the information after the split, within the equation the $w_i$ is the weight, in the other words, the proportion of each split (as total number of households with larger than and equal to or less than 19% of family members have 10 years and more of the college education) occupied in total number of households. This equation will apply to every possible cutoff that has been contained in 'all_coed' variable, in order to obtain the percentage of the information that presented with using different proportion of family members with 10 and more years of college education, compare them and selection the one with maximum percentage of information to start on the splitting process. The same process is applied to all variables to decide which has the maximum information gain among all variables.The encoding information that would be gained by branching on explanatory variable is, gain(x)=I(x)-E(x), where I(x) is the information before split and E(x) is the information after split, as in the example I(x)=0.2575 and E(x)=0.22.

Up till here only the root node has been selected and explained, for the next layer of variable selection the same process will be carried out except the information after the split would now be the information before the split. The same process is repeated until all the observations belong to the same class which each of them have been classified as either poor or non-poor, or all the observations have identical attribute values. For the poverty data, the initial split involves the variable 'all_coed' followed by 'famsize' (family size), 'roofN' (roof type), 'Hou_own_tel' (proportion of households who have telephone) and 'Hou_own_ref' (proportion of households who have refrigerator), etc. Check the Figure 3 in Appendix which gives the optimal Gini tree based on un-weighted data, for the convenience of reading the specific rules for the tree has been listed in Figure 1 and 2. There are other trees which been built with different data base show in Appendix.

The same process applies to the Gini method as well, the only difference would be the equation we applied for the initial variable selection and the equation we applied with weights. Here again we will use the poverty data to illustrate. Use the equation that presented earlier, refers back to equation (2.2).

Therefore for Gini method the information gain equation would be state as below, and the proportion is 0.4032.

$$\begin{aligned} G(\text{`poverty data'}) &= 1 - \sum_k (p_k^2) = 1 - (p_1^2 + p_2^2) \\ &= 1 - (0.28^2 + 0.72^2) \\ &= 0.4032 \end{aligned} \tag{2.3}$$

The following step is to select the variable with the maximum information gain among all the variables. Take example as variable 'all_coed' with proportion of 16.67 here (in Table 2.3), since this is the first split occurred in the Gini classification tree, to calculate the information gain after the split would be:

$$\begin{aligned} G(x) = w_1 I(x_1 < k) + w_2 I(x_1 \geq k) = (24224/39529)(1 - (0.59^2 + 0.407^2) \\ +(15305/39529)(1 - (0.93^2 + 0.07^2)) = 0.3482 \end{aligned} \tag{2.4}$$

Table 2.3: All_coed Split Table

| Poverty | all_coed $< 16.67\%$ | all_coed $\geq 16.67\%$ |
|---|---|---|
| Poor | 9858 | 1076 |
| Non-Poor | 14366 | 14229 |
| Total | 24224 | 15305 |

Generally, the procedure of binary split and selection of variable to form the split is the same as described previously. The only difference would be the variable selection method that applied to form two models. The extended part of trees is obvious on the graph. Again, here the encoding information that would be gained by branching on x variable is, gain(x)=I(x)-E(x), where I(x) is the information before split and E(x) is the information after split, as in the example I(x)=0.4032 and E(x)=0.3482.

## 2.2 Optimizing the Tree

First of all the whole data set is randomly split into two subsets training and testing (the proportion of two sets is 80 and 20). It is the act of partitioning available data into two portions, the training data is used to develop a predictive model and the testing data to evaluate the model's performance. The resultant model is, with certainty, too complex, and the question arises as it does with all procedures when to stop. The second stage of the procedure consists of using cross-validation to trim back the full tree.

The partitioning method can be applied to many different kinds of data. Here we concentrate on classification problems. The sample population consists of 39529 observations from 2 classes. We break these observations into 10 terminal groups; to each of these groups is assigned a predicted class (this will be the response variable). In other words, building model on 9 terminal groups and leave one group out, run the same procedure 10 times each time leave one different group out. Hence there are ten models being built. The purpose of having one group left out is to test the model for its precision, the number of misclassification that been occurred by fitting the predictive model divided by the total number of classified is treated as the misclassification error. The cross-validation is applied to provide an unbiased estimate of misclassification error. The misclassification error is related with the tree size in a plot (Figure 2.1). Building ten models with different partition sets, full trees have been developed and pruned back in order to obtain simpler models. Different numbers of leaves obtain different values for misclassification.

To construct a decision tree the most common approach is to grow a full tree and prune it back. The reason for pruning to be desirable is because the tree may grown overfit the data by inferring more structure than is justified by the training set. Therefore, a good choice of CP (CP is the cost complexity pruning the measurement of the complexity of the model) for pruning is often the lowest value for which the mean lies below the horizontal line (The horizontal line is drawn one standard error above the minimum of the curve), as show in Figure 2.1 and 2.2. The CP table is a set of possible cost-complexity pruning of a tree from a nested set. For the geometric means of the intervals of values of CP for which a pruning is optimal, a cross-validation has been done in the initial construction by RPART (recursive partitioning). The CP table in the fit contains the mean and standard deviation of the errors in the cross-validated prediction against each of the geometric means,

and these are plotted by RPART function. The table is printed from the smallest tree to the largest one. We find it easier to compare one tree to another when they start at the same place.

We define the cost complexity criterion:

$$C_\alpha(T) = \sum_{m=1}^{|T|} N_m Q_m(T) + \alpha|T| \tag{2.5}$$

where $Q_m(T) = (1/N_m)\sum_{x_i \in R_m}(y_i - c_m)^2$ is the squared-error node impurity measure, and $|T|$ denotes the number of terminal nodes in T. We index terminal nodes by m, with node m representing region $R_m$. $c_m$ is the average of $y_i$ in region $R_m$. $N_m$ is the number of observations check the equation 2.5.

In order to help understand it better here we going to illustrate it with an example, say when models with node number equals four, there will be ten different models produced with same number of nodes. Each model will have misclassified observations here in total it counted as misclassification, the number of misclassification showed up on the tree size plot, to obtain the misclassification error by plus all misclassifications together and average it. On the CP table for easier reading, the error columns have been scaled so that the first node has an error of 1 (for example see the Table 2.4).

The plot of Entropy CP obtaining the optimal number of splits is show in Figure 2.1. As the Figure 2.1 displayed by holding the smallest misclassification error with relatively small CP value, which in other words prune the classification tree back in order to obtain simpler model without losing much information and with the maximum precision. Therefore, according to the plot with the minimum misclassification error cross up to the top we have the number of splits for the tree is 24. Looking at the Entropy CP table, we see that the best tree has 25 terminal nodes (24 splits) with CP value 0.001111 (see in Table 2.5), based on cross-validation, which it takes the smallest cross validation error (xerror). This subtree is extracted with a call to prune.

For the classification tree which we obtained by using Gini method, it shows the lowest xerror corresponding to the CP value as 0.0006207 with corresponding number of splits as 33 (show in Table 2.4). The same method as we illustrated above

will applied here for obtaining the optimal tree size.

Table 2.4: Gini classification CP Table

| | CP | nsplit | rel error | xerror | xstd |
|---|---|---|---|---|---|
| 1 | 9.0843e-02 | 0 | 1.00000 | 1.00000 | 0.0091280 |
| 2 | 1.2241e-02 | 3 | 0.72747 | 0.73126 | 0.0081941 |
| 3 | 1.0632e-02 | 5 | 0.70299 | 0.71552 | 0.0081273 |
| ... | | | | | |
| 13 | 6.8966e-04 | 31 | 0.62517 | 0.65609 | 0.0078613 |
| 14 | 6.2069e-04 | 33 | 0.62379 | 0.65563 | 0.0078592 |
| 15 | 4.5977e-04 | 39 | 0.61977 | 0.65598 | 0.0078608 |
| ... | | | | | |
| 20 | 0.00000000 | 55 | 0.61402 | 0.66161 | 0.0078870 |

Table 2.5: Entropy classification CP Table

| | CP | nsplit | rel error | xerror | xstd |
|---|---|---|---|---|---|
| 1 | 9.0843e-02 | 0 | 1.00000 | 1.00000 | 0.0091280 |
| 2 | 1.2241e-02 | 3 | 0.72747 | 0.73057 | 0.0081912 |
| 3 | 1.0632e-02 | 5 | 0.70299 | 0.71069 | 0.0081066 |
| ... | | | | | |
| 9 | 1.4559e-03 | 21 | 0.63655 | 0.66172 | 0.0078875 |
| 10 | 1.1111e-03 | 24 | 0.63218 | 0.65874 | 0.0078737 |
| 11 | 1.0345e-03 | 27 | 0.62885 | 0.65908 | 0.0078753 |
| ... | | | | | |
| 20 | 0.00000000 | 55 | 0.61402 | 0.67057 | 0.0079283 |

## 2.3 Results

The final output would have the confusion matrix, which is the evaluation of the performance for the classification model, and it is based on the counts of test records correctly and incorrectly predicted by the model on the test data. The Accuracy=Number of correct predictions/Number of total predictions, and the Error Rate= Number of incorrect predictions/Number of total predictions. The table for Entropy Tree is show in the table below Table 2.6.

As the result show in table it appeared that there is a reasonably good rate for correctly classify the non-poor to be non-poor, however, the accuracy for classifying poor is rather low as about half of the poor would classified as non-poor, therefore it would occur the over classification for the poverty rate. For this reason we try the Gini Tree method and see if we can have any improvement from the Table 2.7.

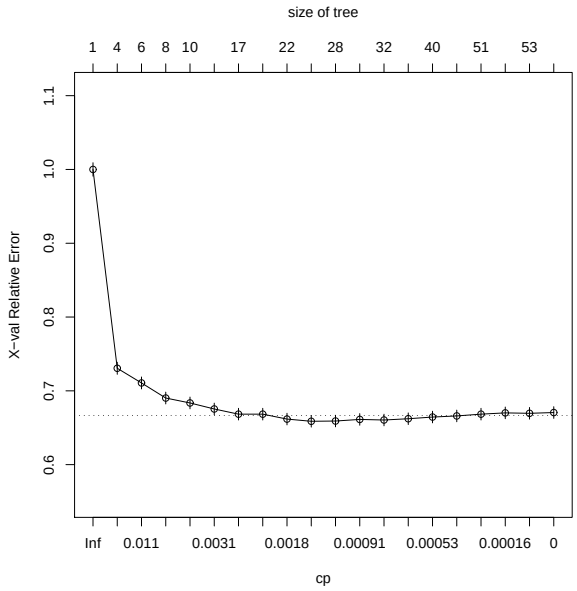


Figure 2.1: Plot for Entropy Tree Size

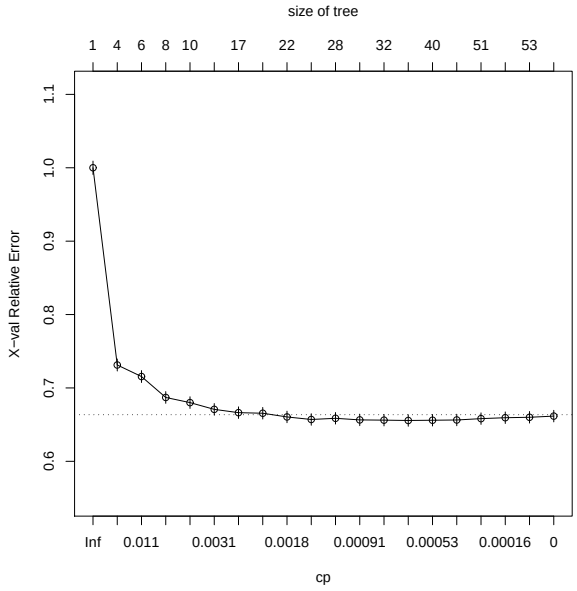


Figure 2.2: Plot for Gini Tree Size

Table 2.6: Confusion Matrix for Entropy method

| | PredPov | |
|---|---|---|
| ActualPov | non-poor | poor |
| non-poor | 0.916079 | 0.08392102 |
| poor | 0.431513 | 0.56848702 |

Table 2.7: Confusion Matrix for Gini method

| | PredPov | |
|---|---|---|
| ActualPov | non-poor | poor |
| non-poor | 0.9204866 | 0.0795134 |
| poor | 0.4440466 | 0.5559534 |

Both trees are basically the same from one to the other, most of the parent nodes and child nodes are the same, even the similar proportion for classification has applied for both trees. Comparing both confusion matrices we obtained from two different methods, the accuracy did not get improved by changing Entropy method into Gini, as we can tell there is still a reasonably large proportion of misclassification for the poor household into non-poor (as the misclassification rate is 0.444 for Gini and 0.432 for Entropy, see Table 2.6 and 2.7).

In addition it is worth mentioning that we have using the proportional balanced data between poor and non-poor, for both training and testing data set, which in other words set the proportions of poor and non-poor to be 27.7% and 72.3% respectively in both training and testing. Tables below are the confusion matrix for Entropy and Gini method Table 2.8 and Table 2.9:

Table 2.8: Confusion Matrix for Entropy method based on balanced data

| | PredPov | |
|---|---|---|
| ActualPov | poor | non-poor |
| non-poor | 0.9108236 | 0.08917643 |
| poor | 0.4252401 | 0.57475995 |

Table 2.9: Confusion Matrix for Gini method based on balanced data

| | PredPov | |
|---|---|---|
| ActualPov | poor | non-poor |
| non-poor | 0.9148453 | 0.08515475 |
| poor | 0.4334705 | 0.56652949 |

It is obvious that there is a slight improvement of the prediction, therefore it seems to be more precise as less proportion of non-poor household been classified as

poor.

Here we have tested our model based on census data for two provinces, the comparison result shown as two sets of plots. In the graph it has symbols as 'Soft' and 'Hard', it means we used different methods to predict the category of each household fall in. 'Soft' has been calculated using posterior probability (which is the probability of falling poor or non-poor).

For getting the estimation we apply the Bernoulli distribution function. The equation follows Bernoulli distribution, the equation show as below.

$$E[y_i] = n_i p_i$$
$$E[X] = \sum p_i n_i$$

where $n_i$ is the family size; $p_i$ is the posterior probability of being poor; $y_i$ represents the number of poor people in household i; and $X$ is the sum of all the poor people in the municipality.

If the total number of people in the municipality is $\sum n_i = n$, then the poverty incidence $P_0$ is given by:

$$E[P_0] = \sum p_i(n_i/n)$$

In the 'soft' poverty calculation, we estimate the poverty incidence by $\sum p_i n_i/n$. The 'hard' classification, each $p_i$ is replaced by 0 if $p_i < 0.5$ or 1 if $p_i \geq 0.5$. To illustrate the figures shown below, the dots in plots is classified poor household cumulated in municipality level. Moreover, the symbol 'Hard' represents the cumulate numbers of households which has classified as poor into a municipality level. Meanwhile, 'Soft' means the cumulative posterior probabilities have been applied in the calculation steps.

The graph in Figure 2.3 shows poverty incidences obtained using two classification methods (Gini and Entropy), comparing it with the results obtained using small area estimation (in the whole thesis when we metion small area estimation method we mean small area estimation by the ELL method) in Province One (There are 17 municipalities in Province One). Solid line is a 45 degree line right across the plot, and the dotted line is the regression line we obtained using small area estimation versus the result got with entropy or Gini method. Looking at the plots of

Entropy versus Small Area Estimation, the Entropy Soft results tend to be lower than the Small Area Estimations, taking the first point locate at the left corner of the plot as an example, with the same municipality different predictions has been obtained using different methods, the small are estimation has the value close to 0.2, as Entropy using soft way of calculation it around at 0.1. Such an circumstances existing at most of the municipalities, therefore most of the points are locate above the solid line, however there are two exceptions, which are extremely high for Entropy (it could round to 0.8 and 1), and the Small Area Estimation results are only 0.3 and 0.4 respectively. The second plot shows two series of results obtained from two methods are quite similar to one and another, as they all line close to the 45 degree line. Still the Small Area Estimation results are slightly higher than the Entropy one. The comparison between Gini method using hard calculation method and Small Area Estimation the hard calculation shows about half of the predictions are above the line and the other half are below the 45 degree line, the lower half as its above the line which means the SAE results is higher and the upper half locate below the line means the SAE results is lower than the results we obtained suing Gini method. A different scenario exist in the comparison of soft Gini and SAE, all of the results are above the 45 degree line, therefore here the SAE is having higher predictions on poor population.

Again the comparison has done between SAE result and classification tree method for Province Two (There are 47 municipalities in Province Two) Figure 2.4. Two methods applied in classification tree: Gini and Entropy. Same as Figure 2.3, solid line is a 45 degree line across the plot, the dotted line is formed up by the regression between SAE and classification outcomes. The top set of plots has shown symbols of 'Soft' and 'Hard' which represents different calculation methods using Entropy information criteria; and the bottom set of plots are using Gini information criteria. The majority of points are locate above the 45 degree line for Entropy criteria using Soft calculation method, only three dots are below the line, in the other words, most of estimations tend to be higher as obtained using SAE method and lower for Entropy one. A similar conclusion came up for the hard calculation with Entropy criteria. For Gini criteria, a small proportion of points has fall below the solid line, and a contrast situation exists in the municipalities with lower percentage of poor population (lower estimations in 'GiniHard' and a relatively higher estimation in SAE for the same municipality). As the estimations gets higher two results from two methods are getting closer. The last plot shows most points are above the solid line, hence, most of it having higher estimations using SAE method.

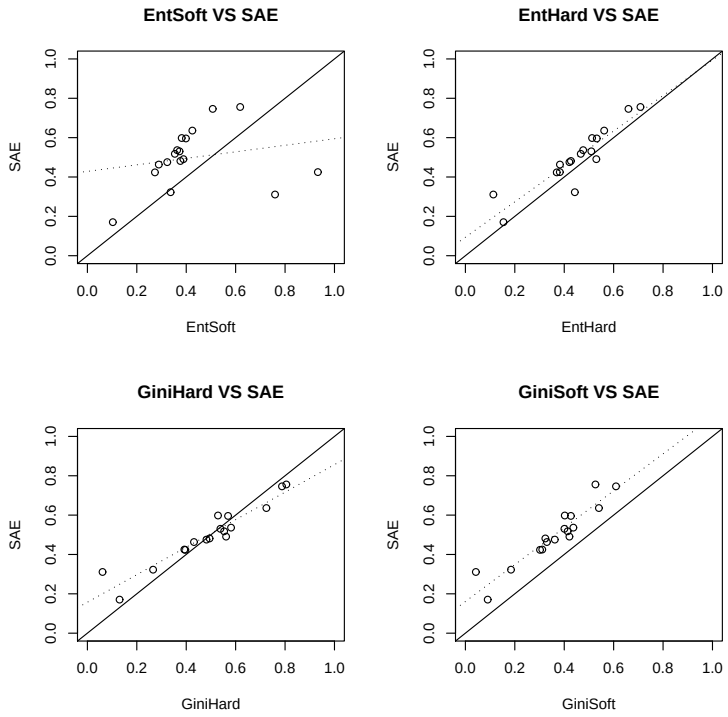


Figure 2.3: Poverty Incidences obtained from two classification methods Vs Small Area Estimation result(Prediction on Province One)

As described in previous paragraph, treat the small area estimation result as the standard answer, then the best combination is using Gini information criteria with 'hard' calculation method. To a degree the estimation obtained from the best combination does relatively consist with the SAE results.

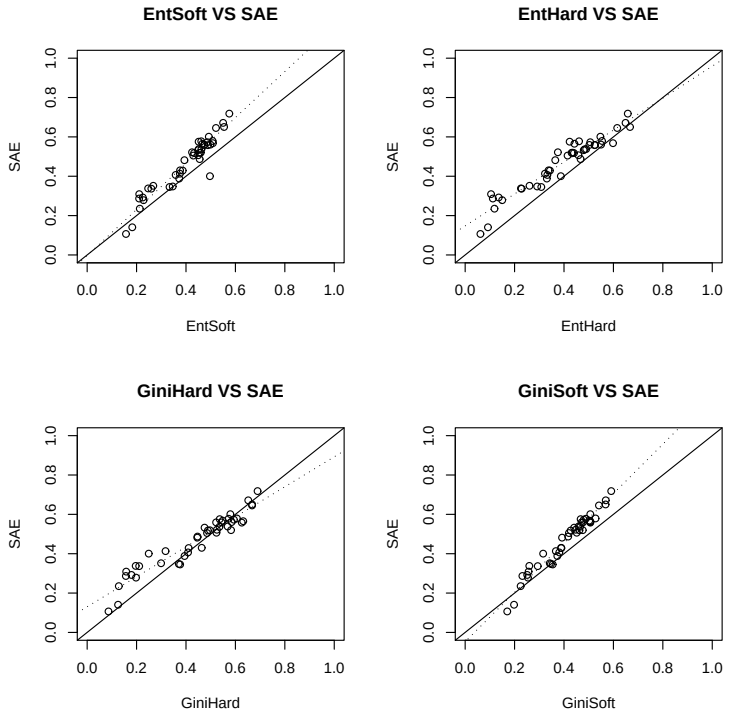


Figure 2.4: Poverty Incidences obtained from two classification methods Vs Small Area Estimation result(Prediction on Province Two)

# Chapter 3

# Classification Trees with Weighted Data

The design of survey been described in Chapter One, the sample data has been selected out of the population using known probabilities so that one may estimate something about the whole population. The probability of a household being randomly selected for the survey data is considered as survey weight, in other words the weight is inverse of the probability (Cole & Hernan, 2008). Therefore, in this chapter we incorporate survey weight into building the classification trees .

## 3.1 Weighting by Duplication

The motivation for incorporating weights is that it should reduce bias. As the evaluation results of model shown in the previous chapter, the precision of classification for non-poor is 0.9 (on average) whereas the precision classification on poor has only about 52%. There may be bias from not using the survey weights. They affect the analysis in two ways. First is to incorporating weights into building the tree; second is to incorporate the weights into the calculation of confusion matrices. Here we will apply the method based on Gini information gain criteria, as the confusion matrix showed more precise classification than the Entropy method. Check on the comparison which been done at the end of Second Chapter.

The original data, collected proportionally out of population, has two relevant survey weight variables: 'sswgtpp' and 'sswgthh', the variable 'sswgthh' stands for the proportion of a household has been selected out of a region, say if one household been selected out of hundred households then this one household is representing for one hundred households. The connection between 'sswgthh' and 'sswgtpp' can be

illustrated using equation:

$$\textit{'sswgtpp'} = \textit{'sswgthh'} \times \textit{number of people in a household} \tag{3.1}$$

which in other words, the survey weight in household level times number of people in a household and it become the survey weight in per-person level. The household survey weight is for mapping out the proportion of poor in household level, whereas the survey weight for per-person level is applied for getting the proportion of poor in per-person level. Here we try incorporating weights into the analysis, using a duplication method based on one of these two variables.

As the previous chapter built classification tree models based on the un-weighted data, the corresponding confusion matrix shows highly precise prediction on non-poor (which is 91.5%), on the other hand, a less precise classification for poor (there is only 56.4% of poor been correctly classified). We now want to see if using weight will improve the classification for poor.

### 3.1.1 Duplication with household Weights (Gini and Entropy)

Each household in the data may represent a number of households in a region, therefore we adjust the data by duplicating each household a number of times, roughly in proportion to its weight. The whole data set is divided into five parts according to the ascending order of variable 'sswgthh', we calculate the mean weight for each part, and divide each mean by the mean obtained from the first part, the times of duplication for each part is carried out according to this number which it rounded to the nearest integer first. Splitting the duplicated data set into two parts: 'training' (80%) and 'testing' (20%), again build the model based on training data set and evaluate it in testing data. To be specific an example is illustrated here. Before duplication we divided the data into five parts, ascending the data with respect to 'sswgthh' variable. The times of duplication applied to each part is obtained by calculating the mean of variable 'sswgthh' in part 1 and divided all the means by the mean from part 1, giving $\mu_1$=1, $\mu_2$=1.7, $\mu_3$=2.5, $\mu_4$=3.6, $\mu_5$=5.9. The round number of duplications is used. Hence, part 1 stays unchanged, part 2 will be duplicate two times, three times in part 3, four in part 4, at last, part 5 will have six

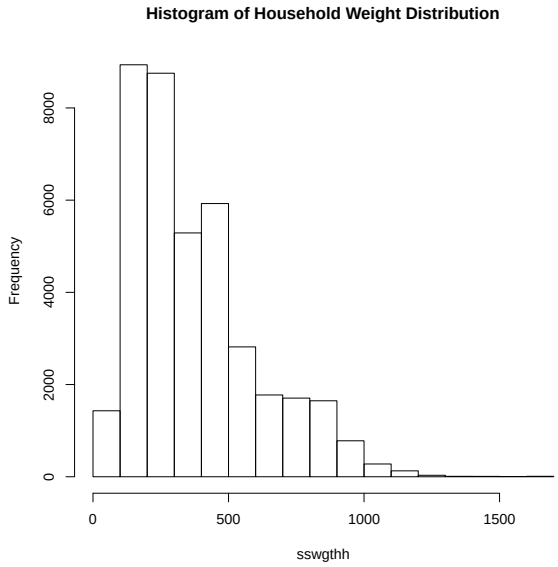


Figure 3.1: Histogram of Household Weight Distribution

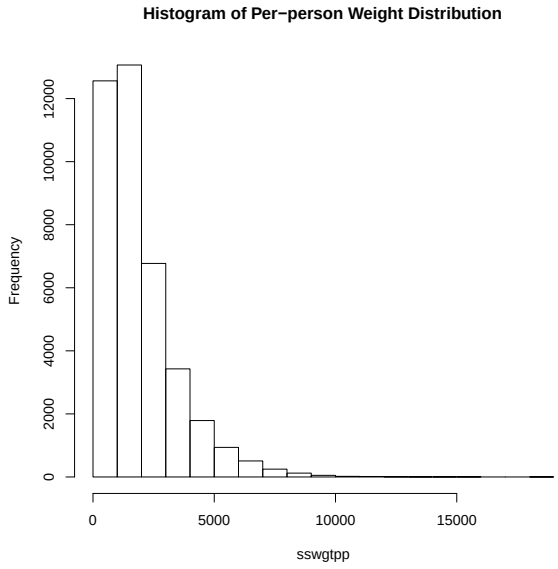


Figure 3.2: Histogram of Per-person Weight Distribution

times of duplication. The histogram of household and per-person weight distribution showed in Figure 3.1 and 3.2. The figure for most of weighted households locate at range 200 to 500, and survey weight based on person concentrates at 1000 to 3000.

Below presents the CP table (Table 3.1) for duplicated data with the data parted into five parts based on variable 'sswgthh'. The minimum split used in the tree is 800, in other words, the minimum number of observations that must exist in a node

is 800, whereas for an un-weighted tree the minimum is set to 200. The reason for that is due to the duplication of the data. The average number for times that each part has been duplicated we approximate to 4. The number of split here is 22, which is a relatively large number compared to the un-weighted Gini tree. The confusion matrix shown in Table 3.2 has the accuracy of classification in non-poor is about 90% and poor is only 58%. Obviously this result is quite close to the one we obtained from the un-weighted data.

Table 3.1: The data duplicated based on variable sswgthh using Gini method classification CP Table (five parts)

| | CP | nsplit | rel error | xerror | xstd |
|---|---|---|---|---|---|
| 1 | 0.09391669 | 0 | 1.00000 | 1.00000 | 0.0050287 |
| 2 | 0.01067422 | 3 | 0.71825 | 0.71920 | 0.0044925 |
| 3 | 0.01062146 | 5 | 0.69690 | 0.71041 | 0.0044719 |
| ... | | | | | |
| 21 | 0.00022275 | 43 | 0.60964 | 0.62136 | 0.0042471 |
| 22 | 0.00015241 | 46 | 0.60898 | 0.62195 | 0.0042487 |
| 23 | 0.00000000 | 49 | 0.60852 | 0.62076 | 0.0042455 |

Table 3.2: Confusion Matrix for Gini method (duplicate based on variable sswgthh five parts)

| | PredPov | |
|---|---|---|
| ActualPov | poor | non-poor |
| poor | 0.58128161 | 0.41871839 |
| non-poor | 0.09762211 | 0.9023779 |

Table 3.3: Confusion Matrix for Entropy method (duplicate based on variable sswgthh five parts)

| | PredPov | |
|---|---|---|
| ActualPov | poor | non-poor |
| poor | 0.57117278 | 0.4288272 |
| non-poor | 0.09150212 | 0.9084979 |

Based on the same duplication data we have applied with the Entropy information criteria in order to know which criteria seems to be better. Therefore, for comparison purpose we provide the confusion matrix shown in Table 3.3, it is clear that Gini confusion matrix (Table 3.2) perform better as the correct classification of the poor and non-poor is slightly better.

### 3.1.2 Duplication with per-person Weights (Gini and Entropy)

In this study we have tried several different ways of applying duplication. A similar procedure we have applied to duplicate the data according to variable 'sswgtpp' where this variable has per person basis. The data has been sorted in ascending order using to 'sswgtpp' variable. Again divide the data into five parts, calculate mean of variable 'sswgtpp' for each part, and the times of duplication is been done according to its mean. The CP table shown in Table 3.4 is obtained using Gini classification method, more branches been split here, 30 splits of the classification tree is recognisee as the best split number, it hits the minimum cross validation error. Average rate of duplication here is 3. The corresponding confusion matrix obtained show in Table 3.9 which has the same conclusion where the high percentage on true positives (the accurate classification on non-poor) and relative low classification rate on the poor.

The Table 3.10 is obtained by Entropy method with duplication based on 'sswgtpp' variable, a slightly better classification of the non-poor shown in the confusion matrix, meanwhile, a better classification of the poor in Gini method confusion matrix.

Table 3.4: The data duplicate based on variable sswgtpp using Gini method classification CP Table (five parts)

| | CP | nsplit | rel error | xerror | xstd |
|---|---|---|---|---|---|
| 1 | 1.2329e-01 | 0 | 1.00000 | 1.00000 | 0.0039978 |
| 2 | 9.5268e-02 | 2 | 0.75342 | 0.75423 | 0.0036769 |
| 3 | 1.1084e-02 | 3 | 0.65815 | 0.66164 | 0.0035135 |
| ... | | | | | |
| 29 | 1.1944e-04 | 72 | 0.53195 | 0.55202 | 0.0032829 |
| 30 | 8.7591e-05 | 73 | 0.53183 | 0.55164 | 0.0032821 |
| 31 | 4.7777e-05 | 76 | 0.53157 | 0.55197 | 0.0032828 |
| ... | | | | | |
| 33 | 0.00000000 | 81 | 0.53143 | 0.55264 | 0.0032844 |

Instead of sorting data into five parts we now trying to part the data into ten parts. The reason for increasing the number of parting in the data is we want to know whether by having more specified and detailed duplication number would bring a more precise classification of non-poor, since the prediction of the non-poor is the main focus of this study. Table 3.7 and 3.8 showing both Gini and Entropy information criteria confusion matrices respectively. It is hard to tell which one is

Table 3.5: Confusion Matrix for Gini method (duplicate data based on variable sswgtpp five parts)

| | PredPov | |
|---|---|---|
| ActualPov | poor | non-poor |
| poor | 0.59131603 | 0.4086840 |
| non-poor | 0.08832863 | 0.9116714 |

Table 3.6: Confusion Matrix for Entropy method (duplicate data based on variable sswgtpp five parts)

| | PredPov | |
|---|---|---|
| ActualPov | poor | non-poor |
| poor | 0.58639212 | 0.4136079 |
| non-poor | 0.08638928 | 0.9136107 |

better than the other as the result of precision classification for poor is very close to each other.

Table 3.7: Confusion Matrix for Gini method (duplicate data based on variable sswgtpp ten parts)

| | PredPov | |
|---|---|---|
| ActualPov | poor | non-poor |
| poor | 0.57743957 | 0.4225604 |
| non-poor | 0.09661495 | 0.9033850 |

Table 3.8: Confusion Matrix for Entropy method (duplicate data based on variable sswgtpp ten parts)

| | PredPov | |
|---|---|---|
| ActualPov | poor | non-poor |
| poor | 0.58281110 | 0.4171889 |
| non-poor | 0.09432299 | 0.9056770 |

### 3.1.3 Prediction Comparison between Duplication method and SAE results

For the comparison purposes both weighted and un-weighted model have been used to predict municipal level poverty in Provinces One and Two.

The comparison has been done between the small-area estimation and Gini classification methods, moreover the weighting has been done using 'sswgthh'. The comparison is based on the prediction obtained using different models. Figure 3.3 displays the comparison for Province One, with the 45 degree solid line across the plot and regression line shown as a dotted line in the plot. The top left plot showing the Entropy method applied on un-weighted data, it's clear that most of dots are above the 45 degree line, in the other words most of estimation in municipal level which using small area method has higher proportion of prediction in poor, the prediction proportion of poor concentrates at the range 0.4 to 0.8. The top right plot is the comparison between SAE and Entropy method on weighted data (the weighted data is obtained by duplication with household weight which part data into five parts) showing a similar conclusion as most of dots are above the 45 degree line, however comparing it with the method on un-weighted data the dots seems more spread out. The bottom two plots have been formed based on weighted and un-weighted data using Gini classification method, the bottom left plot shows still most of dots are below the solid line, plot shown on the weighted data more than half of the dots are below the solid line, even though it has been points below the solid line than the Gini method on un-weighted data, thereby here the conclusion can be drawn that using Gini method on weighted data has prediction results relatively lower than the small-area estimation.

In Figure 3.4 the comparison has been done for Province Two. Both Gini and Entropy methods have been used incorporating both weighted and un-weighted data. The top left plot illustrates the comparison between Entropy on un-weighted data and small-area estimation, most of dots are above the line in a relatively concentrated way, dots on the top right plot are spreading out and more points are located below the line. Plots on the bottom are the comparison using Gini method on un-weighted and weighted data, similar spreading of dots is displayed on two plots. Hence, to sum up, the predictions obtained using SAE have higher incidences than the Gini method.

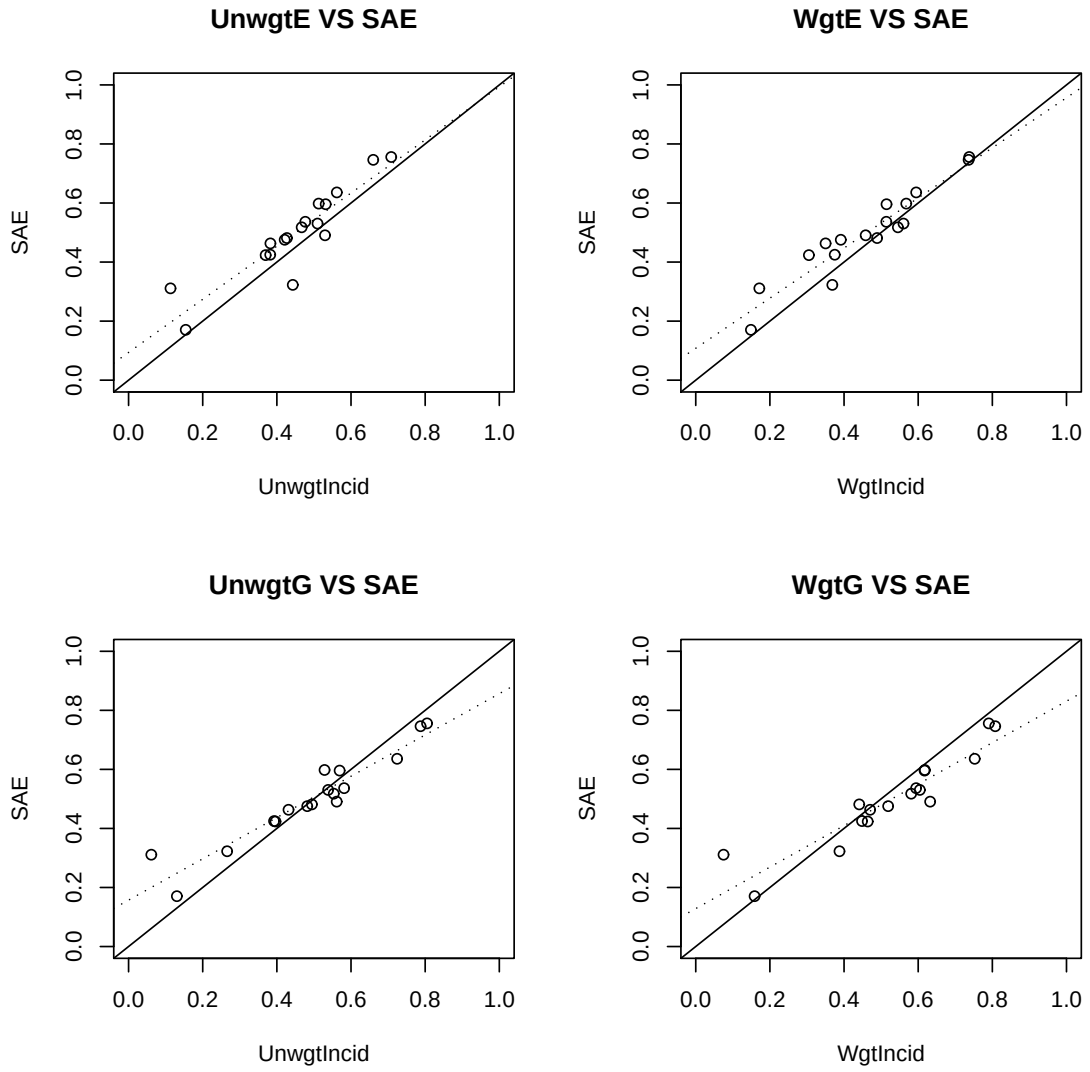


Figure 3.3: Poverty Incidences obtained from two classification methods Vs Small Area Estimation result(Prediction on Province One with duplicated data base)

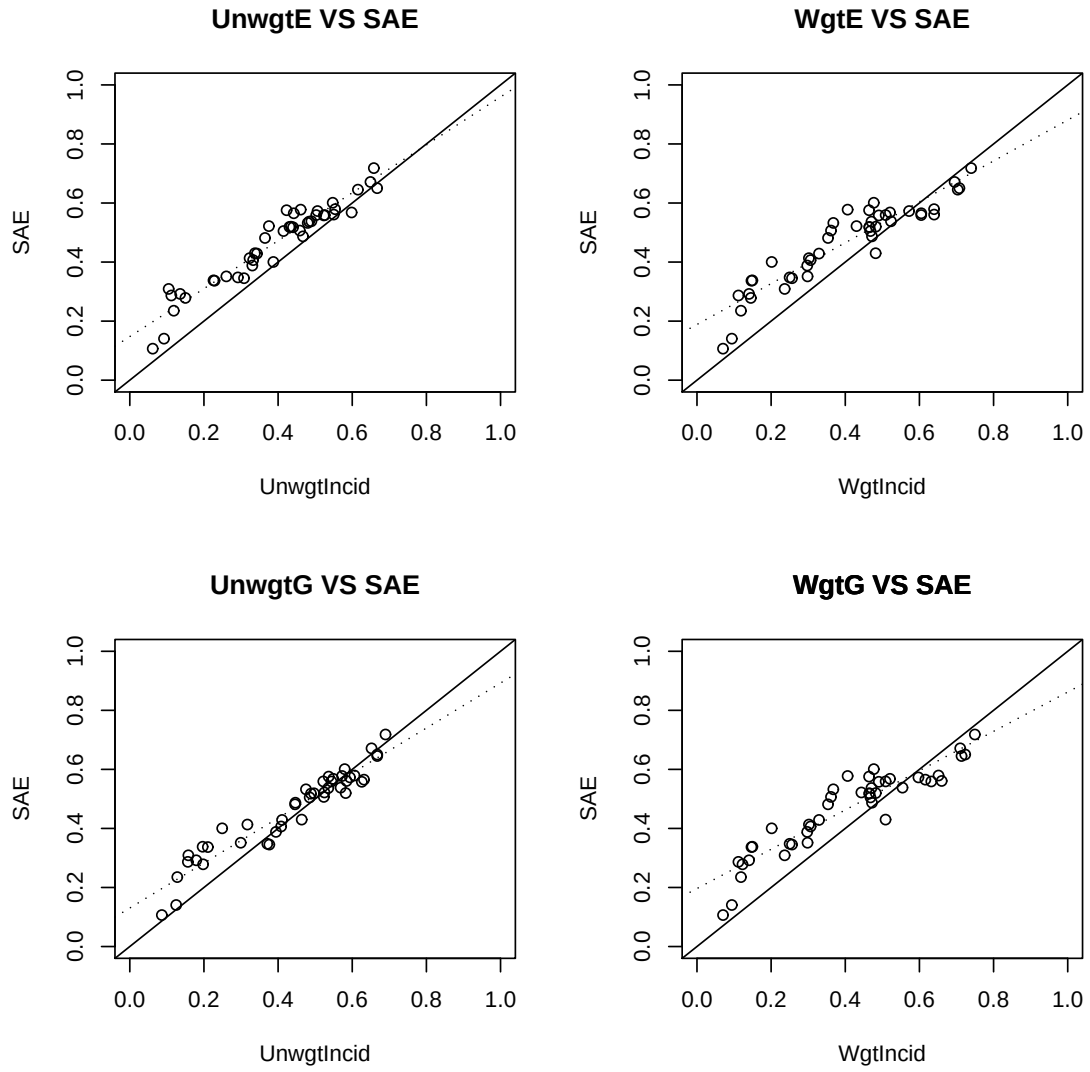


Figure 3.4: Poverty Incidences obtained from two classification methods Vs Small Area Estimation result(Prediction on Province Two with duplicated data base)

Overall, the estimation results from weighted data tend to spread out comparing with the un-weighted plots. The un-weighted plots showing the classification tree method with un-weighted data has outcomes tend to be under estimate comparing with the small area estimation result. A slight improvement showing in the plots of weighted data, in that the better method would be the one based on weighted data with Gini information criteria (as mentioned in Chapter 2, Gini confusion matrix showing a slight better prediction in poor).

## 3.2 Weights Argument in RPART

According to the journal that introduce the recursive partitioning using the RPART routines the weight argument is not supported in the program (Therneau & Atkinson, 1997). However, after finish running the duplication method we found in the weights argument is functioning in the latest RPART program. In the weights argument the user can specify a weight for each observation. The weight are included in the calculation of the information criteria. Both Entropy and Gini been applied here along with the weights argument. Both survey weights have been tried in the weights argument. As explained earlier for the survey weights, it is known as the probability of a household and or person being selected out of a region, for example if a household (person) been selected out of hundred households (people) then this household (person) represents a hundred households (people).

Table 3.9 and 3.10 shows the confusion matrices obtained by using weights argument with respect to Gini and Entropy methods. The survey weight used to form both tables is 'sswgtpp', the reason for using this variable is we want to specifically identify the poverty at per-person level instead of household. Showing the confusion matrices using both information criteria, we can see that, a slightly higher correct classification of poor is shown in Gini method and a higher correct classification of non-poor in Entropy method. The aim in this study is to obtain the classification of the poor, hence the correct classification of the poors is preferred.

Table 3.9: Confusion Matrix for Gini method (weights argument using variable sswgtpp)

| | PredPov | |
|---|---|---|
| ActualPov | poor | non-poor |
| poor | 0.56445837 | 0.4355416 |
| non-poor | 0.08515515 | 0.9148449 |

Table 3.10: Confusion Matrix for Entropy method (weights argument using variable sswgtpp)

| | PredPov | |
|---|---|---|
| ActualPov | poor | non-poor |
| poor | 0.56401074 | 0.4359893 |
| non-poor | 0.08444993 | 0.9155501 |

## 3.3 Setting up priors and obtaining ROC curve

A Receiver Operating Characteristic (or simply named as ROC) curve, is a graphical plot of the sensitivity (true positive rate) against false positive rate (1 − specificity), for a binary classifier system as its discrimination threshold is varied (Bradley, 1997). The ROC curve is an analysis tool which is used to select possible optimal models and delete other ones independently from the cost context or the classification distribution. Classification model is a technique for sorting instances into a certain class.Here we consider a two-class prediction problem; the outcomes are labeled as either poor (positive) or non-poor (negative). There are four possible outcomes from a binary classifier, which are applied for building the ROC curve. If the outcome from a prediction is poor and the actual group is also poor, then it is called as true positive; however if the actual group is non-poor then it is said to be a false positive. Conversely, a true negative has occurred when both the prediction outcome and the actual value are non-poor, and false negative is when the prediction outcome is non-poor while the actual is poor.

The proportions of false positives and negatives can be affected by giving a prior distribution for the classes. This prior is used in calculating the information and making decisions about poor and non-poor. Figure 3.5 shows all possible combinations of sensitivity and (1 − specificity) obtained by differing priors (the prior of poor set the range from 0.15 to 0.95, meanwhile the prior of non-poor is calculated as 1-prior of poor). Models have been built based on the un-weighted data set using the Gini information criteria. The reason for using this particular method is that in the confusion matrices shown above, the ones using Gini method seems doing a

slightly better job on classification. The black dot displayed in the plot is the sensitivity and (1 − specificity) obtained by setting the prior as the proportion of poor and non-poor households in the data set. The same method is applied for getting the plot show in Figure 3.6, the only difference here is the per-person based survey weight has been used in the data set. Notice that the black dot shown up in the plot is obtained by setting prior as proportions of poor and non-poor on per-person basis. A comparison between two plots is shown in Figure 3.7, two lines have been formed up by connecting dots together with respect to different sets of sensitivity and (1 − specificity), where the solid line is obtained from original data set and the dashed line is based on duplicated data set. The little circle in the plot showing a particular set of sensitivity and (1 − specificity) obtained by setting the prior according to the proportion of poor and non-poor household in the original data set (which is 28% of poor and 72% of poor). Once again the star symbol is obtained by setting prior regard to proportion of poor and non-poor people in the duplicated data (34.15% of poor and 65.85% of non-poor). In the way that, the prior combination as 34.15% versus 65.85% has relatively better percentage of accurate prediction rate without loosing much information.

In Figure 3.7, the two ROC curves seem similar as they almost overlap with each other. Notice that there are two symbols (one in star shape, the basis prior probability here is per-person survey weight, and the little circle has the base prior probability of poor and non-poor on un-weighted data set) on the graph, the eye catching point here is these two symbols the little circle locating at a relatively higher position than the star shaped one. The little circle shows higher true positive rate as well as the false positive rate, in which the prior is set according to household on un-weighted data.

## 3.4 Comparison between Confusion Matrices

Before comparing confusion matrices presented in tables below, first let us know how they have been calculated. There are three methods for calculating the confusion matrices. The un-weighted one uses the un-weighted test data for getting the number of households classified as poor or non-poor, cross classifies it with the real poverty property and then we can calculate the probabilities of correct and incorrect classification. Second, two kinds of survey weights can be incorporated into the calculation of confusion matrices, one is on household level and the other is per-person level. The calculation steps are the same as stated above except when calculating

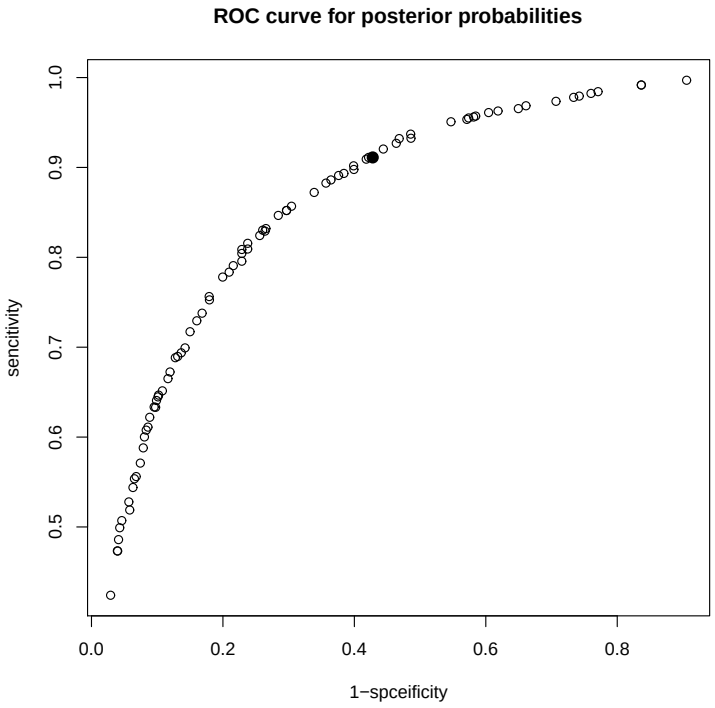


Figure 3.5: The ROC curve obtained from unweighted data

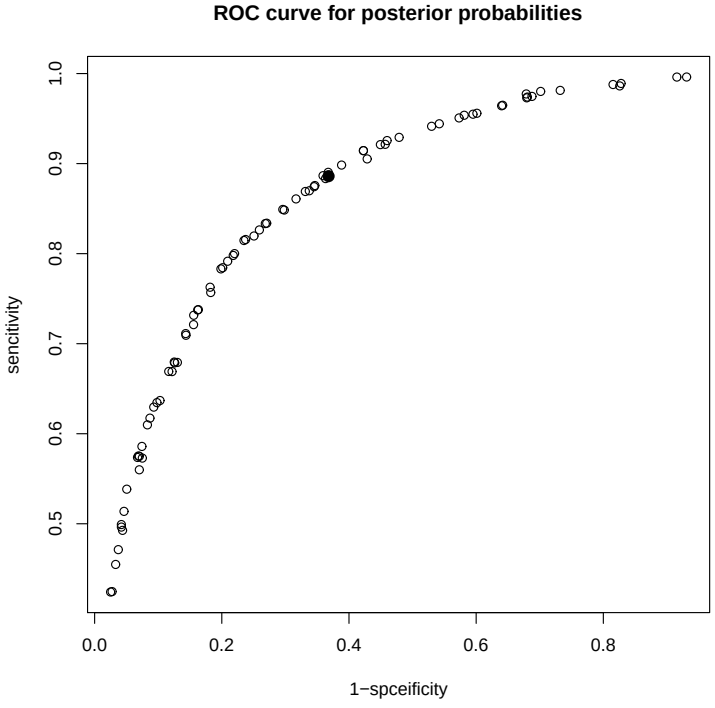


Figure 3.6: The ROC curve obtained from data weighted on per-person scale

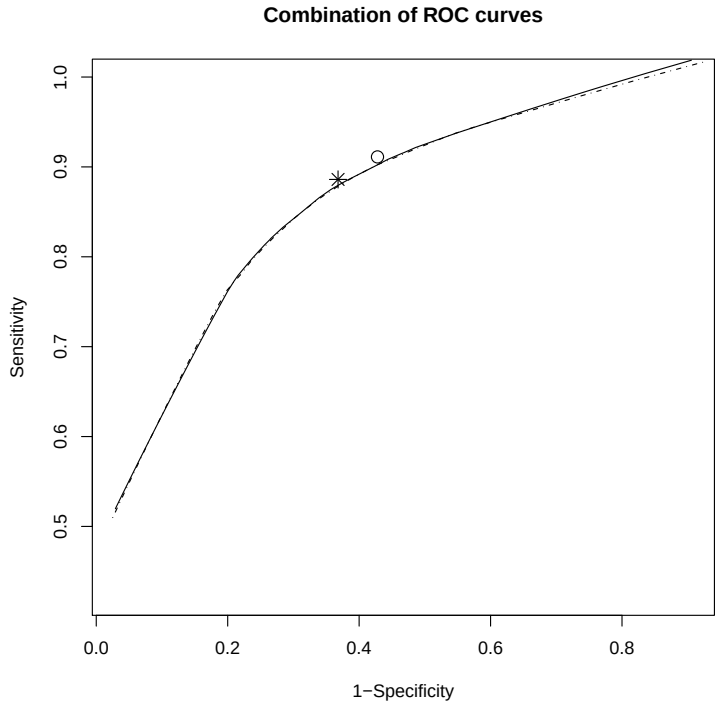


Figure 3.7: comparison between unweighted ROC curve and weighted ROC curve

the confusion matrices instead of using classification based on un-weighted original training data, we apply the variables 'sswgtpp' or 'sswgthh' as it shows each household in the training data represents for a number of households or people. We then calculate corresponding correct and miss classification rates on basis of these two variables. In the five tables presented below, each table has its own way of tree building method however they share the same confusion matrix calculation methods. Table 3.11 building the tree based on un-weighted data; Table 3.12 building the tree based on household duplication data in which it been parted in 5 parts; Table 3.13 tree build based on per-person duplication data in which it parts data into 5 parts; again, Table 3.14 building the tree based on per-person duplication data in which it parts data into 5 parts; at last,Table 3.15 trees build based on per-person weights argument. Five tables share the three common confusion matrix calculation method, which are called un-weighted, 'sswgthh' and 'sswgtpp'. One thing need to point out is that all of confusion matrices been build using Gini criteria.

Compare all the confusion matrices below obtained by incorporating the survey weights in various ways. In total there are five sets of confusion matrices, the first set showing large true positive rate in all three different methods of calculating the confusion matrices.

Tables stated below showing different sets of confusion matrices obtained based on different data set and incorporating with survey weights. Overall, the precision of prediction for non-poor is very high as its round at 0.89; on the other hand, the precise classification of the poor reaches only about 0.6. As the purpose of knowing the poverty property of households is to be able to allocate aid sufficiently and precisely therefore we expect to have high classification precisions on poor instead of non-poor. The aid is expected to go to the most needed ones first therefore about 60% percision of classification of the poors may be acceptable. As the aid is allocated at municipal level, therefore the error rate at household level is not as important as the municipal-level accuracy.

Five sets of confusion matrices listed below, each has contained three different ways in calculating the matrices, and each table which one is based on un-weighted original data set, one is based on weight of household level and last is survey weight based on the per-person level.

The comparison of the un-weighted confusion matrix method of five tree building methods, Table 3.11 shows higher correct classification of non-poor (0.915) and the lowest value exist in Table 3.14; meanwhile, the higher correct classification of poor exist in Table 3.13 and the lowest one is in Table 3.11. The 'sswgthh' confusion matrices between 5 tables average value for the correct classification of non-poor is 0.885, and the average percentage for the poor is 0.582. At last, the 'sswgtpp' confusion matrices giving the higher proportion of non-poor in Table 3.11 (which is 0.886) and poor in the Table 3.13 (0.654); on the other hand, the lowest proportion of non-poor is 0.872 (Table 3.14) and poor is 0.639 (Table 3.11). Overall, the highest proportion of non-poor 0.915 and poor is 0.654; the lowest value of non-poor is 0.872 and poor is 0.564. To conclude, in consideration of the trace of trade off between accurate prediction of poor and non-poor, the better prediction would be using per-person scaled survey weight, as comparative having high accurate prediction of non-poor meanwhile the accurate prediction of poor was not too low, check on the Table 3.13.

Table 3.11: Confusion Matrices obtained from original data set (where the NP(P) is non-poor predicted; NP(A) is non-poor actual; P(P) is poor predicted and P(A) is poor actual)

| Gini Tree Building | | | | |
|---|---|---|---|---|
| | | | un-weighted | |
| | | | NP(P) | P(P) |
| Confusion Matrix | un-weighted | NP(A) | 0.915 | 0.085 |
| | | P(A) | 0.436 | 0.564 |
| | sswgthh | NP(A) | 0.886 | 0.114 |
| | | P(A) | 0.429 | 0.571 |
| | sswgtpp | NP(A) | 0.886 | 0.114 |
| | | P(A) | 0.361 | 0.639 |

Table 3.12: Confusion Matrices obtained by duplicate data based on sswgthh (parts for five)

| Gini Tree Building | | | | |
|---|---|---|---|---|
| | | | Duplicated-hh-5 | |
| | | | NP(P) | P(P) |
| Confusion Matrix | un-weighted | NP(A) | 0.908 | 0.092 |
| | | P(A) | 0.429 | 0.571 |
| | sswgthh | NP(A) | 0.878 | 0.122 |
| | | P(A) | 0.419 | 0.581 |
| | sswgtpp | NP(A) | 0.878 | 0.122 |
| | | P(A) | 0.358 | 0.642 |

Table 3.13: Confusion Matrices obtained by duplicate data based on sswgtpp (parts for five)

| Gini Tree Building | | | | |
|---|---|---|---|---|
| | | | Duplicated-pp-5 | |
| | | | NP(P) | P(P) |
| Confusion Matrix | un-weighted | NP(A) | 0.911 | 0.089 |
| | | P(A) | 0.409 | 0.591 |
| | sswgthh | NP(A) | 0.909 | 0.091 |
| | | P(A) | 0.406 | 0.594 |
| | sswgtpp | NP(A) | 0.885 | 0.115 |
| | | P(A) | 0.346 | 0.654 |

Table 3.14: Confusion Matrices obtained by duplicate data based on sswgtpp (parts for ten)

| Gini Tree Building | | | | |
|---|---|---|---|---|
| | | | Duplicated-pp-10 | |
| | | | NP(P) | P(P) |
| Confusion Matrix | un-weighted | NP(A) | 0.903 | 0.097 |
| | | P(A) | 0.423 | 0.577 |
| | sswgthh | NP(A) | 0.872 | 0.128 |
| | | P(A) | 0.415 | 0.585 |
| | sswgtpp | NP(A) | 0.872 | 0.128 |
| | | P(A) | 0.36 | 0.64 |

Table 3.15: Confusion Matrices obtained Using Weights Argument

| Gini Tree Building | | | | |
|---|---|---|---|---|
| | | | Weights argument | |
| | | | NP(P) | P(P) |
| Confusion Matrix | un-weighted | NP(A) | 0.91 | 0.09 |
| | | P(A) | 0.42 | 0.58 |
| | sswgthh | NP(A) | 0.88 | 0.12 |
| | | P(A) | 0.421 | 0.579 |
| | sswgtpp | NP(A) | 0.88 | 0.12 |
| | | P(A) | 0.357 | 0.643 |

# Chapter 4

# Estimating the Standard Errors

This chapter demonstrates method of calculating the standard errors for classification tree model. There are three approaches we attempted to compute the standard errors using different ways.

## 4.1 Calculation Method without consider the Municipal Variances

### 4.1.1 Standard Error Calculation Method One

To begin with, we calculated the standard errors based on the Bernoulli distribution which is a discrete probability distribution, it takes value 1 with success probability $p$ and value 0 with failure probability $q$ = 1-$p$ (Bain & Engelhardt, 2000), so the equation can be expressed as equation 4.1. The reason for using this method is we have to include the variance component in standard errors.

$$f(k;p) = p^k \times (1-p)^{1-k}, \qquad k \in \{0,1\} \tag{4.1}$$

Since the prediction is done for households and each household has a different number of family members. The predicted number of poor people for household $i$ is given by:

$$\begin{aligned} Y_i &\sim Bernoulli(P_i)n_i \\ E[Y_i] &= n_i P_i \\ V[Y_i] &= n_i^2 P_i(1-P_i) \end{aligned}$$

where $P_i$ is the posterior probability of a household being classified as poor, $n_i$

represents family size. The equations above show the expected value and variances for every household.

The next set of equations show the predicted total of households identified as poor in a municipality, in that the summation method has been implemented here, with the equations for expected value and variance:

$$X = \sum Y_i$$
$$E[X] = \sum P_i n_i$$
$$V[X] = \sum n_i^2 P_i(1 - P_i)$$

Finally to identify the proportion of poor in a municipality and obtain the standard error, we use the equations below:

$$P_0 = \frac{X}{\sum n_i}$$
$$E[P_0] = \frac{\sum n_i P_i}{\sum n_i}$$
$$V[P_0] = \frac{\sum n_i^2 P_i(1 - P_i)}{(\sum n_i)^2}$$
$$SE[P_0] = \sqrt{\frac{\sum n_i^2 P_i(1 - P_i)}{(\sum n_i)^2}}$$

The above equations show the cumulative way of calculation at municipal level. Finally we obtain the standard errors in an accumulative way of calculation for each municipality. The reason for obtain the standard errors in municipal level is first we want to attain the municipal variance as the municipal difference exist between municipalities; second, in use of municipal level data we would have reasonable size data for calculating the standard errors.

We employed the calculation method we have illustrated above to produce the predictions and standard errors for Province One and Two. Tables below display both prediction results attained by using methods of classification and small area estimation. First there is some general information about these two provinces. There are 17 municipalities in Province One (see Table 4.1), the total population is 464473, in which the smallest municipality size is 14927 (which is the fifth municipality), and the largest municipal population size is 59759 (the seventh municipality). Meanwhile, Province Two is a larger province which has 44 municipalities and a total population of 1860728. In this province, the largest municipality is 348461 (munic-

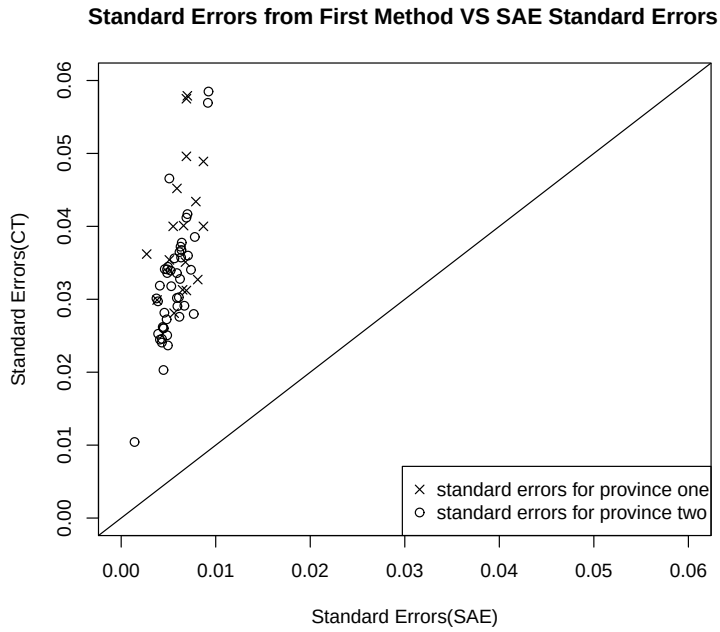


Figure 4.1: Standard Errors obtained using the First calculation method VS SAE Standard Errors for combining both Provinces

ipality 22) and the smallest one is 11765 (municipality 10).

The proportion of poor at municipal level obtained using the method of classification tree is larger than the ones from small area estimation method, in other words, a larger proportion of poor been estimated when using classification tree. After comparing the two estimation results let's now focus on the standard errors. Initially, we calculated the standard errors for Province One however the result was not optimistic as they were too small by comparing with the SAE standard errors. As the size of this province is quite small, it might have an effect on the estimation results therefore we employed the same procedure on Province Two as it is larger than Province One. Nevertheless, standard errors for Province Two is still less than 0.01. The standard errors calculated for classification tree method using the Bernoulli distribution function are way too small as the average values are 0.006 and 0.005 corresponding to Province One and Two respectively. Check on the Tables below and corresponding graph been plotted show in Figure 4.8. Where, '$P_0$soft': is calculated by taking means of all the prior probabilities in municipal level. '$P_0$hard': is calculated by taking means of all the hard classifications (1 if it's poor, 0 if it's non-poor) in municipal level.

Table 4.1: Results from Bernoulli function for Province One

| MunCode | People | Poor People | V(X) | E(X) | $P_0$hard | $V(P_0)$ | $P_0$soft | SE($P_0$) | $(P_0)$sae | SE(sae) |
|---|---|---|---|---|---|---|---|---|---|---|
| 60401 | 21678 | 12171 | 22308.16 | 9127.402 | 0.5614 | 4.75E-05 | 0.421 | 0.0069 | 0.4907 | 0.0496 |
| 60402 | 22606 | 16376 | 24431.65 | 12213.82 | 0.7244 | 4.78E-05 | 0.5403 | 0.0069 | 0.6359 | 0.0575 |
| 60403 | 30338 | 13089 | 28859.13 | 10005.62 | 0.4314 | 3.14E-05 | 0.3298 | 0.0056 | 0.4632 | 0.0281 |
| 60404 | 27191 | 14637 | 25961.73 | 10894.57 | 0.5383 | 3.51E-05 | 0.4007 | 0.0059 | 0.5304 | 0.0452 |
| 60405 | 14927 | 8498 | 16852.22 | 6372.756 | 0.5693 | 7.56E-05 | 0.4269 | 0.0087 | 0.5959 | 0.0489 |
| 60406 | 37968 | 18299 | 37074.5 | 13736.78 | 0.482 | 2.57E-05 | 0.3618 | 0.0051 | 0.4755 | 0.0354 |
| 60407 | 59759 | 7786 | 52237.04 | 5397.068 | 0.1303 | 1.46E-05 | 0.0903 | 0.0038 | 0.1705 | 0.0299 |
| 60408 | 11914 | 4722 | 10667.56 | 3598.263 | 0.3963 | 7.52E-05 | 0.302 | 0.0087 | 0.4235 | 0.04 |
| 60409 | 24990 | 19683 | 26663.84 | 15231.79 | 0.7876 | 4.27E-05 | 0.6095 | 0.0065 | 0.7462 | 0.0313 |
| 604010 | 16678 | 13427 | 18154.21 | 8777.214 | 0.8051 | 6.53E-05 | 0.5263 | 0.0081 | 0.7558 | 0.0327 |
| 604011 | 22171 | 10964 | 22554.32 | 7174.416 | 0.4945 | 4.59E-05 | 0.3236 | 0.0068 | 0.4814 | 0.0351 |
| 604012 | 22977 | 6109 | 25758.5 | 4238.139 | 0.2659 | 4.88E-05 | 0.1845 | 0.007 | 0.3227 | 0.0579 |
| 604013 | 22492 | 11890 | 24270.12 | 9046.675 | 0.5286 | 4.8E-05 | 0.4022 | 0.0069 | 0.598 | 0.0312 |
| 604014 | 24275 | 13433 | 25902.27 | 10045.27 | 0.5534 | 4.4E-05 | 0.4138 | 0.0066 | 0.5177 | 0.0401 |
| 604015 | 31866 | 12504 | 30349.78 | 9912.79 | 0.3924 | 2.99E-05 | 0.3111 | 0.0055 | 0.4249 | 0.04 |
| 604016 | 55650 | 3412 | 22328.26 | 2355.77 | 0.0613 | 7.2E-06 | 0.0423 | 0.0027 | 0.311 | 0.0362 |
| 604017 | 16993 | 9880 | 18108.42 | 7425.726 | 0.5814 | 6.27E-05 | 0.437 | 0.0079 | 0.5364 | 0.0434 |

Table 4.2: Results from Bernoulli function for Province Two

| MunCode | People | Poor People | V(X) | E(X) | $P_0$hard | V($P_0$) | $P_0$soft | SE($P_0$) | ($P_0$)sae | SE(sae) |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 43262 | 25295 | 45796.68 | 21887.23 | 0.584693 | 2.45E-05 | 0.505923 | 0.004947 | 0.560755 | 0.034446 |
| 2 | 34062 | 15183 | 31710.52 | 13359.04 | 0.445746 | 2.73E-05 | 0.392198 | 0.005228 | 0.481609 | 0.033933 |
| 3 | 21391 | 13388 | 22484.89 | 10837.55 | 0.625871 | 4.91E-05 | 0.506641 | 0.00701 | 0.557982 | 0.041679 |
| 4 | 21664 | 10044 | 19273.26 | 8412.15 | 0.463626 | 4.11E-05 | 0.388301 | 0.006408 | 0.429892 | 0.037767 |
| 5 | 24809 | 14379 | 24293.95 | 12565.39 | 0.579588 | 3.95E-05 | 0.506485 | 0.006283 | 0.600908 | 0.037213 |
| 6 | 26236 | 14073 | 26189.38 | 12261.3 | 0.5364 | 3.80E-05 | 0.467346 | 0.006168 | 0.57559 | 0.036467 |
| 7 | 44714 | 18364 | 42511.3 | 17410.73 | 0.410699 | 2.13E-05 | 0.38938 | 0.004611 | 0.428761 | 0.034132 |
| 8 | 35344 | 18447 | 35042.32 | 16852.77 | 0.521927 | 2.81E-05 | 0.476821 | 0.005296 | 0.559243 | 0.031794 |
| 9 | 16581 | 10048 | 16669.18 | 8750.778 | 0.605995 | 6.06E-05 | 0.527759 | 0.007787 | 0.579249 | 0.038549 |
| 10 | 11765 | 6398 | 11640.78 | 5542.208 | 0.543816 | 8.41E-05 | 0.471076 | 0.009171 | 0.559062 | 0.056937 |
| 12 | 44644 | 16800 | 40420.18 | 15813.46 | 0.37631 | 2.03E-05 | 0.354212 | 0.004503 | 0.345295 | 0.026021 |
| 13 | 47600 | 25491 | 45386.17 | 22148.87 | 0.535525 | 2.00E-05 | 0.465312 | 0.004476 | 0.536852 | 0.020302 |
| 14 | 50825 | 35056 | 53891.83 | 30044.19 | 0.689739 | 2.09E-05 | 0.59113 | 0.004568 | 0.718201 | 0.028156 |
| 15 | 33010 | 21524 | 34510.66 | 18782.18 | 0.652045 | 3.17E-05 | 0.568985 | 0.005628 | 0.671298 | 0.035612 |
| 16 | 37629 | 14834 | 33551.12 | 14036.24 | 0.394217 | 2.37E-06 | 0.373017 | 0.004868 | 0.388228 | 0.033602 |
| 17 | 29750 | 15618 | 31304.79 | 13297.46 | 0.524975 | 3.54E-05 | 0.446973 | 0.005947 | 0.521847 | 0.029053 |
| 18 | 53765 | 17055 | 48293.42 | 19729.77 | 0.317214 | 1.67E-05 | 0.366963 | 0.004087 | 0.413208 | 0.031869 |
| 19 | 34427 | 15376 | 30733.58 | 14355.22 | 0.446626 | 2.59E-05 | 0.416976 | 0.005092 | 0.487139 | 0.046562 |
| 20 | 26940 | 4256 | 25199.88 | 6917.063 | 0.159781 | 3.47E-05 | 0.256758 | 0.005893 | 0.308988 | 0.033587 |
| 21 | 26849 | 15635 | 27909.73 | 12775.76 | 0.582331 | 3.87E-05 | 0.475837 | 0.006222 | 0.519751 | 0.032786 |
| 22 | 348461 | 30099 | 243011.7 | 59349.06 | 0.086377 | 2.00E-06 | 0.170318 | 0.001415 | 0.106652 | 0.010429 |
| 23 | 52378 | 24889 | 50860.48 | 23102.04 | 0.47518 | 1.86E-05 | 0.441064 | 0.00431 | 0.53252 | 0.024567 |
| 25 | 60035 | 34141 | 60613.74 | 27534.01 | 0.568685 | 1.68E-05 | 0.458633 | 0.004101 | 0.538152 | 0.024469 |

Table 4.3: Results from Bernoulli function for Province Two cont.

| MunCode | People | Poor People | V(X) | E(X) | $P_0$hard | V($P_0$) | $P_0$soft | SE($P_0$) | ($P_0$)sae | SE(sae) |
|---|---|---|---|---|---|---|---|---|---|---|
| 26 | 22829 | 4485 | 20943.97 | 5926.153 | 0.196461 | 4.02E-05 | 0.259589 | 0.006339 | 0.338034 | 0.035727 |
| 27 | 23104 | 15419 | 23843.8 | 12510.27 | 0.667374 | 4.47E-05 | 0.541476 | 0.006683 | 0.645267 | 0.029112 |
| 28 | 41981 | 20369 | 43026.8 | 17628.58 | 0.485196 | 2.44E-05 | 0.419918 | 0.004941 | 0.505032 | 0.023668 |
| 29 | 29826 | 14575 | 30887.24 | 12734.41 | 0.488668 | 3.47E-05 | 0.426957 | 0.005892 | 0.517835 | 0.030179 |
| 30 | 55128 | 13739 | 56180.96 | 17400.02 | 0.24922 | 1.85E-05 | 0.315629 | 0.0043 | 0.40035 | 0.024062 |
| 31 | 17563 | 7170 | 16803.42 | 6690.454 | 0.408245 | 5.45E-05 | 0.380094 | 0.007381 | 0.406624 | 0.034045 |
| 32 | 18729 | 6947 | 16696.61 | 6504.914 | 0.370922 | 4.76E-05 | 0.347318 | 0.006899 | 0.347954 | 0.041191 |
| 34 | 63745 | 11423 | 56545.21 | 16003.1 | 0.179198 | 1.39E-05 | 0.251049 | 0.00373 | 0.292117 | 0.030115 |
| 35 | 66705 | 34909 | 68010.29 | 30114.94 | 0.523334 | 1.53E-05 | 0.451465 | 0.00391 | 0.506718 | 0.025264 |
| 36 | 31500 | 3941 | 22920.67 | 6229.766 | 0.125111 | 2.31E-05 | 0.19777 | 0.004806 | 0.140605 | 0.027241 |
| 37 | 58725 | 17569 | 52259.49 | 20111.76 | 0.299174 | 1.52E-05 | 0.342474 | 0.003893 | 0.351496 | 0.029717 |
| 38 | 27870 | 18603 | 28885.63 | 15838.84 | 0.667492 | 3.72E-05 | 0.568311 | 0.006098 | 0.650342 | 0.030266 |
| 39 | 27282 | 17233 | 28223.3 | 13814.03 | 0.631662 | 3.79E-05 | 0.506342 | 0.006158 | 0.565221 | 0.027599 |
| 40 | 46350 | 26539 | 50832.33 | 22635.26 | 0.572578 | 2.37E-05 | 0.488355 | 0.004864 | 0.577584 | 0.025064 |
| 41 | 20527 | 2630 | 17122.93 | 4608.641 | 0.128124 | 4.06E-05 | 0.224516 | 0.006375 | 0.235176 | 0.036755 |
| 42 | 12154 | 6669 | 12598.48 | 6103.48 | 0.548708 | 8.53E-05 | 0.502179 | 0.009235 | 0.568022 | 0.058488 |
| 43 | 44558 | 6980 | 38386.75 | 10320.94 | 0.15665 | 1.93E-05 | 0.231629 | 0.004397 | 0.286753 | 0.026199 |
| 44 | 41182 | 20461 | 39930.41 | 19046.41 | 0.496843 | 2.35E-05 | 0.462494 | 0.004852 | 0.518999 | 0.034101 |
| 45 | 48772 | 10269 | 47302.09 | 14325.97 | 0.210551 | 1.99E-05 | 0.293733 | 0.004459 | 0.33685 | 0.026047 |
| 46 | 18293 | 10863 | 19716.89 | 8782.008 | 0.593834 | 5.89E-05 | 0.480075 | 0.007676 | 0.572842 | 0.027985 |
| 47 | 17764 | 3513 | 15741.46 | 4514.625 | 0.19776 | 4.99E-05 | 0.254145 | 0.007063 | 0.278454 | 0.03599 |

## 4.2 Calculation Methods count in Municipal Variances

Note, the standard errors we computed from Section One is extremely small, one of the reason for small standard error is we didn't count the municipal variance component. For including the municipal variance it require a numerous simulation, each of them done in municipal level obtain the relevant figures afterwards. There are two approaches we have tried to count in the municipal variances, which been listed below.

Generally, the standard errors calculated using these two approaches are more sensible and closer to the standard errors gained using Small Area Estimation method. To be specific, one method is calculated based on bootstrapping the training data at municipal level (Method Two); whereas, the other is computed by a model based method using cross-validation errors (Method Three).

Comparing the standard errors from two calculation methods, the small area estimation method's standard errors been treated as the standard value, generally estimates obtained from random selection have higher values comparing the standard errors from small area estimation method (has the range from 0.02 to 0.09); whereas, the standard errors computed with model based method are relatively small (as the range is 0.015 to 0.038).

### 4.2.1 Standard Error Calculation Method Two

In consideration of very small standard errors we calculated above, to overcome this problem we have come up with a method to produce a large set of predictions and calculate mean, standard deviation out of it.

This procedure applies the random selection of municipalities with replacement in survey data, in other words the data that are used for building the classification tree model have been formed by selecting 1250 municipalities randomly from the survey data with replacement, in which this is the same number of municipalities in survey data. Subsequently, we obtain the classification tree and use it for predict Province One and Two. Getting the prediction on each household in the test data (Province One and Province Two), we would be able to compute the proportion of poor in municipal level. This procedure will repeated for 50 times, in that we would have

50 sets of predicted proportion of poor in municipal level for both provinces. The 50 bootstrap estimates for each municipality were summarized by their mean and standard deviation, giving a point estimate and a standard error for each municipality.

Now let's move on to the discussion of the results we attained using this approach. For the purpose of counting the municipal variance components, a number of simulations operated, a set of proportion of prediction been calculated in municipal level. Unlike the standard error show in Section One, the standard errors calculated using this method is considerably closer to the standard errors computed using Small Area Estimation (specific figures illustrated in Table 4.4 and 4.5, which are corresponding to Province One and Two). The scatter plot has drawn between standard errors, Figure 4.2 illustrates the standard errors in Province One, it appears standard errors obtained using random selection method are generally higher than ones from Small Area Estimation method, as most of dots located above the 45 degree line. Figure 4.3 plots standard errors from Province Two, again, the standard errors produced by the random selection method are generally higher than Small Area Estimation Method's standard errors. Graph 4.4 plots combination of two Provinces' standard errors together.

Table 4.4: Random Selection results for Province One

| mcode | mean | standard error | SE(sae) |
|---|---|---|---|
| 1 | 1.520153 | 0.05915361 | 0.0496 |
| 2 | 1.83246 | 0.05289653 | 0.0575 |
| 3 | 1.209553 | 0.04424858 | 0.0281 |
| 4 | 1.557589 | 0.04186352 | 0.0452 |
| 5 | 1.626165 | 0.04250010 | 0.0489 |
| 6 | 1.389456 | 0.04812457 | 0.0354 |
| 7 | 0.424887 | 0.03513671 | 0.0299 |
| 8 | 1.133748 | 0.04239786 | 0.04 |
| 9 | 2.152355 | 0.04480459 | 0.0313 |
| 10 | 2.161707 | 0.04212351 | 0.0327 |
| 11 | 1.337961 | 0.05463948 | 0.0351 |
| 12 | 1.093242 | 0.07233704 | 0.0579 |
| 13 | 1.596827 | 0.04052988 | 0.0312 |
| 14 | 1.545286 | 0.04265507 | 0.0401 |
| 15 | 1.154184 | 0.05046312 | 0.04 |
| 16 | 0.6781623 | 0.05553521 | 0.0362 |
| 17 | 1.584469 | 0.04897079 | 0.0434 |

Table 4.5: Random Selection results for Province Two

| mcode | mean | standard error | SE(sae) |
|---|---|---|---|
| 1 | 0.6438224 | 0.04877424 | 0.034446 |
| 2 | 0.5203062 | 0.05720704 | 0.033933 |
| 3 | 0.6103374 | 0.0489852 | 0.041679 |
| 4 | 0.4629213 | 0.06295801 | 0.037767 |
| 5 | 0.6322944 | 0.03607345 | 0.037213 |
| 6 | 0.5703431 | 0.04938258 | 0.036467 |
| 7 | 0.4125208 | 0.04107428 | 0.034132 |
| 8 | 0.6113954 | 0.03950223 | 0.031794 |
| 9 | 0.683404 | 0.0354239 | 0.038549 |
| 10 | 0.5887629 | 0.05573433 | 0.056937 |
| 12 | 0.3382787 | 0.06206538 | 0.02602 |
| 13 | 0.5979526 | 0.03630998 | 0.020302 |
| 14 | 0.7875214 | 0.03642715 | 0.028156 |
| 15 | 0.7488626 | 0.03083764 | 0.035612 |
| 16 | 0.3815614 | 0.05006894 | 0.033602 |
| 17 | 0.5410183 | 0.08338134 | 0.029053 |
| 18 | 0.3778713 | 0.03862623 | 0.031869 |
| 19 | 0.539315 | 0.03598054 | 0.046562 |
| 20 | 0.268781 | 0.06235302 | 0.033587 |
| 21 | 0.5940225 | 0.05754758 | 0.032786 |
| 22 | 0.0815435 | 0.02230296 | 0.010429 |
| 23 | 0.5394206 | 0.05313095 | 0.024567 |
| 25 | 0.575336 | 0.05253858 | 0.024469 |
| 26 | 0.2945904 | 0.04899486 | 0.035727 |
| 27 | 0.6924372 | 0.05499278 | 0.029112 |
| 28 | 0.5426887 | 0.05715067 | 0.023668 |
| 29 | 0.5580351 | 0.04773331 | 0.030179 |
| 30 | 0.3768457 | 0.06235881 | 0.024062 |
| 31 | 0.3923188 | 0.06256528 | 0.034045 |
| 32 | 0.3398534 | 0.06572763 | 0.041191 |
| 34 | 0.2786664 | 0.052242 | 0.030115 |
| 35 | 0.5614899 | 0.05102043 | 0.025264 |
| 36 | 0.1547807 | 0.0358126 | 0.027241 |
| 37 | 0.3422338 | 0.04572671 | 0.029717 |
| 38 | 0.756731 | 0.03285061 | 0.030266 |
| 39 | 0.6182581 | 0.05359791 | 0.027599 |
| 40 | 0.5895489 | 0.05677785 | 0.025064 |
| 41 | 0.2443766 | 0.05132485 | 0.036755 |
| 42 | 0.6461397 | 0.04893429 | 0.058488 |
| 43 | 0.2460093 | 0.05251585 | 0.026199 |
| 44 | 0.5642773 | 0.04834979 | 0.034101 |
| 45 | 0.3088625 | 0.0450099 | 0.026047 |
| 46 | 0.6181947 | 0.05523134 | 0.027985 |
| 47 | 0.2737418 | 0.0537859 | 0.03599 |

### 4.2.2 Standard Error Calculation Method Three

Using the method we discussed in Section One, it is clear that the standard errors are too small to consider, as we comparing it to the standard errors from Small Area Estimation method. In this section we will attempt to over come this problem by discovering a new approach to compute the standard errors for Province One and Two.

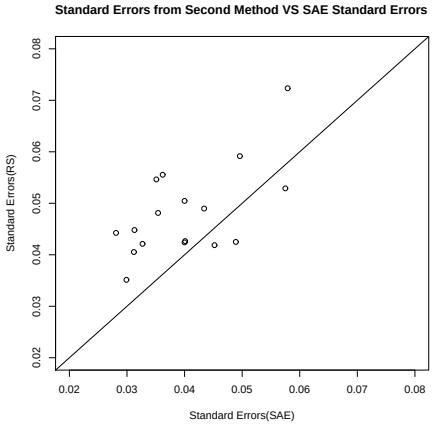


Figure 4.2: Standard Errors obtained using the Second calculation method VS SAE Standard Errors for Province One


Figure 4.3: Standard Errors obtained using the Second calculation method VS SAE Standard Errors for Province Two

For standard errors estimated at municipal level, each of them is affected by municipal features. The sample data has 1250 municipalities in total, each municipality different from one to another. To count the variation at municipal level we leave one municipality out of the training data and build the model based on the rest of the municipalities. To test the new model on the remaining municipality, predict each household and obtain the proportion of poor for each municipality. Per-person scale survey weights have been used in calculating the proportion of predicted poor people in a municipality. The same procedure is repeated 1250 times, for every municipality in the training data, so that we have 1250 models built for all the municipalities. The prediction errors are calculated by subtracting the actual from the predicted proportion of poor in each municipality. A relationship between standard

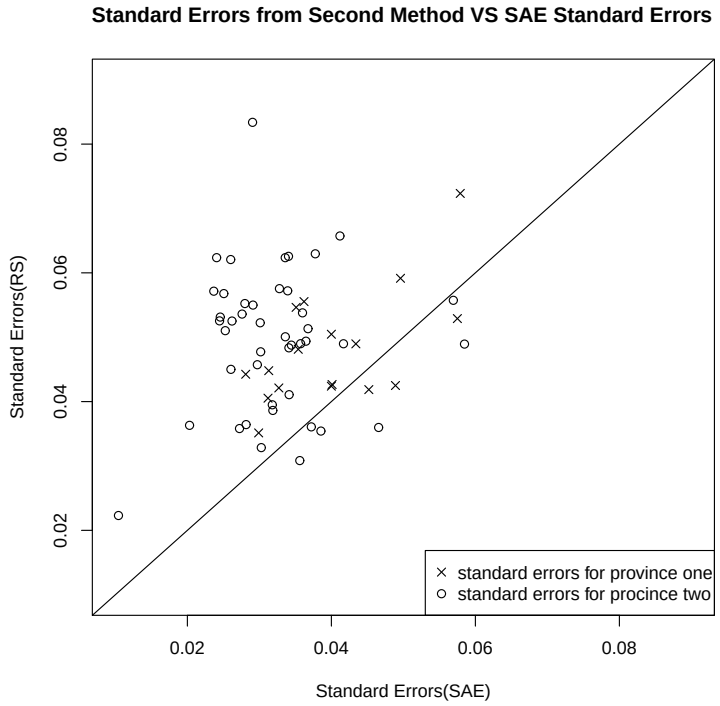


Figure 4.4: Standard Errors obtained using the Second calculation method VS SAE Standard Errors for combining both Provinces

error and poverty estimates, municipal capacity measured by household size was found by Haslett and Jones (Haslett & Jones, 2005). Employ a log transform on squared errors, household size and proportion of actual poor and non-poor. The logarithmic scale reduces wide-ranging quantities to small scopes. The logarithmic of the proportions of actual poverty estimates and the municipal capacity measured in household size been used to model the logged square of standard errors using regression equation. We assume that the relationship between error and proportion of poor, proportion of non-poor, household size can be expressed as an equation (expressed in equation 4.2):

$$Error^2 \propto \frac{P(1-P)}{sizehh} \tag{4.2}$$

Taking log on both sides. The equation has become (equation 4.3):

$$log(Error^2) = log(P) + log(1-P) - log(sizehh) \tag{4.3}$$

A model containing these three terms was fitted to the municipal prediction errors. As expected both log(P) and log(1-P) have positive coefficients and negative

Figure 4.5: Regression equation on Municipal level logged error square

```
> Sum.lm <- lm(ErSq.log ~ propRP.log + RNP.log + sizehh.log, data=sumTable)
> summary(Sum.lm)

Call:
lm(formula = ErSq.log ~ propRP.log + RNP.log + sizehh.log, data = sumTable)

Residuals:
     Min       1Q   Median       3Q      Max
-13.5349  -1.0217   0.3489   1.5017   4.2024

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.94144    0.31041  -9.476  < 2e-16 ***
propRP.log   0.44730    0.13984   3.199  0.00142 **
RNP.log      0.02895    0.15499   0.187  0.85184
sizehh.log  -0.37440    0.06139  -6.099 1.45e-09 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 2.082 on 1164 degrees of freedom
Multiple R-squared: 0.06522,    Adjusted R-squared: 0.06281
F-statistic: 27.07 on 3 and 1164 DF,  p-value: < 2.2e-16
```

on log(sizehh). One important thing here is there are about 200 municipalities that obtained the prediction results exactly the same as actual categories. Therefore the subtraction between predicted and actual would be zero. Moreover, the negative infinite result computed by taking log of these zero values. Consequently, these about 200 values were withdrawn from the sample data for the purpose of building regression equation (display as Figure 4.5). Thereafter, we run regression on these variables ($log(Error^2) = \beta_0 + \beta_1 log(P) + \beta_2 log(1-P) - \beta_3 log(sizehh)$). The regression result is displayed in Figure 4.5 in which all the variables are estimated in municipal level. All the variables have been log transformed in the equation below, and each variable been calculated at municipal level. There is strong evidence of negative relationship between variable 'sizehh.log' (log transformed household size in municipal level) and log of squared errors (the p-value is 1.45e-09); and a positive relationship between response variable and logged actual poor ('propRP.log',significant p-value: 0.00142). However, the p-value of 'RNP.log' (logged actual proportion of non-poor) is 0.87259 which considered as very insignificant.

Now apply the regression equation above to both Province One and Two to estimate the standard errors at municipal level. Since a log transform was applied in modeling standard errors, poverty estimates (it produced from the most suitable

classification tree model which it using per-person survey weight and set the prior as 0.34 and 0.66) and municipal capacity, we now undo this transformation by exponentiating. Unlike the standard error calculated in Section One, an acceptable value of the standard error is produced using this approach. In addition, both standard errors (one attained by small area estimation and the other by classification tree method) are very close to each other).

$$MerSq.log = -2.94144 + 0.4473propRP.log + 0.02895RNP.log - 0.3744sizehh.log \tag{4.4}$$

Table 4.6: Table for Province One

| MunCode | Sizehh | $P_0$(sae) | SE(sae) | $P_0$(ct) | SE(ct) |
|---|---|---|---|---|---|
| 60401 | 4277 | 0.4907339 | 0.0495909 | 0.7836516 | 0.03256946 |
| 60402 | 4365 | 0.6359334 | 0.0575189 | 0.8741042 | 0.03310338 |
| 60403 | 6139 | 0.4631762 | 0.0280642 | 0.6659635 | 0.02958854 |
| 60404 | 5748 | 0.5303736 | 0.0451611 | 0.7760288 | 0.03133691 |
| 60405 | 2728 | 0.5959101 | 0.0488527 | 0.8375427 | 0.03555290 |
| 60406 | 7613 | 0.4755336 | 0.0354011 | 0.7358302 | 0.02859419 |
| 60407 | 12154 | 0.1704645 | 0.0299470 | 0.3165214 | 0.02256199 |
| 60408 | 2402 | 0.4235236 | 0.0400233 | 0.6308545 | 0.03511099 |
| 60409 | 4456 | 0.7461784 | 0.0313370 | 0.8984794 | 0.03256225 |
| 604010 | 2959 | 0.7557867 | 0.0327414 | 0.8733062 | 0.03502674 |
| 604011 | 4231 | 0.4813572 | 0.0351459 | 0.7057417 | 0.03200935 |
| 604012 | 4532 | 0.3227345 | 0.0578767 | 0.6966967 | 0.03169481 |
| 604013 | 4175 | 0.5979895 | 0.0312395 | 0.7466773 | 0.03223952 |
| 604014 | 4582 | 0.5177162 | 0.0400600 | 0.7676210 | 0.03202785 |
| 604015 | 6341 | 0.4248556 | 0.0400234 | 0.6828909 | 0.02964384 |
| 604016 | 4713 | 0.3109612 | 0.0362379 | 0.4277245 | 0.02919110 |
| 604017 | 3219 | 0.5364162 | 0.0434048 | 0.8085682 | 0.03449981 |

Three standard error plots have been drawn according to the Table 4.6 and 4.7, and the last one by merging results from both provinces together. To interpret the two separate plots (Plot 4.6 and 4.7), standard errors by both provinces of classification tree concentrate in the range from 0.02 to 0.04; simultaneously, there is a wide range of standard error computed for small area estimation method (mostly between 0.02 and 0.05). A 45 degree line has been drawn across the graph for the convenience of comparison between standard errors. In Figure 4.6 most of dots lie on the left and middle part of graph, in other words, standard errors for SAE are higher than the standard errors of classification tree. On the contrary, for Province Two it seems both the standard errors are spread symmetrically and concentrated (both have a similar range, both are around 0.02 to 0.04); by way of explanation, the two set of estimated standard errors are close to each other, the reason for this might because Province Two has relatively larger populations in each municipality. Three extreme cases need to be pointed out on the plot, one in the left lower corner

Table 4.7: Table for Province Two

| MunCode | Sizehh | $P_0$(sae) | SE(sae) | $P_0$(ct) | SE(ct) |
|---|---|---|---|---|---|
| 1 | 7966 | 0.584693 | 0.034446 | 0.8274467 | 0.02909138 |
| 2 | 5825 | 0.445746 | 0.033933 | 0.7471358 | 0.03014251 |
| 3 | 4124 | 0.625871 | 0.041679 | 0.8166519 | 0.03314074 |
| 4 | 4479 | 0.463626 | 0.037767 | 0.7252123 | 0.03252839 |
| 5 | 4988 | 0.579588 | 0.037213 | 0.8020880 | 0.03204049 |
| 6 | 5109 | 0.5364 | 0.036467 | 0.7769477 | 0.03155912 |
| 7 | 8558 | 0.410699 | 0.034132 | 0.6826721 | 0.02779965 |
| 8 | 6779 | 0.521927 | 0.031794 | 0.7682209 | 0.02978021 |
| 9 | 3175 | 0.605995 | 0.038549 | 0.8491647 | 0.03494599 |
| 10 | 2370 | 0.543816 | 0.056937 | 0.8268593 | 0.03722391 |
| 12 | 8785 | 0.37631 | 0.026021 | 0.6810546 | 0.02779313 |
| 13 | 9131 | 0.535525 | 0.020302 | 0.7618277 | 0.02809881 |
| 14 | 9386 | 0.689739 | 0.028156 | 0.9001476 | 0.02851806 |
| 15 | 6239 | 0.652045 | 0.035612 | 0.8771887 | 0.03083095 |
| 16 | 7423 | 0.394217 | 0.033602 | 0.6785458 | 0.02872331 |
| 17 | 5604 | 0.524975 | 0.029053 | 0.8000336 | 0.03097498 |
| 18 | 10864 | 0.317214 | 0.031869 | 0.6715893 | 0.02636998 |
| 19 | 6818 | 0.446626 | 0.046562 | 0.7567897 | 0.02985029 |
| 20 | 5176 | 0.159781 | 0.033587 | 0.5518560 | 0.02987043 |
| 21 | 5393 | 0.582331 | 0.032786 | 0.7953741 | 0.03162197 |
| 22 | 69408 | 0.086377 | 0.010429 | 0.2220076 | 0.01503437 |
| 23 | 9946 | 0.47518 | 0.024567 | 0.7574745 | 0.02759073 |
| 25 | 10980 | 0.568685 | 0.024469 | 0.7641376 | 0.02700943 |
| 26 | 4379 | 0.196461 | 0.035727 | 0.4254676 | 0.02960463 |
| 27 | 4575 | 0.667374 | 0.029112 | 0.8711911 | 0.03297831 |
| 28 | 7775 | 0.485196 | 0.023668 | 0.7436459 | 0.02868378 |
| 29 | 5243 | 0.488668 | 0.030179 | 0.7273855 | 0.03045425 |
| 30 | 10602 | 0.24922 | 0.024062 | 0.6448810 | 0.02701996 |
| 31 | 3335 | 0.408245 | 0.034045 | 0.7038091 | 0.03331351 |
| 32 | 3626 | 0.370922 | 0.041191 | 0.6559880 | 0.03246573 |
| 34 | 12609 | 0.179198 | 0.030115 | 0.4341931 | 0.02446199 |
| 35 | 12450 | 0.523334 | 0.025264 | 0.7474102 | 0.02631272 |
| 36 | 6329 | 0.125111 | 0.027241 | 0.3278730 | 0.02560163 |
| 37 | 11350 | 0.299174 | 0.029717 | 0.6461814 | 0.02614884 |
| 38 | 5375 | 0.667492 | 0.030266 | 0.8725870 | 0.03179702 |
| 39 | 5231 | 0.631662 | 0.027599 | 0.8316472 | 0.03174036 |
| 40 | 8226 | 0.572578 | 0.025064 | 0.7936570 | 0.02849215 |
| 41 | 4080 | 0.128124 | 0.036755 | 0.5326643 | 0.03145133 |
| 42 | 2331 | 0.548708 | 0.058488 | 0.8276288 | 0.03688343 |
| 43 | 8581 | 0.15665 | 0.026199 | 0.5216347 | 0.02721882 |
| 44 | 8263 | 0.496843 | 0.034101 | 0.7861687 | 0.02920025 |
| 45 | 9507 | 0.210551 | 0.026047 | 0.4829 | 0.02629662 |
| 46 | 3618 | 0.593834 | 0.027985 | 0.8399934 | 0.03444196 |
| 47 | 3422 | 0.19776 | 0.03599 | 0.4473655 | 0.03156289 |

as both methods calculated very low values of the standard error, two are on the mid-right of the plot, both of them have extreme value from SAE method (reach to 0.06) and for classification tree these two are estimated only about 0.035. Finally, the last plot (Figure 4.8) is just a combination of standard errors from two provinces, where the cross is the standard errors from Province One and the dots are standard errors from Province Two. This combination plot displays that most of standard

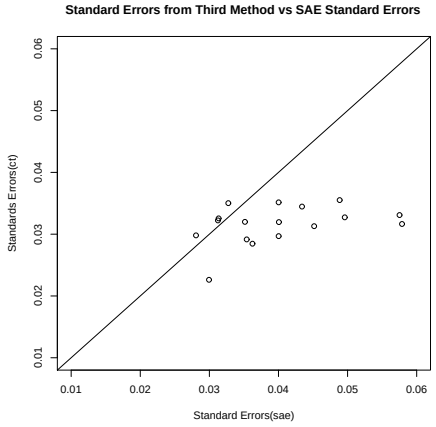


Figure 4.6: Standard Errors obtained using the Third calculation method VS SAE Standard Errors for Province One

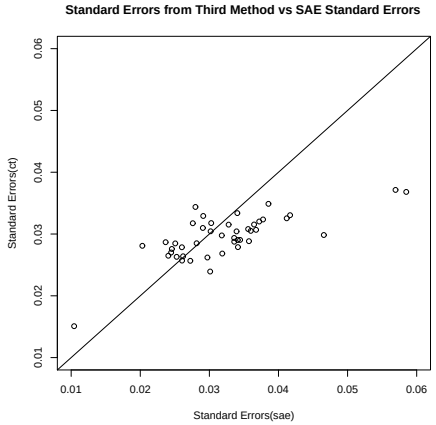


Figure 4.7: Standard Errors obtained using the Third calculation method VS SAE Standard Errors for Province Two

errors lying at lower right part of graph, by way of explanation, small range for standard errors of classification tree model (0.025 to 0.04) whereas a relative large range occurred for small area estimation method (0.02 to 0.06).

To conclude, this standard error calculation method has two shortfalls. One, as there are about 200 municipalities been removed from the data that used to form the regression equation for standard errors this may affect the outcome for calculating the standard errors. Second, the data set that we employed for setting up standard error regression equation is way too small for the whole census data. These are two main limitations for the standard errors we attained using this method.

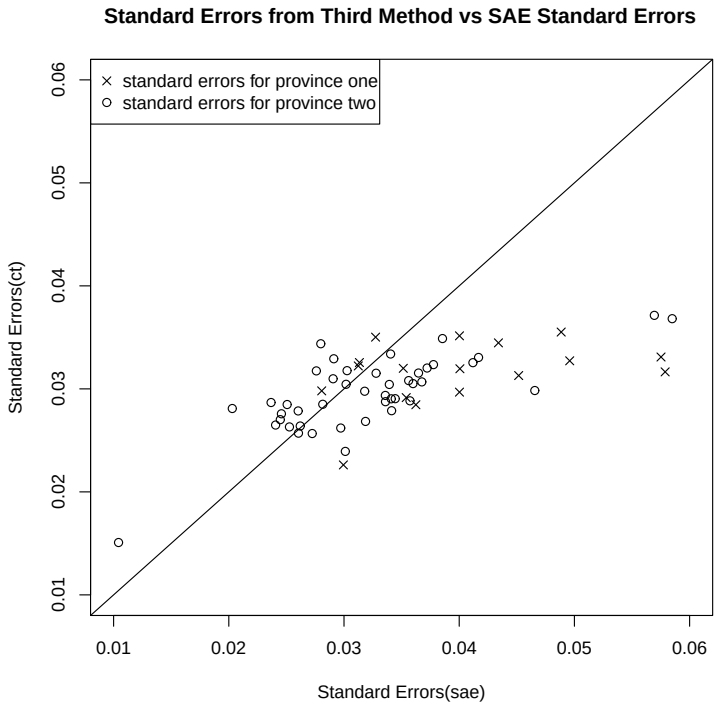


Figure 4.8: Standard Errors obtained using the Third calculation method VS SAE Standard Errors for combining both Provinces

# Chapter 5

# Conclusions and Further work

The adaptation of the classification tree method for survey data seems not have been considered before, further it seems promising to apply this method for poverty estimation because results obtained by classification tree are similar to the established method (small area estimation method).

The sample data used for poverty mapping is a small proportion selected out of the whole population. There is only a small amount of data collected for each small region which may be too small to generate accurate estimates. The established way to analyze such data would the small area estimation method, which can produce small-area estimates of poverty at municipality level by combing survey data with auxiliary data derived from the complete 2000 census short-form and 10% sample of long-form census data. The strategy of for choosing appropriate regression models for the target variable is not usually made explicit, it appeared as using separate models in each stratum, sometimes with large number of strata, and that variables have been selected from a very large pool of possibilities including all interaction terms, criterion as adjusted $R^2$ or AIC will penalize for fitting too many variables, but do not make proper account when there are a large number of variables to selected from. It is difficult and time consuming process for obtaining the appropriate model. A single model supplemented by different urban and rural effects within each region was found to be adequate for predicting log average per capita household expenditure and the poverty measures derived from it. The municipality-level estimates obtained have acceptably low standard errors, where they are obtained by using small area estimation method (Jones and Haslett, 2005).

In this study the tree-structured method has been employed for classifying the poverty of a household. The benefit of the classification tree model is that it is

simpler and efficient. The strategy of forming an appropriate classification tree for the target variable is according to a certain rule as every household eventually falls into the category either poor or non-poor. Instead of using a multiple regression equation we apply the classification tree incorporating survey weights for the prediction. The estimate result produced from our method is close to ones obtained by the small area estimation method. Intuitively, as similar results are obtained using the classification tree method, comparing it to the established method, small area estimation, the classification tree method is rather simpler and efficient. In general, the information criteria applied mostly here is Gini method, as we attained better results by using this method. Generally, the confusion matrices computed by all methods we demonstrated in this study showed a high accuracy rate for the prediction of non-poor (around 89%) and a relatively low percent for the prediction of poor (approximately 60%). A Receiver Operating Characteristic curve was employed to graph the sensitivity against false positive rate for such a binary classification, showing all possible combinations of sensitivity and (1 − specificity) attained by differing priors. To a degree, the best prior combination displayed in plot is 0.3415 and 0.6585 (the prior sets as the proportion of poor and non-poor respectively in per-person unit) which has high a percentage of true positives without losing much information.

The provision of standard errors with the classification tree method is seen as important because it gives the user an impression of how much accuracy is being claimed. Among all three methods of calculating standard errors the most acceptable result of the standard errors would be the one based on subtraction between predicted and actual results in municipal level, thereafter, building a regression equation based on taking log of all of variables. The equation is shown as:

$$log(Error^2) = \beta_0 + \beta_1 log(P) + \beta_2 log(1 - P) - \beta_3 log(sizehh) \tag{5.1}$$

Finally, apply this regression equation (Equation 5.1) to estimate the standard errors for the predicating data set. However, there are two major shortfalls from the calculation of standard errors. First, there are about 200 municipalities has the prediction output exactly the same as the actual classification, therefore the subtraction for these 200 municipalities would be 0 and log these series of zeros are negative infinite, it cannot included to form the regression equation, so that it has been removed from the basic data, to a degree this might affect the estimation of standard errors. Second, a limited range of municipal sizes in the basic data set was used to build the regression equation for standard errors, therefore using this

to predict on census data was out of range for log(sizehh).

There are a number of specific areas of further work which need to be carried out. These are all extensions of the new approach and do not represent any new work. A number of suggestions are included below.

- Looking at the calculation of the standard errors for classification tree models, as there are major drawbacks to proposed methods. Further work can be done for getting more accurate standard errors. Perhaps we could have larger data for building the regression equation for model-based variance by taking random subsets of the survey data.

- In the third method of calculating standard errors for classification tree model, the regression model we applied is based on the relationship conclude from SAE method, for the classification tree it could be different, further research can be done for forming a model for classification tree's standard errors.

- Two particular information criteria have been applied for building an appropriate tree. More research can be carried out for finding the best rule of selecting leaves, nodes for such data.

- A small proportion been selected from the whole population, in order to have more precise estimations a larger data set is required. Therefore a larger data base can be selected as increase the amount of data collected from each small region for building the classification tree.

- Since the prediction on non-poor is fairly accurate using classification tree method, therefore a reverse approach may be more efficient especially in calculating the standard errors, predicting the non-poor instead and have the total minus the non-poor to calculate the poor.

# Appendix : X Variables and Full Trees

# List of X Variables for the Philippines data

Table 6.1: List of X Variables for the Philippines data

| Variable Name | Variable Label |
|---|---|
| Prov | Province code |
| Regn | Region code |
| Mun | Municipality code |
| Urb | 1 if urban |
| bcode | Barangay code |
| sswgthh | survey weighted in household level |
| sswgtpp | survey weighted in per-person level |
| incpp | income per-person |
| famsize | number of persons in household |
| famsizesqc | square of mean-adjusted family size |
| head_male | 1 if head is male |
| no_spouse | 1 if no spouse in family |
| per_kids | proportion of household members who are sons/daughters of head |
| per_61up | proportion of members ages 61 and up |
| dom_help | 1 if household has domestic help |
| all_noed | proportion of all members 10 years and over with no education |
| all_eled | proportion of all members 10 years and over with only elememtary education |
| all_hsed | proportion of all members 10 years and over with only high school education |
| all_coed | proportion of all members 10 years and over with college education |
| hou_9600 | % of dwellings in municipality built in 1996-2000 |
| hea_rel_mus | % of heads in municipality who are Muslim |
| hou_own_ref | % of households who have refrigerator |
| hou_own_tel | % of households who have telephone |
| hou_coelpg | % of households that use electricity or lpg for cooking |
| per_ind_52 | % of persons employed in retail trade |
| per_wor_prh | % who worked for private household |
| per_eng | % of persons 5 and older who speak English |
| typeN | type of housing (single house, duplex, apart., comm. or other) |
| wallN | type of material that walls are made of (light materials, salvaged materials or other materials) |
| roofN | type of materials that roofs are made of (light materials, salvaged materials or other materials) |
| faN | lot floor area size (xs is less than 18.6sq m; s is 18.6-32.5sq m; m is 32.5-44.6sq m; l is 44.6-65sq m; xl is 65-83.6sq m; xxl is 83.6-139.4sq m) |
| undpovN | estimation of poverty on per-person level |

# Full Trees Built with Various Data Base

Figure 6.1: Split Rules for Gini Tree with un-weighted data (Figure 3)

```
n= 31623

node), split, n, loss, yval, (yprob)
      * denotes terminal node

   1) root 31623 8700 non-poor (0.72488379 0.27511621)
     2) all_coed>=0.1909091 12282  851 non-poor (0.93071161 0.06928839)
       4) roofN=oth,strong 11206  580 non-poor (0.94824201 0.05175799) *
       5) roofN=light,salvaged 1076  271 non-poor (0.74814126 0.25185874)
        10) hea_rel_mus< 0.36848 952  200 non-poor (0.78991597 0.21008403)
          20) all_coed>=0.3038461 546   64 non-poor (0.88278388 0.11721612) *
          21) all_coed< 0.3038461 406  136 non-poor (0.66502463 0.33497537)
            42) all_hsed>=0.1180556 334   96 non-poor (0.71257485 0.28742515) *
            43) all_hsed< 0.1180556 72   32 poor (0.44444444 0.55555556) *
        11) hea_rel_mus>=0.36848 124   53 poor (0.42741935 0.57258065) *
     3) all_coed< 0.1909091 19341 7849 non-poor (0.59417817 0.40582183)
       6) famsize< 4.75 7786 1753 non-poor (0.77485230 0.22514770)
        12) Hou_own_tel>=0.0912905 3498  390 non-poor (0.88850772 0.11149228) *
        13) Hou_own_tel< 0.0912905 4288 1363 non-poor (0.68213619 0.31786381)
          26) famsize< 3.25 2481  572 non-poor (0.76944780 0.23055220)
            52) faN=l,m,xl,xxl,xxxl 1077  155 non-poor (0.85608171 0.14391829) *
            53) faN=s,xs 1404  417 non-poor (0.70299145 0.29700855)
             106) famsize< 1.75 311   49 non-poor (0.84244373 0.15755627) *
             107) famsize>=1.75 1093  368 non-poor (0.66331199 0.33668801)
               214) all_hsed>=0.4166667 251   48 non-poor (0.80876494 0.19123506) *
               215) all_hsed< 0.4166667 842  320 non-poor (0.61995249 0.38004751)
                 430) all_eled>=0.4166667 197   47 non-poor (0.76142132 0.23857868) *
                 431) all_eled< 0.4166667 645  273 non-poor (0.57674419 0.42325581)
                   862) hea_rel_mus>=0.000282 428  161 non-poor (0.62383178 0.37616822)
                    1724) hou_9600< 0.519896 359  123 non-poor (0.65738162 0.34261838) *
                    1725) hou_9600>=0.519896 69   31 poor (0.44927536 0.55072464) *
                   863) hea_rel_mus< 0.000282 217  105 poor (0.48387097 0.51612903)
                    1726) wallN=oth,salvaged,strong 96   40 non-poor (0.58333333 0.41666667) *
                    1727) wallN=light 121   49 poor (0.40495868 0.59504132) *
          27) famsize>=3.25 1807  791 non-poor (0.56225789 0.43774211)
            54) roofN=oth,strong 1064  376 non-poor (0.64661654 0.35338346)
             108) all_hsed>=0.3095238 494  129 non-poor (0.73886640 0.26113360) *
             109) all_hsed< 0.3095238 570  247 non-poor (0.56666667 0.43333333)
               218) Hou_coelpg>=0.2031039 242   79 non-poor (0.67355372 0.32644628) *
               219) Hou_coelpg< 0.2031039 328  160 poor (0.48780488 0.51219512)
                 438) Per_wor_prh>=0.0665455 136   53 non-poor (0.61029412 0.38970588) *
```

Figure 6.2: Split Rules for Gini Tree with un-weighted data Cont.

```
          439) Per_wor_prh< 0.0665455 192   77 poor (0.40104167 0.59895833) *
       55) roofN=light,salvaged 743  328 poor (0.44145357 0.55854643)
        110) Per_wor_prh>=0.07779 191   85 non-poor (0.55497382 0.44502618) *
        111) Per_wor_prh< 0.07779 552  222 poor (0.40217391 0.59782609)
          222) all_noed< 0.08333335 463  201 poor (0.43412527 0.56587473)
            444) Hou_own_ref< 0.092581 116   51 non-poor (0.56034483 0.43965517) *
            445) Hou_own_ref>=0.092581 347  136 poor (0.39193084 0.60806916) *
          223) all_noed>=0.08333335 89   21 poor (0.23595506 0.76404494) *
  7) famsize>=4.75 11555 5459 poor (0.47243617 0.52756383)
   14) Hou_own_ref>=0.2994995 4904 1585 non-poor (0.67679445 0.32320555)
     28) all_hsed>=0.2752526 3165  735 non-poor (0.76777251 0.23222749)
       56) wallN=strong 1893  317 non-poor (0.83254094 0.16745906) *
       57) wallN=light,oth,salvaged 1272  418 non-poor (0.67138365 0.32861635)
        114) per_kids< 0.6833333 972  274 non-poor (0.71810700 0.28189300) *
        115) per_kids>=0.6833333 300  144 non-poor (0.52000000 0.48000000)
          230) Per_eng>=0.8221655 101   30 non-poor (0.70297030 0.29702970) *
          231) Per_eng< 0.8221655 199   85 poor (0.42713568 0.57286432) *
     29) all_hsed< 0.2752526 1739  850 non-poor (0.51121334 0.48878666)
       58) wallN=oth,strong 1254  501 non-poor (0.60047847 0.39952153)
        116) Hou_coelpg>=0.7312998 564  167 non-poor (0.70390071 0.29609929) *
        117) Hou_coelpg< 0.7312998 690  334 non-poor (0.51594203 0.48405797)
          234) all_coed>=0.07738095 175   52 non-poor (0.70285714 0.29714286) *
          235) all_coed< 0.07738095 515  233 poor (0.45242718 0.54757282)
            470) per_kids< 0.6458333 239  104 non-poor (0.56485356 0.43514644) *
            471) per_kids>=0.6458333 276   98 poor (0.35507246 0.64492754) *
       59) wallN=light,salvaged 485  136 poor (0.28041237 0.71958763) *
   15) Hou_own_ref< 0.2994995 6651 2140 poor (0.32175613 0.67824387)
     30) all_hsed>=0.3875 1624  794 non-poor (0.51108374 0.48891626)
       60) faN=l,m,xl,xxl,xxxl 951  383 non-poor (0.59726604 0.40273396)
        120) wallN=strong 532  175 non-poor (0.67105263 0.32894737)
          240) Per_wor_prh>=0.022761 450  132 non-poor (0.70666667 0.29333333) *
          241) Per_wor_prh< 0.022761 82   39 poor (0.47560976 0.52439024) *
        121) wallN=light,oth,salvaged 419  208 non-poor (0.50357995 0.49642005)
          242) famsize< 6.75 261  109 non-poor (0.58237548 0.41762452)
            484) all_hsed>=0.4642857 149   51 non-poor (0.65771812 0.34228188) *
            485) all_hsed< 0.4642857 112   54 poor (0.48214286 0.51785714) *
          243) famsize>=6.75 158   59 poor (0.37341772 0.62658228) *
       61) faN=s,xs 673  262 poor (0.38930163 0.61069837)
        122) roofN=oth,strong 409  195 poor (0.47677262 0.52322738)
          244) Hou_own_tel>=0.034225 188   73 non-poor (0.61170213 0.38829787) *
          245) Hou_own_tel< 0.034225 221   80 poor (0.36199095 0.63800905) *
        123) roofN=light,salvaged 264   67 poor (0.25378788 0.74621212) *
     31) all_hsed< 0.3875 5027 1310 poor (0.26059280 0.73940720)
       62) all_coed>=0.1055555 742  361 non-poor (0.51347709 0.48652291)
        124) wallN=oth,salvaged,strong 540  223 non-poor (0.58703704 0.41296296)
          248) Hou_own_ref>=0.071632 459  168 non-poor (0.63398693 0.36601307)
            496) famsize< 6.25 176   42 non-poor (0.76136364 0.23863636) *
            497) famsize>=6.25 283  126 non-poor (0.55477032 0.44522968)
              994) Hou_own_tel< 0.0193125 105   31 non-poor (0.70476190 0.29523810) *
              995) Hou_own_tel>=0.0193125 178   83 poor (0.46629213 0.53370787) *
          249) Hou_own_ref< 0.071632 81   26 poor (0.32098765 0.67901235) *
        125) wallN=light 202   64 poor (0.31683168 0.68316832) *
       63) all_coed< 0.1055555 4285  929 poor (0.21680280 0.78319720)
        126) famsize< 6.75 2222  640 poor (0.28802880 0.71197120)
          252) roofN=oth,strong 1209  432 poor (0.35732010 0.64267990)
            504) per_kids< 0.6125 686  290 poor (0.42274052 0.57725948)
             1008) all_hsed>=0.0625 410  199 poor (0.48536585 0.51463415)
               2016) all_eled>=0.3666666 83   28 non-poor (0.66265060 0.33734940) *
               2017) all_eled< 0.3666666 327  144 poor (0.44036697 0.55963303)
                 4034) Hou_own_tel>=0.033833 140   66 non-poor (0.52857143 0.47142857) *
                 4035) Hou_own_tel< 0.033833 187   70 poor (0.37433155 0.62566845) *
             1009) all_hsed< 0.0625 276   91 poor (0.32971014 0.67028986) *
            505) per_kids>=0.6125 523  142 poor (0.27151052 0.72848948) *
          253) roofN=light,salvaged 1013  208 poor (0.20533070 0.79466930)
            506) Hou_own_ref< 0.0162245 80   39 non-poor (0.51250000 0.48750000) *
            507) Hou_own_ref>=0.0162245 933  167 poor (0.17899250 0.82100750) *
        127) famsize>=6.75 2063  289 poor (0.14008725 0.85991275) *
```

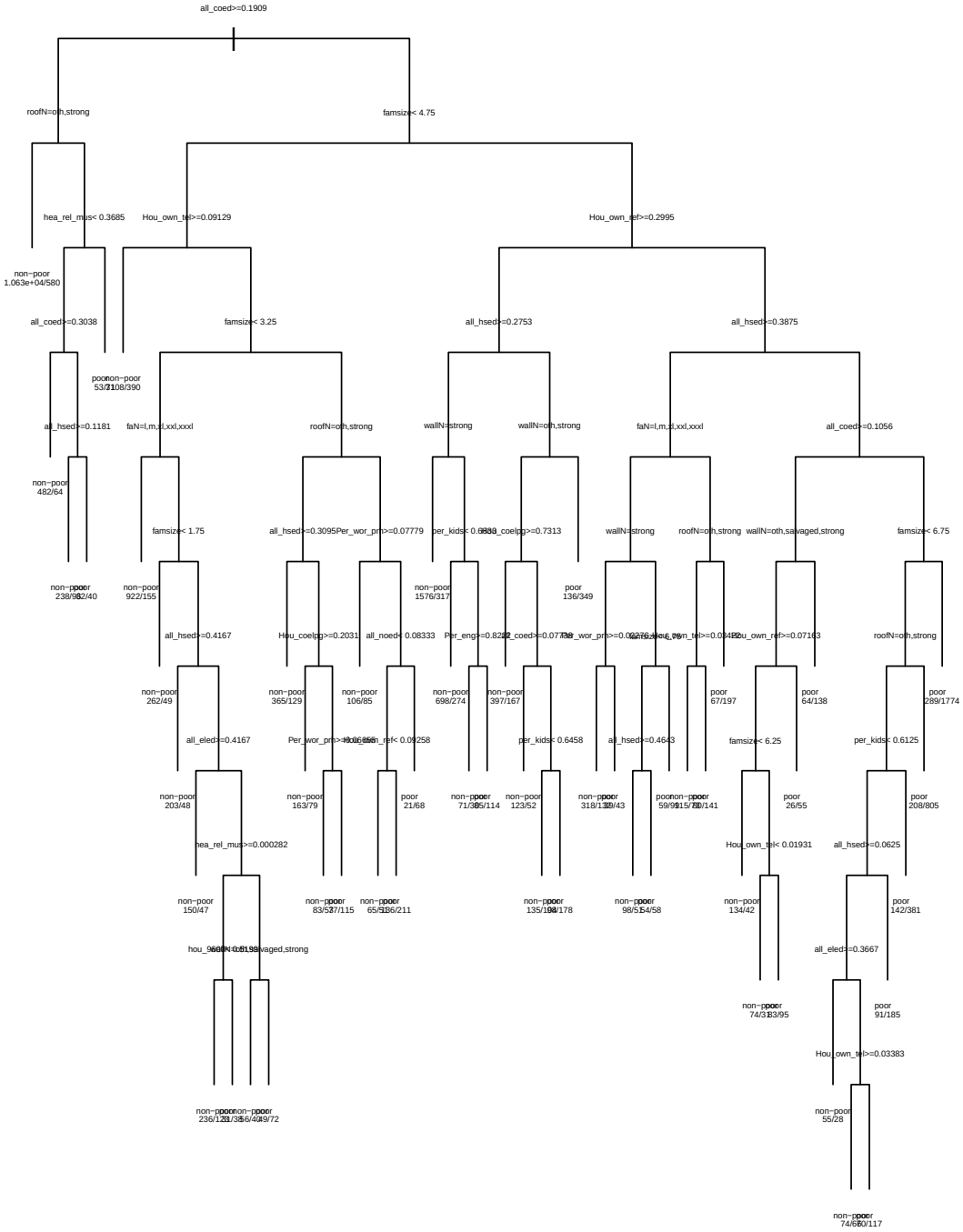


Figure 6.3: Plot for Gini Tree with un-weighted data

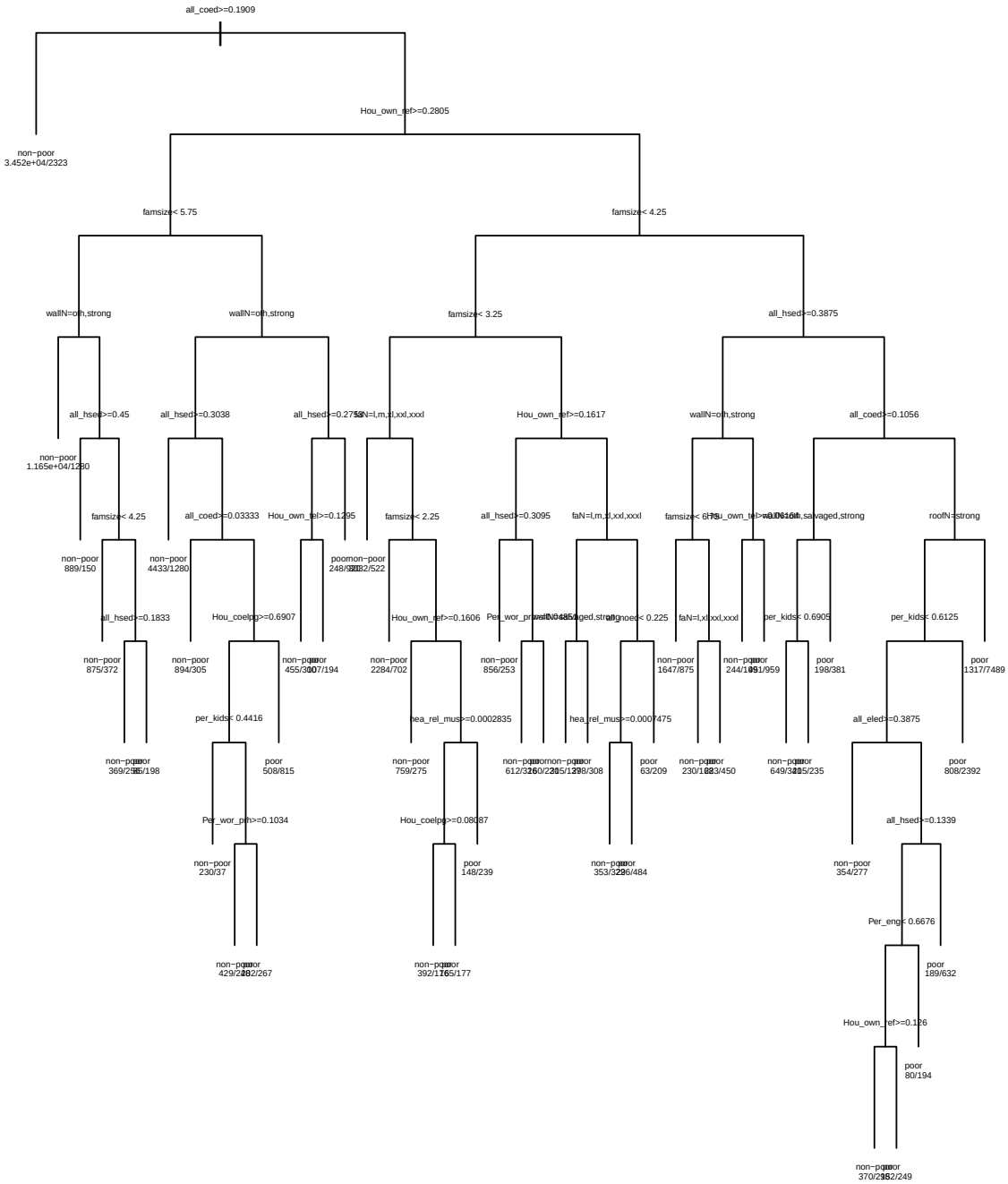


Figure 6.4: Plot for Gini Tree with household duplication data

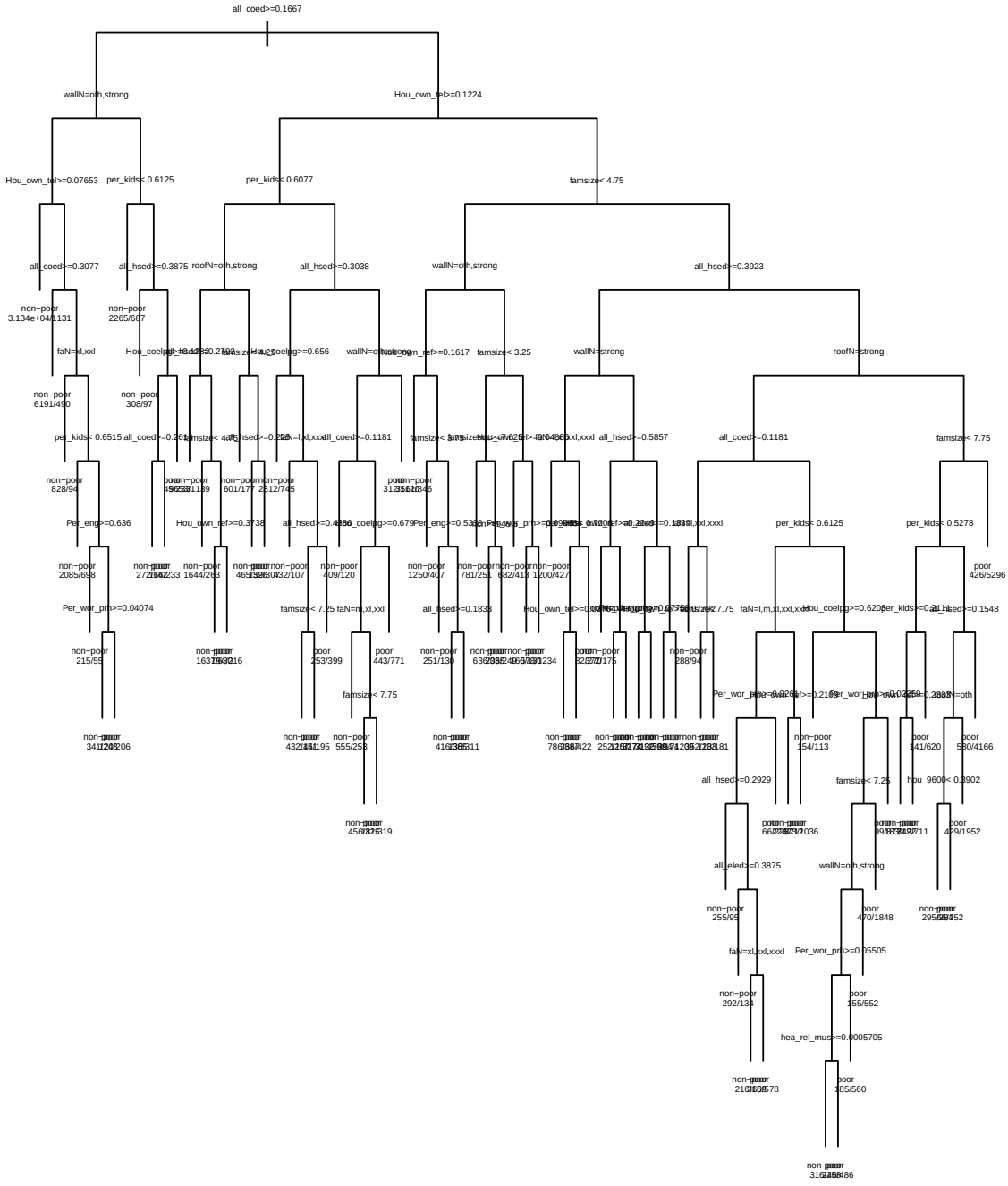


Figure 6.5: Plot for Gini Tree with perperson duplication data (part data into 5 parts)

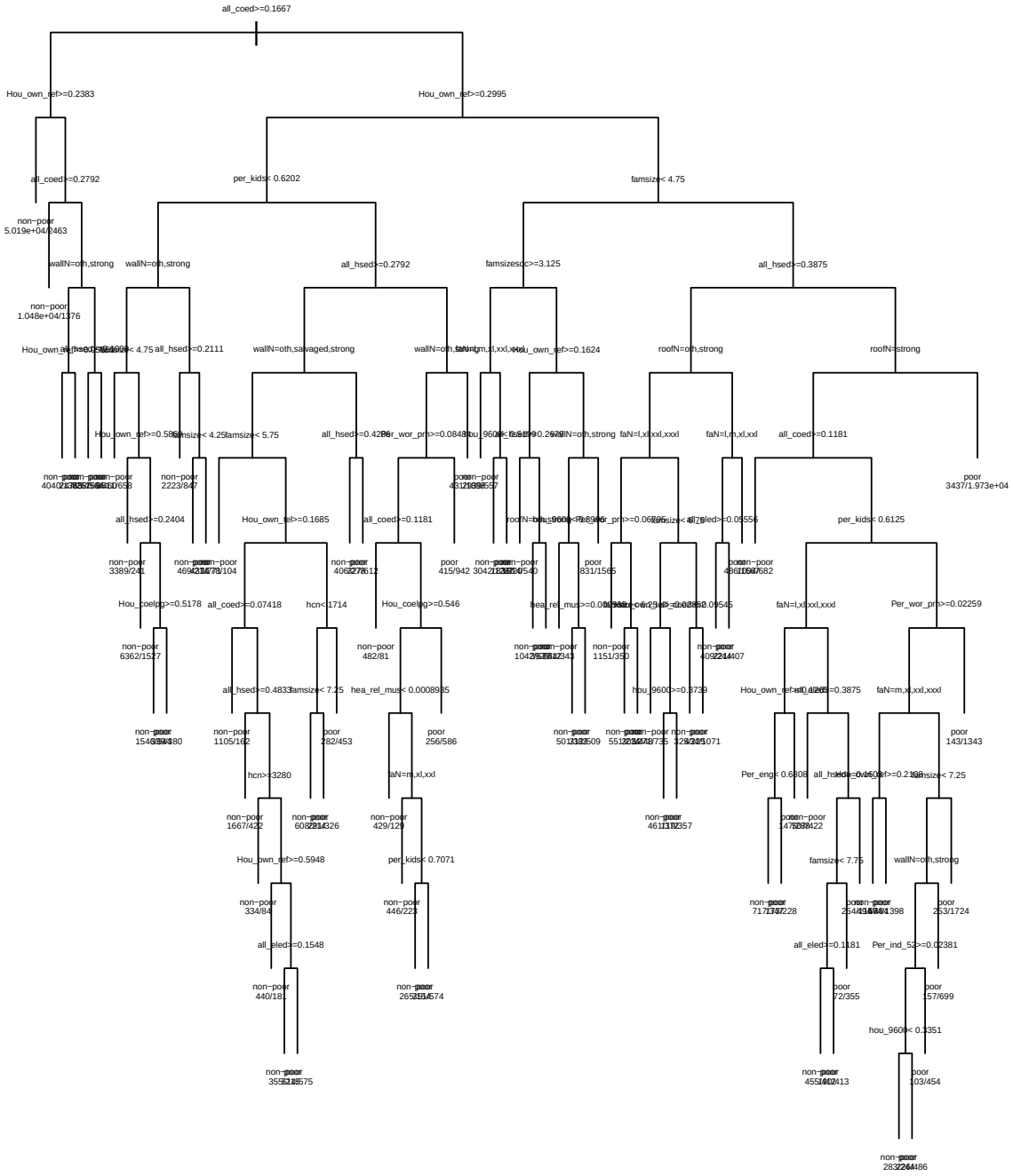


Figure 6.6: Plot for Gini Tree with perperson duplication data (part data into 10 parts)

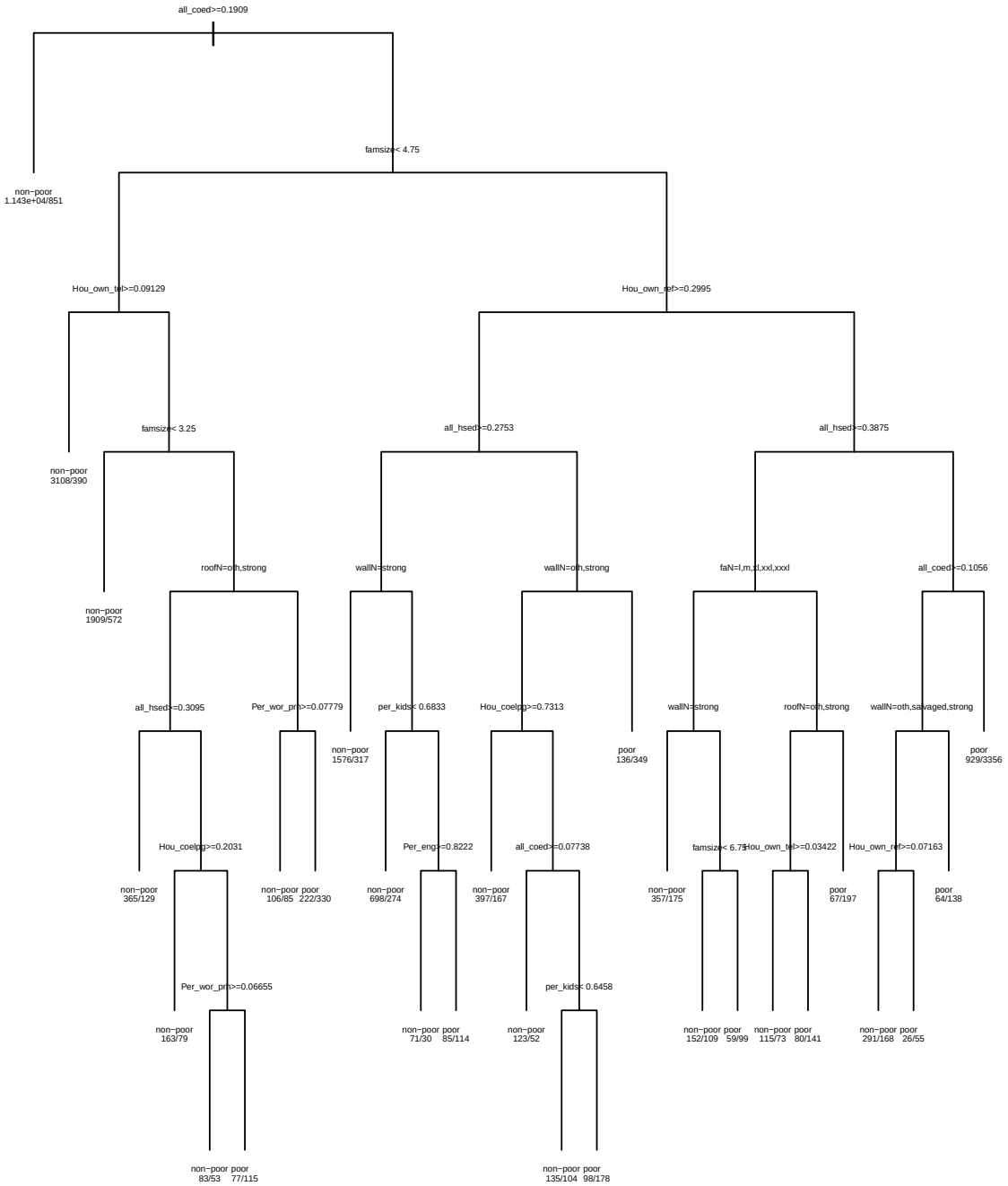


Figure 6.7: Plot for Entropy Tree with un-weighted data

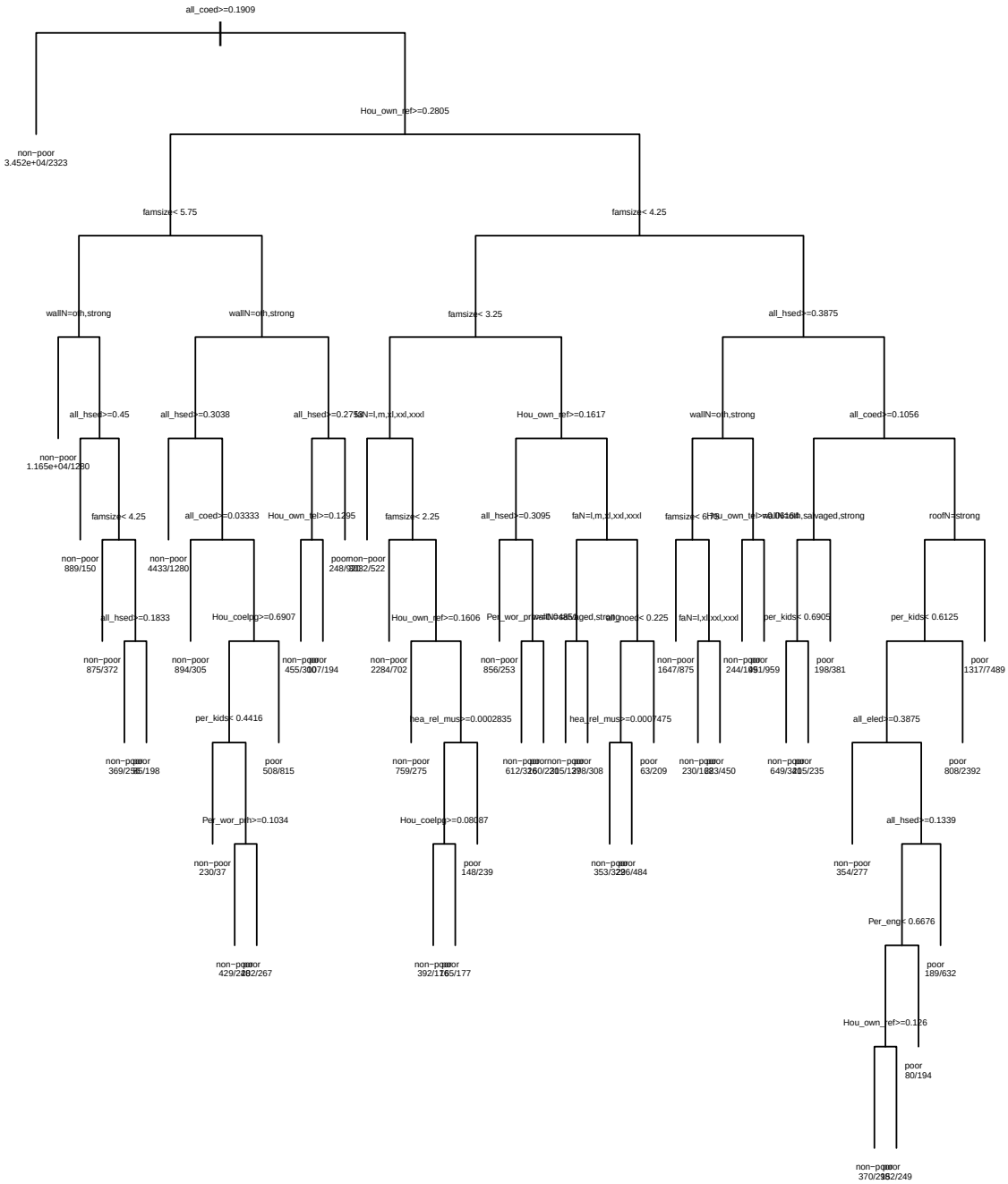


Figure 6.8: Plot for Entropy Tree with household duplication data

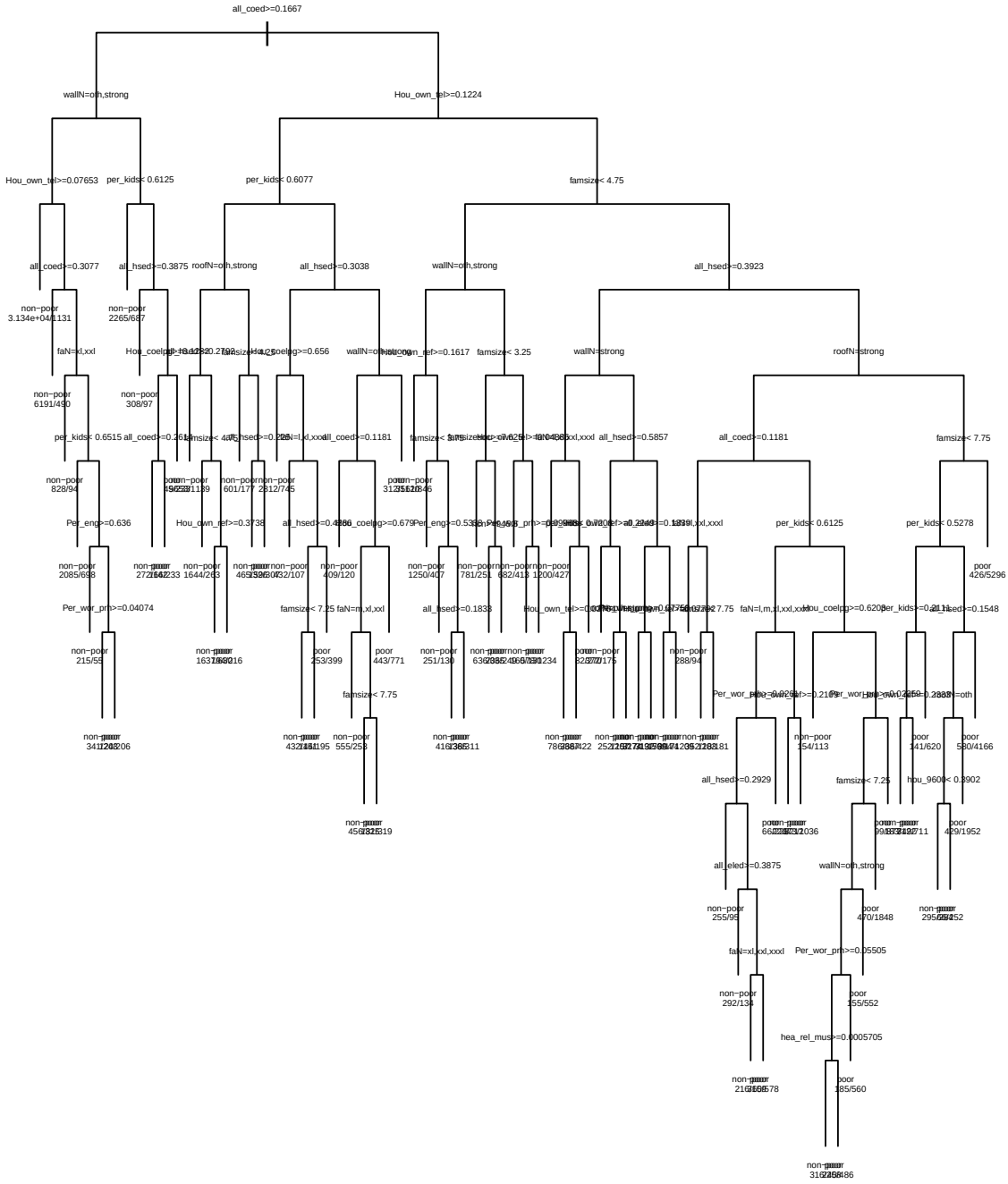


Figure 6.9: Plot for Entropy Tree with perperson duplication data (part data into 5 parts)

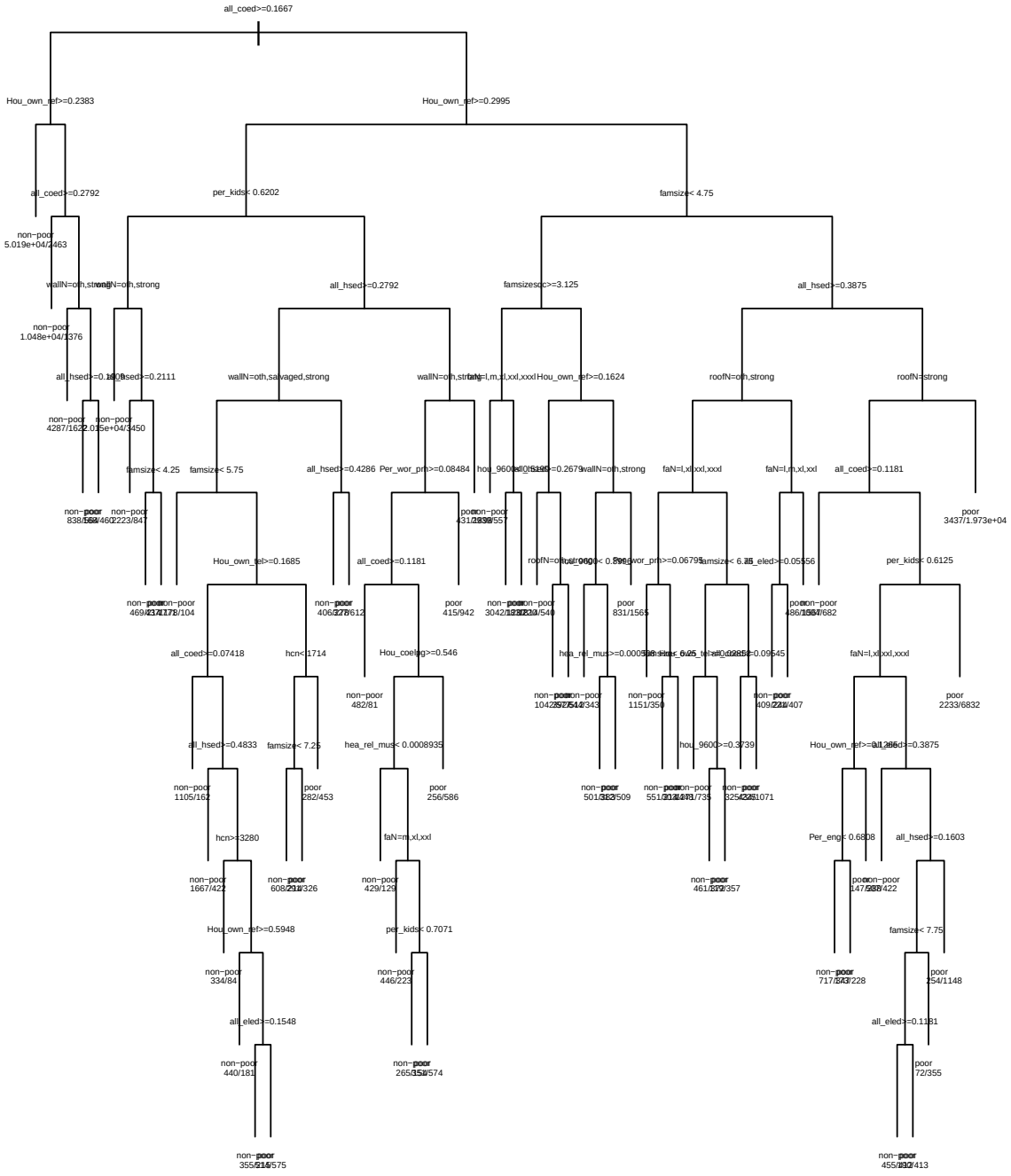


Figure 6.10: Plot for Entropy Tree with perperson duplication data (part data into 10 parts)

# Bibliography

[1] Battese, G. E., Harter, R. M.,& Fuller, W. A. (1988). An Error-Components Model for Prediction of County Crop Areas using Survey and Satellite Data. *Journal of the American Statistical Association*, 83(401), 28-36.

[2] Bain, L. J., & Engelhardt, Max. (2000). Introduction to Probability and Mathematical Statistics. *Boston : PWS-KENT Pub., c1992*. Second Edition.

[3] Bradford, J. P., Kunz, C., Kohavi, R., Brunk C., & Brodley, C. E. (1998). Pruning Decision Trees with Misclassification Costs. *Machine Learning: ECML-98*, 1398/1998, 131-136.

[4] Bradley, P., Andrew. (1997). The use of the Area Under the ROC curve in the evaluation of Machine Learning Algorithms. *Pattern Recognition Society. Published by Elsevier Science Ltd.* 30(7), 1145-1159.

[5] Breiman, L., Friedman, J. H., Olshen, R., & Stone, C. J. (1984). Classification and Regression Trees. *Chapman & Hall, New York*.

[6] Chambers, R. L., & Skinner, C. J. (2003). Analysis of Survey Data. *John Wiley & Sons Ltd*. England.

[7] Cochran, W. G. (2007). Sampling Techniques. *Wiley India Pvt. Ltd.*, NY, 3rd Edition.

[8] Cole, Stephen R., & Hernan, Miguel A. (2008). Constructing Inverse Probability Weights for Marginal Structural Models. *American Journal of Epidemiology*, 168(6), 656-664.

[9] Elbers, C., Lanjouw, J. O., & Lanjouw, P. (2003). Micro-level Estimation of Poverty and Inequality. *Econometrica*, 71(1), 355-364.

[10] Fawcett, Tom. (2006). An Introduction to ROC Analysis. *Pattern Recognition letters*. 27, 861-874.

[11] Gershunskaya, J. B. & Lahiri, P. (2005). Variance Estimation for Domains in the U.S. Current Employment Statistics Program. *Proceedings of the American Statistical Association*

[12] Montgomery, D. C. (2009). Design and Analysis of Experiments. *John Wiley & Sons, Inc.*, NJ, 7th Edition.

[13] Ghosh, M., & Rao, J. N. K. (1994). Small Area Estimation: An Appraisal. *Statistical Science*, 9(1), 55-76.

[14] Jones, G. & Haslett, S. (2005). Local Estimation of Poverty in the Philippines. *The National Statistics Coordination Board of the Philippines.*

[15] Juan-Albacea, Z. V. (2009). Small Area Estimation of Poverty Statistics. *Philippine Institute for Development Studies*, Discussion Paper Series No. 2009-16.

[16] Kearns, M. (1997). A Bound on the Error of Cross Validation Using the Approximation and Estimation Rates, with Consequences for the Training-Test Split. *Neural Computation*, 9(5), 1143-161.

[17] Lohr, Sharon L. (1998). Sampling: Design and Analysis. *Cengage Learning.*

[18] Mingers, John. (1989). An Empirical Comparison of Pruning Methods for Decision Tree Induction. *Machine Learning*, 4, 227-243.

[19] Molina, I. & Rao, J. N. K. (2009). Small area estimation of Poverty indicators. *The Canadian Journal of Statistics*

[20] Rao, J. N. K. (1999). Some recent advances in model-based small area estimation. *Survey Methodology*, 23, 175-186.

[21] Rao, J. N. K. (2003). Small Area Estimation. *Hoboken, N.J. : John Wiley*

[22] Sing, T., Beerenwinkel, N., & Lengauer, T. (2004). Learning Mixtures of Localized Rules by Maximizing the Area Under the ROC Curve.

[23] Therneau, Terry M. & Atkinson, Elizabeth J. (1997). An Introduction to Recursive Partitioning Using the RPART Routines. *Mayo Foundation.*

[24] Thuiller, W., Araujo, M. B., & Lavorel, S. (2003). Generalized Models Vs. Classification tree Analysis: Predicting Spatial Distributions of Plant Species at different Scales. *Journal of Vegetation Science*, 14(5), 669-680.