A Non-parametric Analysis of Micro-data using Classification Trees

Dr. William Chow

23 January, 2013

1. Introduction

1.1 In formulating public policy, certain facts about targeted impact subjects may not be always observable. To be able to gather scattered information and deduce from it patterns, traits and characteristics of the subjects will be of particular importance, especially when desired survey data are not available.

1.2 This paper presents a pattern recognition exercise of local household tenure dispersion based on information extracted from the 2011 census data. Economic and social variables are screened systematically to facilitate the differentiation of household tenure forms. The results allow categorization of the tenure class of a subject individual based on a limited set of observables.

1.3 When analyzing micro data, one often has to deal with categorical (or dichotomous) variables which classify the subject variables into types, categories or classes; e.g. $Y = 1$ for a home purchase decision, and $Y = 0$ for a rental decision.

1.4 If Y is an explanatory variable, it could be incorporated into standard regression models via specification of dummy variables. If Y is the dependent variable, standard regression would not work and logit (binomial or multinomial depending on the number of categories specified) models are natural solutions.

1.5 Logit modeling may not always be preferred. First, interpretation is linked to the probability of occurrence instead of a point-blank answer or outcome. Second, an N-category explanatory variable requires fitting at least $N - 1$ dummy variables in the logit model, and the degrees of freedom will deplete rapidly if the data set contains many qualitative variables.

The views and analysis expressed in the paper are those of the author and do not necessarily represent the views of the Economic Analysis and Business Facilitation Unit.

1.6 Of course, in this latter case with mostly qualitative variables, one can always make pairwise comparison of the factors concerned using simple diagrams. However, ranking the priority of the involved factors will not be easy or possible.

1.7 An alternative is to use classification methods, e.g. CART, to generate interpretable results. The acronym CART means "Classification and Regression Trees" which is a nonparametric method that groups and classifies data to unveil underlying patterns of observed decisions or outcomes.

1.8 The task we have here fits into the situation described above as the census data contains mostly categorical variables. Section 2 outlines the CART methodology. Section 3 describes the particulars of our model and the data. Section 4 presents the findings.

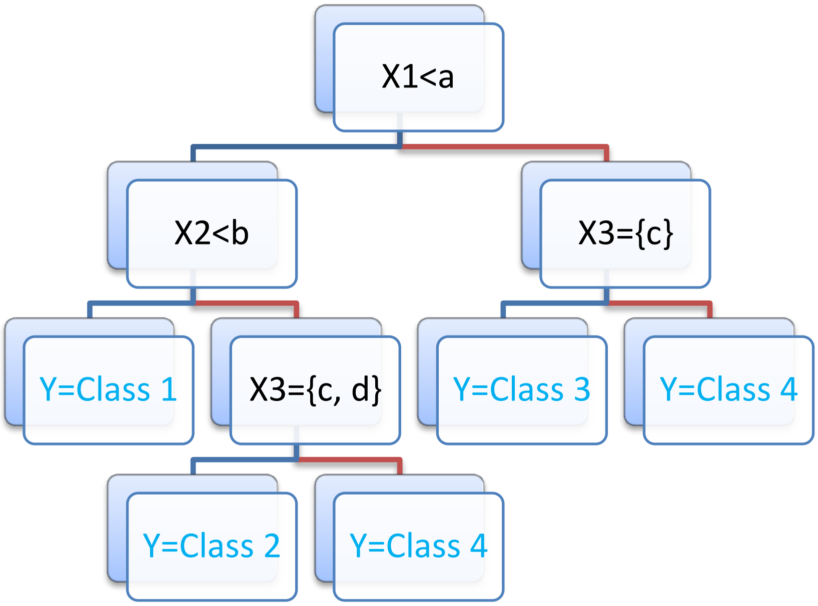
2. CART

2.1 CART is a scheme of sequential data partitioning that facilitates classification of the subject variable – housing tenure decision in our context. When the decision variable is categorical as in here, the model outcome will be configured as a classification tree. The output will be called a regression tree if the decision variable is continuous.

2.2 In essence, the scheme asks a series of questions one at a time until the responses from a partitioned subsample become (more or less) homogeneous. They can then be grouped under the same class. The ordering of the questions is important in that the most "statistically relevant" questions should be raised first. As long as the subsample remains sufficiently heterogeneous, further questions will be asked.

2.3 Fig.1 is an illustration of the data partitioning. The schema resembles a tree with nodes and branches (hence the name classification trees). At each node, the responses of the sample to the highlighted question are sorted. Those with affirmative answer go to the left branch (blue) and those with negative answer go to the right (red). When the partitioned data Y in a node becomes relatively homogeneous, there will be no more splitting and a terminal node stipulates the majority class.

Fig.1 Sample Classification Tree



2.4 In the root node, the first question asks if the continuous variable X1<a. Those who answered yes are grouped to the left branch where a second question (whether X2<b) is raised. If the answer to this second question is yes, the respondent is classified as a type 1 subject. If the answer is a no, a third question (whether the categorical variable X3 is within the set {c, d}) follows, and the subjects answering yes and no are classified as Type 2 and Type 4, respectively.

2.5 To conclude, a Type 2 subject is associated with the observation: X1<a, X2≥b, and X3={c, d}. Interpretation of the right hand side branch is similar. By checking on the response patterns regarding the predictor variables (Xs) one can infer from the classification tree which class the respondent belongs to.

2.6 Major issues concerning the CART methodology include:

- How many steps in general for CART to be applied? Two – a tree growing step as explained above, and a pruning step which refines the tree by eliminating unnecessary/mediocre branches.

- How many questions to raise in the tree growing step? This depends on the number of relevant predictor variables available and their information content.

- How to determine the cut-off points or sets (the values a, b, c and d in the example)? Find from all possible intervals or sets of values observed from the sample that most reduce data heterogeneity at each node.

- How to decide if a node should be split further? To see if incremental reduction in heterogeneity is significant. If negative, stop splitting and stop growing the tree from that node.

- Why prune the tree? Typically, when the test sample or new data sets are fitted to the tree, its accuracy will increase with the complexity (= the number of terminal nodes) up to a certain point before declining again. In a way, there is an eventual cost to over-fitting.

- How to do model checking? There are various methods, and the one used in this paper is to break the sample into two parts – the learning sample and the test sample. The tree will be grown using the learning sample, and it will be pruned to avoid excessive complexity. The test sample will then be applied to the pruned tree to see how much prediction error there is.

- Is it possible to accommodate non-binary splits (answers)? Yes, but it is more involved and will not be pursued here.

2.7 Outline of the Algorithm:

2.7.1 Break the entire sample into 2 parts. Use the learning sample to growth the tree. Let K be the number of predictor variables.

2.7.2 For a continuous predictor variable with M distinct values, there are $M - 1$ possible cutoff (e.g. midpoints between two successive values) points for constructing the inequalities needed. For a categorical predicator variable with M discrete types, the number of possible sets will be $2^{M-1}-1$. E.g. if $X_m = \{1,2,3\}$ is a 3-category variable, the 3 cutoff sets will be:

$$\begin{array}{ccc} \{1\} & vs. & \{2,3\} \\ \{1,2\} & vs. & \{3\} \\ \{1,3\} & vs. & \{2\} \end{array}$$

Note that the number of possible combinations explodes as M increases. Evaluate all possible cutoff points and sets for all K predictor variables.

2.7.3 Let $p(t|j)$ be the probability of a class j subject (i.e. $Y = j$) reaching node t, and t_L , t_R be the left child node and right child node of t respectively. Define the Gini Impurity function as:

$$i(t) = 1 - \sum_{j=1}^{J} p(t|j)^2$$

and the Goodness of split as:

$$\Delta i(t) = i(t) - p(t_L)i(t_L) - p(t_R)i(t_R)$$

The Gini impurity is bounded by [0,1] and the larger the value, the more heterogeneous the sample[1]. The Goodness of split measures the improvement in impurity given a further split. This is the criterion used to gauge if a split should be executed at a certain node.

2.7.4 Start with the root node. Screen through all K predictor variables across all corresponding cutoff points/sets. Evaluate the Goodness of split for each of these possibilities and choose the variable and the associated cutoff point/set that has the largest $\Delta i(t)$. The first question is then asked.

[1] For continuous dependent variable Y, a possible impurity function is the sum of squared errors as encountered in ordinary least squares regressions. The other procedures in growing and pruning the tree are similar. The end product is a regression tree instead of a classification tree.

2.7.5 Move to the two children (left and right) of the root node. Repeat the above assessment for the $K - 1$ variables not yet incorporated in the splitting process.

2.7.6 When a certain node becomes pure or homogeneous, or when the improvement is impurity reduction becomes too small, stop further splitting. Assign a class label to the terminal node that reflects the majority type/class. When this is done, we have what we call a maximal tree.

2.7.7 To avoid undue complexity, prune the maximal tree by collapsing certain branches/subtrees to obtain the optimal tree. This is reminiscent of avoiding overfitting and spurious regressions in econometric analysis.

2.7.8 How the tree should be pruned depends on the cost complexity function[2]. This is expressed as:

$$R_{\alpha}(T) = R(T) + \alpha\left|\tilde{T}\right|$$

where $R(T)$ is the misclassification error of the tree T, α is a parameter that penalizes the complexity of trees, and $\left|\tilde{T}\right|$ is the extent of tree complexity measured as the number of terminal nodes of the tree T. Based on the learning sample alone, the misclassification error gets smaller for larger trees. But the errors would be quite different as new data (the test sample for instance) is fit to a big tree. So the second component of the cost complexity function introduces a penalty to oversized trees.

2.7.9 To find the optimal tree, an algorithm is designed to identify the α and the corresponding tree size. It starts with the maximal tree and loops through all sizes up to the minimal tree containing only the root node. Find the α and the $R(T)$ for all candidate pruned trees. Each of these pruned trees results from the cutting off of the "weakest link" – the node where the elimination of the subtree beneath it results in the smallest change in.

2.7.10 Then fit the test sample to each of these nesting pruned trees and do either of the following: (i) find the one with the minimum estimated error $\hat{R}(T)$, or (ii) find the smallest tree within one standard deviation of the one found in (i). The outcome is the desired optimal tree.

[2] There are other criteria for pruning, see for instance Fürnkranz (1997).

3. The Data

3.1 The raw data of this exercise comes from the 1% sample of the 2011 census. The dependent variable is household tenure choice (Y). We discard dubious records, and group the individual observations into households. There are 24,566 observations, of which 1/3 is set aside as test sample. The tree is grown and pruned using the remaining 2/3 of the sample.

3.2 The choice variable Y is a categorical variable, so the product of the exercise will be a classification tree. Y takes on 5 values, namely, (i) purchase of private quarter, (ii) purchase of public quarter[3], (iii) rental of public quarter, (iv) rental of private quarter, and (v) others. While the word "choice" is used here, the underlying decision of the household may not be totally voluntary in a narrow sense except that it reflects possibly the best choice available given the individual restrictions faced by the household.

3.3 The adopted predictor variables are mostly household specific or (household) head specific. Altogether 19 explanatory variables (6 continuous[4] and 13 categorical) are selected. They are:

• [UHSIZE]	Household size
• [DJHHINC]	Monthly domestic household income
• [HHCOMP]	Household composition
• [WORKPP]	No. of working household members
• [DIST]	Current district of residence (by household head)
• [DUR_HK]	Duration of residence in HK (by household head)
• [AGE]	Age of household head
• [MARIT]	Marital status of household head
• [BORNPL]	Place of birth of household head
• [EDUCNH]	Educational attainment (completed) of head
• [FIELDH]	Highest field of education of household head
• [ACTIV]	Economic activity status of household head

[3] These are principally government subsidized sale flats.

[4] This is a misnomer, as we include in here also variables that are discrete but ordinal. Categorical variables, on the other hand, are those that take on unordered class labels which reflect different classes or categories.

[WHETWK]	Whether the household head is working
[OCCUP]	Occupation (ISCO-08) of household head
[WH_SECEM]	Whether head is having secondary employment
[INTMIG]	Pattern of internal migration (5 year comparison)
[SCHCHILD]	Whether having at least 1 child studying FT
[HELPER]	Whether having a live-in domestic helper
[MONEXP]	Monthly expenses on housing – rent or mortgage

3.4 The data ranges of continuous variables are as specified by the census. For categorical variables, reclassification is applied in certain cases to reduce the burden of dimensionality. For instance, there are 58 categories that indicate the educational attainment of the survey subject in the census. There are thus $2^{58-1} - 1 \cong 144{,}120$ trillion different ways to partition the data using census's definitions of categories. A summary of the redefined range and coverage of the variables is in the Appendix.

3.5 Another issue is the adoption of variables that directly indicate a rental/purchase decision. Including the variable X_k"rental to income ratio", for instance, reveal right away whether the surveyed subject is a landlord or a tenant. This certainly helps the classification process, but would inflate the accuracy of the trees as there is endogeneity between X_k and Y. A compromise is made by merging the monthly rental payment and monthly mortgage payment as a single indicator of monthly expenses on the housing unit. This is the MONEXP shown in the list above.

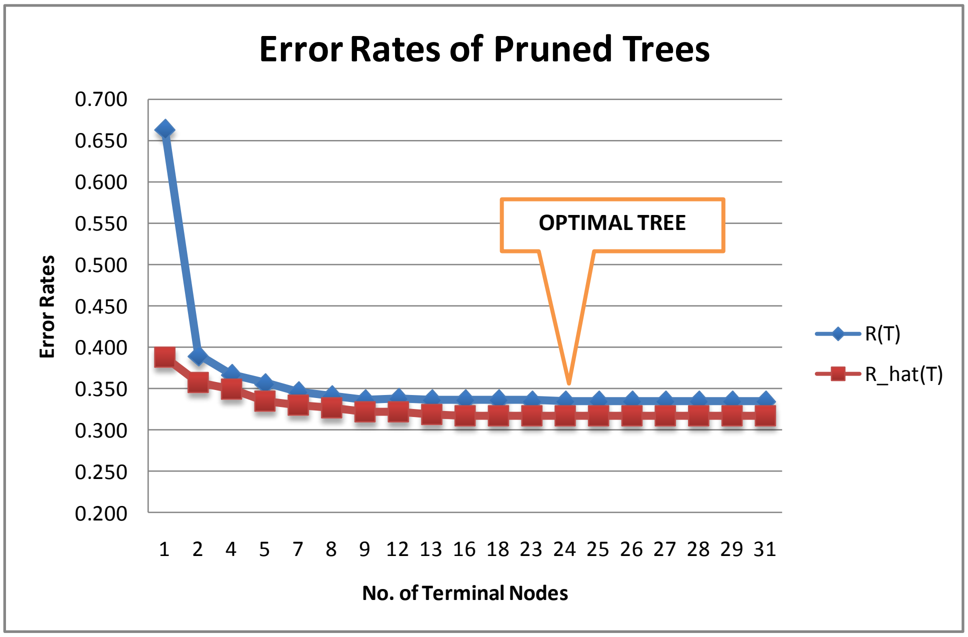
3.6 The tree growing process generates a maximal tree of 61 nodes of which 31 are terminal, i.e. $\left|\tilde{T}_{max}\right| = 31$. As indicated in 3.7.9, the test sample is then applied to the pruned trees and the one with the minimum estimated error rate $\hat{R}(T)$ is chosen as the optimal tree. The results are discussed in the next section.

4. The Findings

4.1 Fig.2 shows the result of the tree pruning process. There are a total of 19 nesting pruned trees, including the maximal tree and the root minimal tree. Based on the minimum estimated misclassification error, the optimal tree is the one with 24 terminal nodes.

4.2 The estimated error rate of the optimal tree, illustrated in Fig.3, is 0.323. This means that the optimal tree has about a 68% chance of correctly classifying a subject regarding his tenure choice. This compares with the naïve probability of 20% of getting it right from wild guess (the prior probability of a correct classification is 1 out of 5).

Fig.2 Error Rates of the Pruned Trees



4.3 The optimal tree is illustrated in Fig.3. The left branches (answer yes) are in blue and the right branches (answer no) are red. The first factor that discriminates tenure choice is monthly expenditure on living quarters (rental or mortgage payment), with a threshold of around $2,800. The majority of those paying less than that amount are tenants of public housing (left branch from the root node), while those paying more than that (right branch from the root node) are mostly homeowners of private housing.

4.4 Moving down the tree sequentially gives a string of characteristics describing a particular class. For instance,

- A household spending less than $2,796 a month on the living unit; whose head either works in the primary sector or is engaged in non-office work or elementary operations; with the residence locating in any districts except Western, Central, Wanchai and Yau Tsim Mong; whose household head has been living in HK for less than 8 years is likely to be a tenant of public housing.

- A household spending more than $2,796 a month on the living unit; has no internal migration pattern that fits with the designated classes; resides in either Wong Tai Sin, Kwun Tong, Sai Kung, Tseung Kwan O, Shatin, Tai Po, or the North; has monthly household income of less than $44,819 has a good chance of being a homeowner of government subsidized housing.

- A household spending more than $2,796 a month on the living unit; has no internal migration pattern that fits with the designated classes; resides in either Wong Tai Sin, Kwun Tong, Sai Kung, Tseung Kwan O, Shatin, Tai Po, or the North; has monthly household income of more than $44,819; whose head works as managers, professionals or is engaged in the primary sector; is likely to be a homeowner of private housing.

- A household spending more than $2,796 a month on the living unit; has no internal migration pattern that fits with the designated classes; resides in any district except Wong Tai Sin, Kwun Tong, Sai Kung, Tseung Kwan O, Shatin, Tai Po, or the North; whose household head has been living in HK for less than 15 years and is born neither in HK or China is likely to be a tenant of private housing.

4.5 In sum, monthly expenditure on the quarter, district of residence, household income, household head's residence duration in HK, his/her occupation and educational attainment are factors that aid the classification. Household size and household composition, on the other hand, are relatively unimportant.

4.6 As the splits in our algorithm are binary, there will just be a single cut-off point instead of a sequence of thresholds. Should a modified scheme be developed to accommodate such flexibility, the classification would make better use of the heterogeneous information contained in the data. In addition, some of the cut-off figures are likely to be biased by the large amount of CSSA recipients who reside in public housing.

5. Conclusion

5.1 CART is a nonparametric tool that classifies subjects into different types based on screening various information and characteristics of the reviewed subjects. The advantage is that it does not rely on probabilistic assumptions and gives interpretable patterns of the classification outcome.

5.2 Since it is nonparametric, test of significance like those used in regression studies is not possible. Also, the identified patterns indicate association which may or may not imply causality.

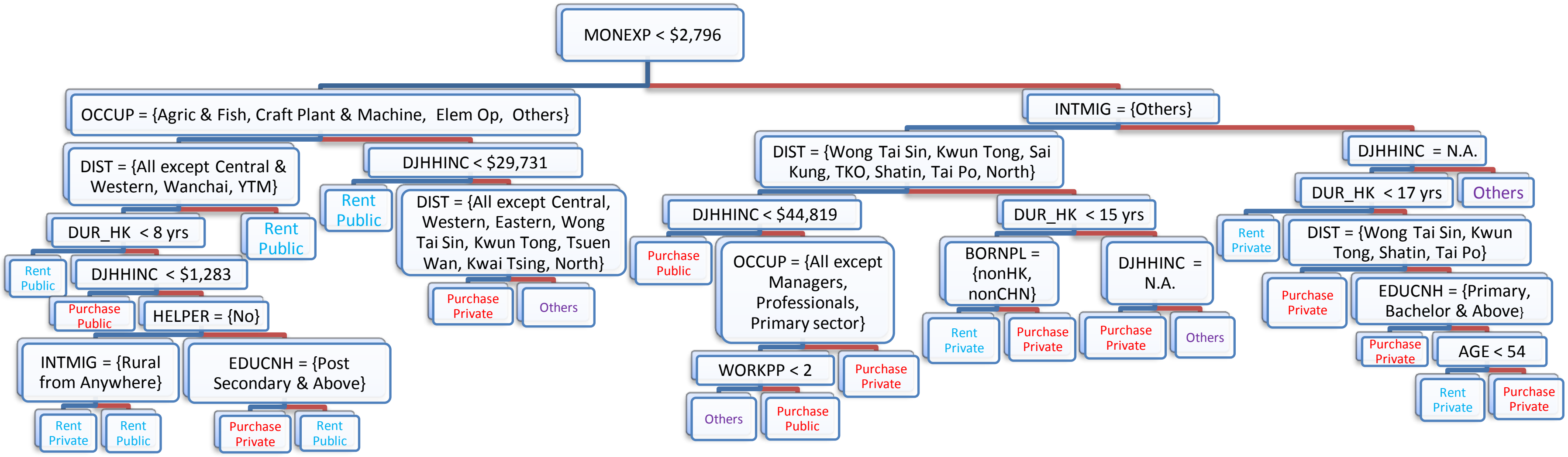
5.3 The quality of the classification depends on the quality of the data. Meaningful patterns stem from inclusion of relevant variables and data coverage.

Reference

Breiman, L., Friedman, J.H., Olshen, R., Stone,C.J. (1998) *Classification and Regression Trees*. Chapman and Hall/CRC, New York.

Fürnkranz, J. (1997) *Pruning Algorithms for Rule Learning*. Machine Learning, 27, 139-172.

Fig.3 The Optimal Pruned Tree



Appendix

Data Range and Categories of Predictor Variables

Variables	Range/Categories
UHSIZE	$1, 2, \cdots, 97, 98$ and over
DJHHINC	$\$1, \cdots, \$150{,}000$ and over; $\$0$ and N.A.
HHCOMP	Couples; Couples with unmarried children; Single parent with unmarried children; Couples with parents; Couples with parents and unmarried children; Others
WORKPP	$0, 1, \cdots, 6$ and over

DIST	Central and Western;
	Wanchai;
	Eastern;
	Southern;
	Yau Tsim Mong;
	Sham Shui Po;
	Kowloon City;
	Wong Tai Sin;
	Kwun Tong;
	Sai Kung, Tseung Kwan O;
	Tsuen Wan, Kwai Tsing;
	Yuen Long, Tuen Mun;
	North;
	Shatin, Tai Po;
	Others

DUR_HK	Less than 1 year;
	1 to less than 2 years;
	$\vdots$
	20 years and over

AGE	$0, 1, \cdots, 100$ and over

MARIT	Never married;
	Now married;
	Widowed;
	Divorced;
	Separated

BORNPL	HK;
	China;
	Elsewhere

EDUCNH	No schooling;
	Pre-primary;
	Primary;
	Secondary;
	Post secondary;

	Diplomas, certificates and sub-degrees; Bachelors and equivalent; Postgraduates

FIELDH	Basic Programmes;
	Arts and Humanities;
	Social Sciences;
	Life and Health Sciences;
	Math and Physical Sciences;
	Education;
	Accounting and Business;
	Computing;
	Engineering – Civil, Mechanical, Electronic and Electrical;
	Law;
	Architecture;
	Transport, Textile, Manufacturing and Processing;
	Journalism, Social Work;
	Others

ACTIV	Employees;
	Employers;
	Self employed;
	Unemployed;
	Inactive

WHETWK	Yes and No

OCCUP	Managers and Administrators;
	Professionals and Associate professionals;
	Clerical support;
	Service and Sales workers;
	Agriculture and Fishery;
	Craft, Plant and Machinery;
	Elementary operations;
	Others

WH_SECEM	Yes, No and N.A.

INTMIG	Urban (from Urban);
	Urban (from New Town);
	Urban (from Rural);
	New Town (from Urban);
	New Town (from New Town);
	New Town (from Rural);
	Rural (from Urban);
	Rural (from New Town);
	Rural (from Rural);
	HK from China;
	Others

SCHCHILD	Yes and No
HELPER	Yes and No
MONEXP	$1, ⋯, $99,998 and over; $0 and N.A.