

# Determinants of House Price: A Decision Tree Approach

Gang-Zhi Fan, Seow Eng Ong and Hian Chye Koh

[Paper first received, August 2003; in final form, January 2006]

**Summary.** The hedonic-based regression approach has been utilised extensively to investigate the relationship between house prices and housing characteristics. However, this approach is subject to criticisms arising from potential problems relating to fundamental model assumptions and estimation such as the identification of supply and demand, market disequilibrium, the selection of independent variables, the choice of functional form of hedonic equation and market segmentation. This study introduces and utilises an alternative approach—the decision tree approach, which is an important statistical pattern recognition tool. Using the Singapore resale public housing market as a case study, the article demonstrates the usefulness of this technique in examining the relationship between house prices and housing characteristics, identifying the significant determinants of housing prices and predicting housing prices. The built tree shows that homebuyers are more concerned about the basic housing characteristics of two- and three-room flats or four-room flats such as floor area, model type and flat age. However, homebuyers of five-room flats pay more attention to floor level in addition to the basic housing characteristics. In addition, homebuyers of executive apartments are less concerned about basic quantitative characteristics and have higher housing consumption expectations and pay more attention to ‘quality’ and service characteristics such as recreational facilities and the living environment.

## 1. Introduction

Over the past three decades, the hedonic-based regression approach has been utilised extensively in the housing market literature to investigate the relationship between house prices and housing characteristics, such as structural, neighborhood and locational characteristics. The primary reasons for such extensive application are for analysing household demand for these characteristics as well as constructing housing price indices (see, for example, Can, 1992; Sheppard, 1999). However, this approach is subject to criticisms

arising from potential problems relating to fundamental model assumptions and estimation such as the identification of supply and demand, market disequilibrium, the selection of independent variables, the choice of functional form of hedonic equation and market segmentation; these problems have been of great concern in the literature (for a review, see Malpezzi, 2003; Sheppard, 1999). This study is motivated by the necessity and importance of examining the relationship between house prices and housing characteristics, while we introduce and utilise a non-parametric approach—the

*Gang-Zhi Fan is in the Research Institute of Economics & Management, Southwestern University of Finance and Economics, 55 Guanghua Cun Street, Chengdu, Sichuan 610074, China. E-mail: gzfan@swufe.edu.cn, and the Department of Real Estate, School of Design and Environment, National University of Singapore, 4 Architecture Drive, Singapore 117566. E-mail: rstfg@nus.edu.sg. Seow Eng Ong is in the Department of Real Estate, School of Design and Environment, National University of Singapore, 4 Architecture Drive, Singapore 117566. E-mail: rstongse@nus.edu.sg. Hian Chye Koh is in the School of Business, SIM University, Clementi Road, Singapore 599491. E-mail: hckoh@unisim.edu.sg.*

decision tree approach, which is an important statistical pattern recognition tool. Using the Singapore resale public housing market as a case study, we demonstrate the usefulness of this technique in examining the relationship, identifying the significant determinants of housing prices and predicting housing prices. Moreover, we also show that this technique provides a probable approach to the empirical examination of the utility tree theory developed by Strotz (1957, 1959).

Jain *et al.* defined statistical pattern recognition as

the study of how machines can observe the environment, learn to distinguish patterns of interest from their background, and make sound and reasonable decisions about the categories of the patterns (Jain *et al.*, 2000, p. 4).

In recent years, pattern recognition has gained much attention and increasing popularity in the commercial world. Pattern recognition techniques and tools have also developed rapidly over the past two decades.<sup>1</sup> However, only a handful of applications in academic research, in particular in real estate academic research, have been reported. In effect, compared to traditional regression models, several advantages of the decision tree-based methodology are remarkable (see, for example, Murthy, 1998; Berry and Linoff, 1997). First, tree approaches can deal with classification problems as well as regression problems, while only a few assumptions with regard to the data distribution and the built model require to be made in the pattern recognition process. Secondly, on the whole, the established decision tree can be easily interpreted by an end user and allows him/her to evaluate the results and identify major attributes in observations, as the tree is produced by straightforward rules that partition data in the target field (variable) into different classes. Thirdly, tree approaches are powerful for analysing the linear or non-linear relationship between the dependent variable and independent variables and for identifying the most significant independent variables in predicting the target

variable and, therefore, can supplement and complement traditional regression methods. Last but not least, tree approaches can not only cope with continuous variables, but also with categorical variables in classification and regression. Due to these advantages, the tree-based method is used to examine the relationship between house prices and housing characteristics and to identify the significant determinants of house prices.

The remainder of this paper is organised as follows. Section 2 provides a brief review of the pertinent literature. Section 3 introduces a theoretical foundation for the use of tree-based models. Section 4 covers the decision tree methodology and its modelling tools. Section 5 briefly introduces the Singapore resale public housing market. Section 6 explores the application of decision tree tools in the real estate domain. Section 7 draws the conclusions and highlights some limitations of tree-based methods.

## 2. Literature Review

Housing as a heterogeneous good can be viewed as a package of inherent characteristics relevant to housing structure, neighborhood and location. By regressing the transaction prices of housing against corresponding housing characteristics, one can estimate the contribution of the characteristics to prices—i.e. the implicit market valuation of these characteristics—and identify the significant characteristics affecting the prices. Studies using the hedonic-based regression approach are too numerous to summarise here. Some of the papers are: Goodman (1978), Li and Brown (1980), Rodriguez and Sirmans (1994), So *et al.* (1997), Benson *et al.* (1998), Fletcher *et al.* (2000a, 2000b), Chau, Ng and Hung (2001), Chau, Ma and Ho (2001), Ong *et al.* (2003) and Berry *et al.* (2003). These studies were carried out within the multiple regression analysis framework, which is often criticised on account of its stringent assumptions. Decision tree modelling provides a potential alternative technique to multiple regression analysis (Breiman *et al.*, 1984).

Over the past two decades, a substantial technical literature on decision trees has emerged. It covers topics such as the derivation of splitting rules (for example, Ben-Bassat, 1987; Breiman *et al.*, 1984; Quinlan, 1986), the construction of univariate and multivariate trees (for example, Breiman *et al.*, 1984; Murthy *et al.*, 1993), the obtaining of the right-sized trees (for example, Breiman *et al.*, 1984; Crawford, 1989; Quinlan, 1987) and the evaluation of tree induction methods (for example, Michie, 1986; Goodman and Smyth, 1988). However, since these studies are beyond the scope of this paper, they are not discussed here (for a comprehensive survey of the technical literature on decision trees, see Murthy, 1998; Safavin and Landgrebe, 1991). Instead, this section focuses on the applications of decision tree methods in relevant academic research areas.

In the financial research domain, a number of studies have attempted to take advantage of decision tree techniques for financial and portfolio analyses. For example, Frydman *et al.* (1985) applied the recursive partitioning algorithm (RPA), a frequently used algorithm for building decision trees, to the classification of the financial distress of firms. More recently, Kao and Shumaker (1999) employed RPA techniques to examine the relationship between several macroeconomic variables and the value–growth spread in order to avoid the stringent assumptions of multiple regression analysis. Sorensen *et al.* (2000) also attempted to use a decision tree approach to deal with the problem of stock selection. It is shown that the constructed stock portfolios using the tree approach perform significantly better than those built by simple stock screening and ranking models.

However, no academic study was found to have used decision tree techniques to examine the issues in real estate research,<sup>2</sup> although some recent studies have applied other pattern recognition techniques such as neural networks and clustering analysis. For example, several studies have compared the predictive performance of artificial neural networks and multiple regression analysis

for property values (for example, Nguyen and Cripps, 2001; Do and Grudnitski, 1992; Tay and Ho, 1991/92; Worzala *et al.*, 1995). Gallagher and Mansour (2000) and Goetzmann and Wachter (1995) attempted to employ clustering algorithms to examine real estate market dynamics and real estate portfolio analysis, respectively. McCluskey and Anand (1999) made use of intelligent hybrid techniques in the mass appraisal of residential properties.

### 3. Theoretical Foundation

Before we introduce the technical features of decision trees, a theoretical framework for our decision tree model is first developed in this section based on Strotz (1957, 1959) and Apps (1973). Strotz (1957, 1959) developed a utility tree theory to help to explain and understand consumer behaviors and preferences, in which the utility function form is assumed to be slightly different from the widely used utility function forms. The theory assumes that the individual or household income allocation decision follows a hierarchical tree structure. That is, an individual or household first allocates income to commodity groups and then to commodities within each group based on their characteristics. Correspondingly, their utility function can be written as a hierarchy of satisfaction functions according to the commodity groups such as food, housing, and education.

Apps (1973) further developed a model of housing demand based on the above utility tree concept. In this model, housing commodities are partitioned into three hierarchical levels: namely, from the single housing group like food and education groups at the same level, to three branches—space, location and internal services or characteristics—and at the third level to the component characteristics of these branches. That is, housing commodities are assumed to be functionally separable so that they can be grouped to form a hierarchical structure. For a given budget constraint, the household housing decision involves this hierarchical structure and depends on the priority ordering and

saturation levels of wants for the housing characteristics within the structure.

Following the theoretical line, consider a separable utility function for housing consumption

$$U = U[V_1(x_1), V_2(x_2), \dots, V_n(x_n)] \quad (1)$$

where,  $x_k = (x_{k1}, x_{k2}, \dots)$ ,  $k = 1, 2, \dots, n$ , represents the vector for a certain type or group of housing attributes, and  $V_k$  represents the corresponding sub-utility function.

This expression suggests that individual housing attributes may be partitioned among broader groups or branches of services such as space, location, internal and external structure features, and that the attributes within any branch constitute the variables of a branch function. Under this specification, the utility for housing depends, in an immediate sense, on a set of intermediate variables or functions,  $V_1, V_2, \dots, V_n$ , each of which is further related to its attribute variables, for example,  $x_{k1}, x_{k2}, x_{k3}, \dots$ .

Expression (1) is analogous to that of the utility tree introduced by Strotz (1957, 1959) and a similar functional form was proposed in Wilkinson (1973), who also viewed housing as a hierarchy of services which are available from a house. Given the assumption that households know their broad requirements for dwelling services when purchasing a dwelling, expression (1) implies that the housing purchasing decision depends on household preference or priority orderings for dwelling attributes or services and that, therefore, the attributes or services probably have different importance in the determination of housing prices.

Maximising this utility function subject to the budget constraint yields the hedonic prices of dwelling attributes. However, for the purpose of this paper, we shall not address the specification of the model and the conditions of utility function maximization and separability. In particular, we do not impose any restrictions on the function relationships between individual attribute variables and housing utility and shall allow any probable relationships, for example from

linear to non-linear relationships, due to a set of intermediate functions being incorporated into expression (1). In the subsequent section, we will demonstrate that decision tree algorithms provide an attractive approach to identifying significant household preference orderings for dwelling attributes or services in housing purchasing decisions.

#### 4. The Decision Tree Algorithm

The purpose of this section is to provide a brief description regarding decision tree methodology, with emphasis on basic techniques and essential features. A decision tree algorithm

works by splitting a dataset in order to build a model that successfully classifies each record in terms of a target field or variable (Woods and Kyril, 1997, p. 42).

Decision trees, built through a so-called recursive partitioning process, provide a powerful tool for the description, classification, regression and prediction of data.

The SAS (Statistical Analysis Software) Institute proposes that a data mining methodology should,<sup>3</sup> on the whole, include five stages: sample, explore, manipulate, model, and assess (SAS Institute, 1998). Specifically, sampling partitions the dataset into different subsets for model construction and model validation, while exploration and manipulation review the data in order to enhance understanding of them and to transform the data, respectively. The modelling and assessing stages are the actual construction of the model and the assessment of the constructed model. In this study, we also basically follow these steps to build a tree-based housing price model, while for the purpose of this paper, our focus is on the modelling stage.

Three of the most extensively used algorithms for building decision trees are CHAID, CART and C4.5/C5.0 algorithms. The chi-squared automatic interaction detection (CHAID) algorithm is restricted to categorical variables, while both the classification and regression trees (CART) and C4.5/C5.0 algorithms are utilised to analyse continuous variables as well as categorical variables and

their fundamentals are also quite analogous. (For a comprehensive survey on these algorithms, see, for example, Berry and Linoff, 1997; Breiman *et al.*, 1984; Quinlan, 1986, 1993.) These algorithms assess a node-splitting criterion using an F-test or a chi-squared test, or based on the reduction in variance, entropy, or Gini impurity measure. This study is to a large extent related to the CART algorithm (Breiman *et al.*, 1984), which can not only be used for classification problems (classification trees), but also, more importantly, cope with regression problems (regression trees).

#### 4.1 Tree Growing

Decision trees are constructed usually via tree growing and pruning. In tree growing, tree algorithms search for the splitter (the independent variable) that partitions the sample in such a way that the difference with regard to the dependent variable is greatest among the divided sub-groups. This process of partition begins from the root node, which contains the entire training sample (see note 4), and continues until no further split can produce statistically significant differences in the dependent variable in the new sub-groups or sub-sub-groups. The sub-groups and sub-sub-groups are known as nodes. When no statistically significant splitter can be identified for a given node, then it is often called a leaf or terminal node; *vice versa*, it is an internal node.

CART makes use of an impurity-based split selection method—more specifically, an impurity function—to determine the best node-splitting criterion or splitter at every node. There are several different functions measuring node impurity such as the Gini index and entropy function but, basically, a low value of impurity measure indicates the predomination of a single class within a set of cases, while a high value suggests an even distribution of classes. Thus, the best splitting criterion at a node should be the one leading to the largest reduction in impurity.

Let  $i(v)$  represent some impurity measure of a root node  $v$ . A candidate splitter  $s$  is assumed

to partition the cases in node  $v$  into left-branch node  $v_l$  and right-branch node  $v_r$  (for clarification, here we only consider the simplest case—a binary structure). That is,  $v$  is the parent of  $v_l$  and  $v_r$ ; and  $v_l$  and  $v_r$  are respectively left child and right child of  $v$ . Then, the impurity change due to binary splitting can be written as

$$\begin{aligned} \Delta i(s, v) = & i(v) - p(v_l)i(v_l) \\ & - p(v_r)i(v_r) \end{aligned} \quad (2)$$

where,  $p(v_l)$ , and  $p(v_r)$  denote respectively the probabilities of a case going into nodes  $v_l$  and  $v_r$  due to the splitter  $s$ ; and  $i(v_l)$  and  $i(v_r)$  represent respectively the impurity measures of nodes  $v_l$  and  $v_r$ .

Expression (2) measures the goodness of the splitter  $s$ . The splitter with maximising  $\Delta i(s, v)$  is ultimately chosen to split node  $v$  from all potential splitters  $s$ . The same procedure can be utilised to split the two child nodes and other offspring nodes until all the leaf nodes contain pure samples from only one class. This process is often known as tree growing.

For regression trees, where the dependent (response) variable,  $Y$ , is continuous, an obvious candidate for node impurity is the mean squared error—i.e. the within-node variance of  $Y$

$$i(v) = \sum_{j \in v} [Y_j - \bar{Y}(v)]^2 \quad (3)$$

where,  $\bar{Y}(v)$  is the mean of  $Y_j$  within node  $v$ .

For any split of node  $v$  into its two child nodes  $v_l$  and  $v_r$  based on a splitter  $s$ , the change in error measure may be simply written as

$$\Delta i(s, v) = i(v) - i(v_l) - i(v_r) \quad (4)$$

where, weights need not be taken into consideration.

Furthermore, similar to classification trees, the best splitter should maximise  $\Delta i(s, v)$ . The same procedure also applies to other node splits (see also Zhang and Singer, 1999).

## 4.2 Tree Pruning

The preceding growing process usually results ultimately in overlarge trees—i.e. the problem of overfitting—which may not make reasonable statistical inference as the leaf nodes contain too few cases. As a result, the constructed trees need to be pruned—that is, surplus leaves and branches are removed in order to obtain right-sized trees. CART makes use of a pruning approach to improve the performance of a constructed tree.

Define the quality, more specifically the misclassification rate, of a tree  $T$  as

$$R(T) = \sum_{v \in \tilde{T}} R(v) = \sum_{v \in \tilde{T}} p(v)r(v) \quad (5)$$

where,  $R(v)$  is the estimated misclassification rate of leaf node  $v$ ,<sup>4</sup>  $r(v)$  the estimated within-node misclassification rate of leaf node  $v$ ; and  $\tilde{T}$  the set of leaf nodes of tree  $T$ . This definition indicates that the overall misclassification rate of a constructed tree can be evaluated from the estimated misclassification rates of its terminal nodes. However, for regression trees, equation (5) should be rewritten as

$$R(T) = \sum_{v \in \tilde{T}} i(v) \quad (6)$$

Furthermore, we also take into consideration a cost-complexity measure of tree  $T$  defined as

$$R_\alpha(T) = R(T) + \alpha|\tilde{T}| \quad (7)$$

where,  $\alpha \geq 0$  denotes the complexity cost of a terminal node; and  $|\tilde{T}|$ , measuring the complexity of tree  $T$ , is the number of leaf nodes.

Accordingly,  $\alpha|\tilde{T}|$  measures all penalising costs for the tree complexity (tree size). That is, the cost-complexity,  $R_\alpha(T)$ , is the sum of the estimated misclassification rate of the tree and the cost penalty for its complexity.

Then the aim of a pruning process is to find a minimising sub-tree  $T(\alpha)$ , for every value of  $\alpha$ , satisfying

$$R_\alpha[T(\alpha)] = \min_{T \subset T_{\max}} R_\alpha(T) \quad (8)$$

where,  $T_{\max}$  represents the fully grown tree; and  $T \subset T_{\max}$  means that  $T$  is a sub-tree of  $T_{\max}$ .

More specifically, for every internal node  $v$  in  $T$ ,  $v \notin \tilde{T}$ , we can find a value for  $\alpha$  from the following formula

$$\alpha_v = \frac{R(v) - R(\tilde{T}_v)}{|\tilde{T}_v| - 1} \quad (9)$$

where,  $R(\tilde{T}_v)$  is the sum of the resubstitution misclassification rates of offspring leaf nodes of node  $v$ ; and  $|\tilde{T}_v|$  is the number of offspring leaf nodes of  $v$ .

Choose the internal node with the smallest  $\alpha_v$ , from those derived  $\alpha_v$  values, as the first target node for tree-shrinking. This procedure is repeated and then we can derive an increasing sequence of  $\alpha_v$  values as thresholds

$$0 < \alpha_1 < \alpha_2 < \dots < \alpha_n \quad (10)$$

Accordingly, a decreasing sequence of nested minimising sub-trees can be produced

$$T_{\max} \supset T_{\alpha_1} \supset T_{\alpha_2} \supset \dots \supset T_{\alpha_n} \quad (11)$$

where,  $T_{\alpha_n}$  denotes the root-node sub-tree.

The sub-tree with the smallest test sample or cross-validation estimate of the misclassification rate or cost is ultimately chosen, from the sub-trees, as the optimally sized tree (see also Zhang and Singer, 1999, for more details).<sup>5</sup>

## 5. The Resale Public Housing Market and Data

In Singapore, the housing market may be broadly partitioned into the public and private housing markets. Public housing refers to houses constructed by the government through the statutory board, the Housing Development Board (HDB), while private housing refers to residential properties built by individuals or private developers on either private or state-tendered land. More than 85 per cent of the Singapore population lives in public housing flats—i.e. HDB flats. The overwhelming majority of these housing units are purchased by the residents themselves. Although new HDB flats are supplied by the government for sale with heavy subsidies so that their prices are not directly subjected to the market forces of supply and demand, HDB homebuyers are allowed to

sell their housing flats in the open market after a time-bar of five years. These housing transactions constitute the resale public housing market in Singapore, which is currently one of the most important components of the Singapore residential property market (see, for example, Ong and Sing, 2002; Ong *et al.*, 2003). In 2003, the volume of housing transactions in the resale public housing market amounted to 33 586 in contrast to only 9950 transactions in the private housing market. That is, the Singapore resale public housing market is about three times as large as the private housing market in terms of housing transaction volume. In particular, given the fact that most of the population resides in public housing, resale prices of public housing (i.e. the transaction prices of housing in the resale public housing market) are one of the most important concerns for most households in their housing consumption and investment.

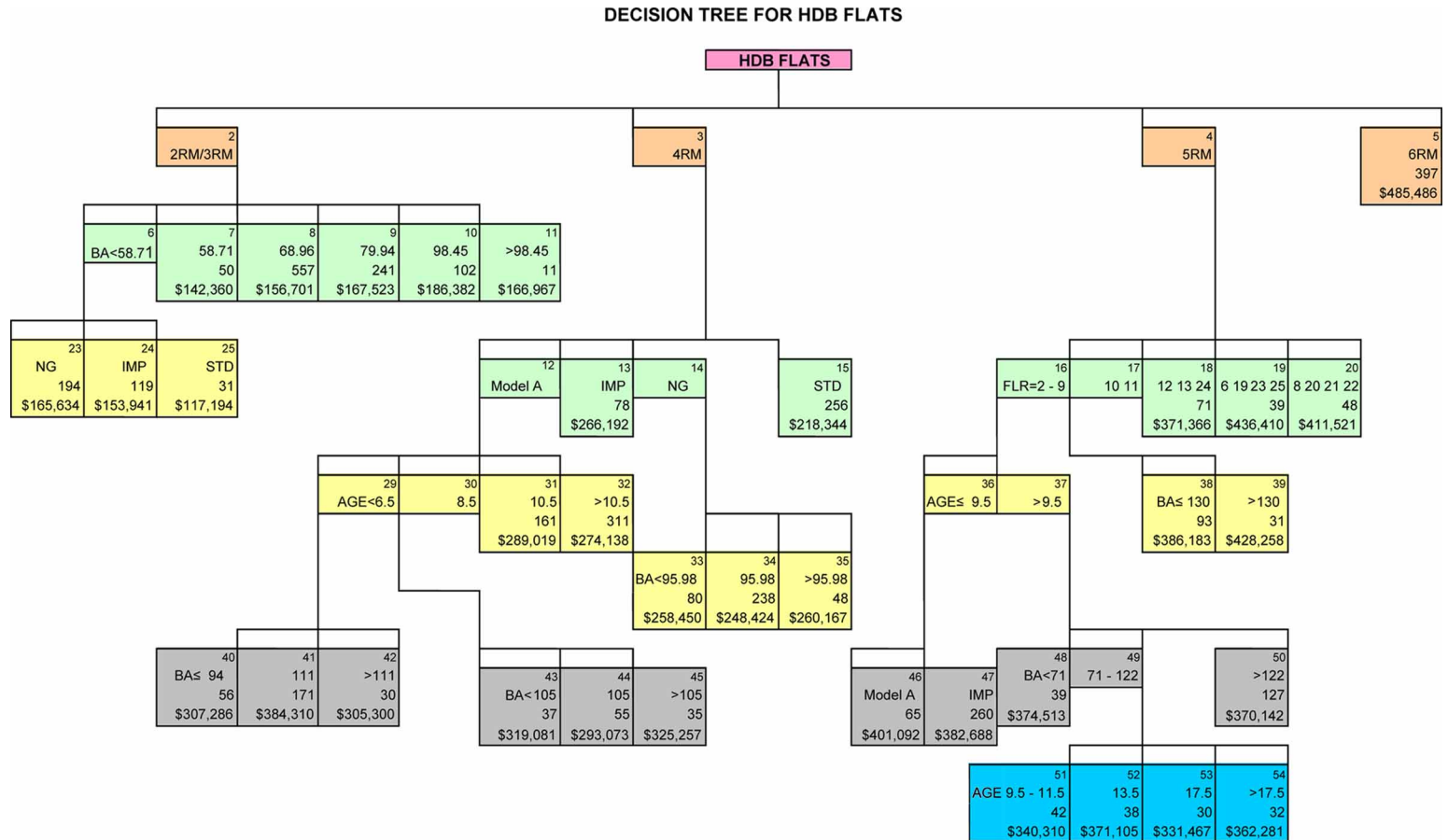
The data used in this study are composed of 5589 resale transactions between January 1997 and December 1998. The sample is obtained from one of the largest real estate agencies in Singapore. The data contain information on the resale price, resale date, floor area, flat type, age of flat, floor level (the floor on which a flat is situated), the number of rooms within a flat, upgrading situation, location and neighborhood. The resale transactions are distributed throughout the various HDB towns of the island country. The average resale price for the 5589 public housing flats in the sample is S\$278 786. The average floor area of the HDB flats is 92 square metres, while their average age is 13 years at the time of sale. The sample contains almost all public housing types, such as two-room, three-room, four-room, five-room and six-room flats (executive apartments), and their sub-types. The average floor level of the HDB flats is 7 floors high and their floor level ranges from 1 (the ground floor) to 25. In addition, the sample contains the information about the status of the HDB Main Upgrading Programme (MUP) of the flats at the time of sale—that is, polling results for their main upgrading, undergoing

the main upgrading, or the completion of their main upgrading.<sup>6</sup> The location and neighborhood characteristics of the flats are measured by their categorised straight-line distances to expressways, mass rapid transit (MRT) stations, bus interchanges, popular primary schools, industrial estates, private housing estates or HDB new towns.

## 6. The Housing Price Tree

The objective of this section is to investigate the relationship between the resale prices of Singapore public housing of all types and housing attributes and identify the significant determinants of the resale prices based on the built housing price tree. To create a housing price tree, we use the Tree Node in SAS (Statistical Analysis Software) Enterprise Miner version 4.1, which integrates the CHAID, CART and C4.5/C5.0 algorithms and extends the algorithms in many aspects.<sup>7</sup> The resale prices of public housing are naturally chosen as the target field or dependent variable, while the housing attributes are treated as independent variables. Prior to building the tree, the Tree Node randomly allocated the sample data into two parts: a training dataset and a test dataset, in the respective proportions of 75 per cent and 25 per cent, in order to help ensure the fit of the constructed tree and its ability to generalise.

Running the Tree Node to approximate the tree-building process described in section 4, we obtained the resulting regression tree displayed in Figure 1 (see also Table 1 for a simplified version of the tree). This tree is composed of 14 internal nodes (also including the root node) and 40 leaf nodes, with non-uniform depth of partitioning. In this tree, each internal node is denoted by a box in which the numbers in the first row represent the sequence of a node and the numbers or characters in the second row are the categories generated by the best splitting criteria. In contrast, each leaf node is represented by a box where the numbers in the first row indicate the sequence of a node, the numbers or characters in the second row are the categories determined by the best splitting criteria, the




**Figure 1.** Decision tree results for the HDB flats




**Table 1.** Summary table of the simplified tree


Initial Split	Two/Three Rooms		Four Rooms						Five Rooms								Executive Apartments	
Second Split	BA<58.71	>58.71	Model A			IMP	NG			STD	FLR=2-9			10 11	12 13 24	6 19 23 25	8 20 21 22	
Third Split			AGE<6.5	8.5	10.5	>10.5		BA<96	96	>96		AGE ≤ 9.5	>9.5	BA≤130	>130			
Fourth Split											Model A	IMP	BA <71	71-122	>122			
Fifth Split																		

 : Floor Area (BA)

 : Model Types (including Model A, New Generation (NG), Improved (IMP) and Standard (STD))

 : Age of HDB Flats(AGE)

 : Floor on which a flat is situated (FLR)

 : No splitting

numbers in the third row are the number of observations allocated into the node and the numbers in the last row are the average resale prices of HDB flats. At the end of the tree-growing process, all 4192 observations in the training set have been assigned to the leaf nodes of the full decision tree.

The average value of the resale prices in a leaf node represents the forecasting value or regression value of the resale price of public housing of a certain category. For example, for node 35, the average price of S\$260 167 indicates the predicted price for the new-generation (NG) type of 4-room flats with more than 95.98 square metres. For node 46, the predicted price of the Model A type of 5-room flat from floors 2–9 and with an age of 9.5 years or less is S\$401 092.

The tree provides a description of how a number of major independent variables differentiate the resale prices of HDB flats within the training set. In Figure 1, the top level or root node of the tree represents the total training set. In the initial split, the Tree Node identifies the *number of rooms* as the best splitter at this node by comparing the classified results produced by all the independent fields (variables), because this variable leads to the largest reduction in node impurity (i.e. node price variance). In other words, this variable can best differentiate between the categories in the target field. Specifically, the training set is partitioned into four subsets—namely, a 2-/3-room subset, 4-room subset, 5-room subset and 6-room (executive apartments) subset (see also the first row in Table 1). In the tree, from the left to the right, the four nodes on the second level represent 2/3 rooms, 4 rooms, 5 rooms and 6 rooms respectively. Such partitioning allows us to compare further the effects of independent variables on the target field among the categories. The splitting rule determining the initial split is commonly viewed as the heart of decision tree algorithms (Berry and Linoff, 2000), which therefore implies that the number of rooms within an HDB unit is the most important determinant of HDB resale housing prices. This is obviously consistent with the transaction practice in the

Singapore resale public housing market. Specifically, an increase in the number of rooms within a flat can significantly raise the construction costs, which actually make up the greatest element in new public housing prices in Singapore, and therefore may further influence the resale prices to a large extent.

After the initial split, the decision tree algorithm further splits the produced four nodes into more nodes using the same partitioning procedure as used at the root node. As to node 2 (the 2- and 3-room node), the *floor area* of these units is identified as the single most significant splitter. Consequently, the dataset of 2- and 3-room units is further partitioned into five leaf nodes and one internal node with three leaf nodes (see also the second column in Table 1). This suggests that floor area can best differentiate between the sub-categories of 2-/3-room units and is the most significant determinant of the resale prices of these units, while model types most significantly affect the resale prices of the HDB units with a floor area of less than 58.71 square metres.

For HDB 4-room flats (node 3), the built tree shows that, of all possible independent variables, *model type* is the single most important factor in predicting the resale prices of the flats. Accordingly, the dataset is partitioned at this node by their model type (see also the third column of Table 1). Also, for new-generation (NG) 4-room units, their floor area is found to be the most significant splitter, while for Model A 4-room flats, the further split is carried out based on their age. Furthermore, floor area is also identified to be the most significant splitter for Model A 4-room flats with an age of less than 6.5 years and with an age of 6.5–8.5 years so that they are further partitioned into six leaf nodes. These imply that model type, floor area and flat age are more important variables in regressing and predicting the resale prices of HDB 4-room units. It is also noticeable that, as the branch of the decision tree grows, the nodes under this branch become purer and more informative about 4-room flats.

Similarly, Figure 1 shows that node 4 (HDB 5-room flats) is further partitioned into 5 internal nodes and 13 leaf nodes (see also the fourth column of Table 1), while for node 5 (executive apartments), no significant splitter can be found (see also the fifth column of Table 1). Such an allocation for node 4 indicates that floor level is identified to be the single most important splitter for 5-room flats and can lead to the greatest reduction in the node price variance of these flats. Further, floor area, flat age and model type are also found to be more important factors in regressing and predicting the resale prices of HDB 5-room flats. However, the decision tree algorithm cannot further distinguish significant splitters of executive apartment resale prices.

Table 1 provides a simplified description of the build tree, which pays attention to the identification of the independent variables better differentiating the categories of HDB flats and, nevertheless, ignores the distribution of the transaction observations in the built tree and the predicted prices of HDB flats. Table 1 demonstrates the identified independent variables and their importance in predicting the resale prices of HDB flats. More specifically, the independent variables that appear higher up in Table 1 may be regarded as more important variables in predicting resale prices. For example, the third column of Table 1 shows that the model type of 4-room flat is identified as the most important explanatory variable, while the next most important explanatory variables are flat age and floor area.

These findings provide several interesting insights into the Singapore resale public housing market. First, the homebuyers are more concerned about the basic housing characteristics of 2- and 3-room flats or 4-room flats such as floor area, model type and flat age. The focus on such basic characteristics is not surprising since public housing usually does not have recreational facilities and living environments comparable with those of private residential properties. Public flats were constructed mainly to alleviate the housing shortage and to satisfy the basic

housing requirements of lower-income and average households in the early years of the public housing programme. As a result, we have reason to believe that the homebuyers pay more attention to basic housing quantitative characteristics in order to satisfy their relatively lower housing requirements due to their limited incomes.

Secondly, homebuyers of 5-room flats pay more attention to floor level in addition to the abovementioned basic housing quantitative characteristics. This is an interesting finding in that to the extent that 5-room flat buyers are usually better off in income terms than buyers of smaller flats, these homebuyers start to have higher dwelling consumption expectations associated with factors related to floor level—such as the view and lucky house numbers (Bourassa and Peng, 1999; Chau, Ma and Ho, 2001)—when their purchasing power improves.

Thirdly, for HDB executive apartments, the result of no splitter being found is also interesting. This indicates that homebuyers of executive apartments are less concerned about basic quantitative characteristics such as floor area, model type, flat age and so on. In Singapore, these apartments were built for those households who possess a total household income exceeding the ceiling of the income qualification for normal HDB flats, but who are priced out of the private housing market (Teo and Kong, 1997). These flats possess better recreational facilities and living environments than common HDB flats and are more similar to private condominiums. As a consequence, we have reason to believe that, with the further rise in purchasing power, the homebuyers have higher housing consumption expectations and pay more attention to ‘quality’ and service characteristics such as recreational facilities and the living environment.

Last but not least, the built decision tree reveals a very interesting order of price-influencing attributes. For instance, *floor area* is a second-order attribute—i.e. second-level split—for 2- and 3-room flats, but is a third-order attribute for 4-room flats other than Model A. In addition, floor area is a

third-order attribute for 5-room flats, but only for flats situated on higher floors. Floor area is a fourth-order attribute for 5-room flats on lower floors (see Figure 1). This result makes intuitive sense in that increases in floor area matter more for smaller flats (a 10 square metre difference for a 70 square metre flat is proportionately higher than a similar difference in floor area for a 120 square metre flat). We can make interesting observations as well for flat age, floor level and model type. The above analysis strongly suggests that the pricing of public flats is highly non-linear, not only in terms of attributes, but also the ordering of these attributes. The implication for parametric pricing models is an interesting area for future research.

Furthermore, we also evaluated the overall performance of the built tree. The relevant estimated results are summarised in Table 2, which reports the source of the data (training set or validation set), the number of observations, the average for the target field, the average squared error and the  $R^2$  for the tree. The results suggest that the built tree performs satisfactorily. For the two datasets, the average values of the target field and the average squared errors are quite close to each other, which implies that the built tree provides good insight into the characteristics of the population. In addition, the  $R^2$  value of about 89 per cent indicates that the constructed price tree has a high model accuracy.<sup>8</sup>

7. Summary and Conclusions

In this study, a decision tree model has been built from the training dataset in order to

help us to examine the relationship between the resale prices of Singapore public houses and housing characteristics and to identify which characteristics are significant in predicting resale prices. Decision tree algorithms perform numerous tests and derive the best sequence for regressing and predicting the dependent variable based on rules expressed in terms of the independent variables. These tests identify the best splitters, which iteratively partition the training data until arriving at terminal (leaf) nodes.

It has been shown that the housing transactions in the Singapore resale public housing market can be visually represented as an interpretable tree-like structure, using the proposed decision tree algorithm, in which each transaction observation is eventually assigned to a leaf node containing the predicted price of public houses in a certain category. Decision trees provide an effective approach to identifying the determinants of public housing resale prices. The built tree shows that, among all the price determinants, the number of rooms, floor area, model type, flat age and floor level are found to be significant in the regression and prediction of the resale prices of HDB flats. In particular, the effects of these variables on resale prices can be analysed further based on the partitioned classes. However, the location and neighborhood characteristics are not identified as the significant splitters in the built tree, because they cannot best differentiate between the housing categories in terms of resale housing prices, given the fact that only a few housing transactions are distributed around a certain amenity or a class of urban amenities over a shorter time-period—for example, two years. For instance, only about 7 per cent of our sample are located near bus interchanges, while about 13 per cent are located near MRT stations. Or, the leaf nodes partitioned by the characteristics probably contain too few transaction cases and are therefore removed in the tree-pruning process. We also demonstrate that the ordering of significant variables/attributes differs for different flat types. Although the results in this study are specific to a housing market segment in

Table 2. Summary table

Statistics	Training set	Validation set
Number of observations	4 192	1 397
Average	279 051.4	277 988.6
Average squared error	1 430 746 262.9	1 498 342 064.9
$R^2$	0.887	0.885

Singapore, it would be equally interesting to apply the same methodology to study other housing markets. In particular, since decision tree models can identify significant household preference orderings for dwelling attributes or services in house-purchase decisions, this therefore provides evidence supporting the application of utility tree theory to housing demand and also suggests a probable approach to looking empirically at utility tree theory.

Additionally, it is quite clear that the decision tree method has a number of noticeable advantages, compared with the hedonic-based regression method, in examining the relationship between housing prices and housing characteristics. Specifically, in decision tree analysis we can clearly identify the most important variables for regression and prediction of the dependent variable at various partitioned levels, almost no strict assumptions with regard to data distribution are required and the built tree is easily interpreted to an end user. Also, this analysis is not undermined by the problems associated with market disequilibrium, the selection of independent variables and the choice of functional form of hedonic equation, and can even identify the non-linear relationship between house prices and housing characteristics. For the above reasons, it is expected that decision trees offer an alternative exploratory data analysis tool for examining the relationship between house prices and a variety of housing characteristics and identifying the significant determinants of housing prices.

Although the decision tree approach can be used as an exploratory data analysis tool to search for patterns, we also note that it suffers from several limitations. First, most decision tree algorithms can only identify the single most significant splitter at a node; although other independent variables may produce a relatively weaker but significant effect on the target variable at the node, their influences cannot be analysed simultaneously within the built tree framework. In other words, it is difficult for decision tree algorithms to carry out a full consideration of the effects of independent variables.

Secondly, it is difficult for decision tree algorithms to analyse and predict the value movement of a continuous variable, while the algorithms are adept at partitioning continuous variables by choosing a number somewhere in their range of values. Therefore, unlike multiple regression models, it is difficult for decision trees to analyse and predict the pure housing price index movement (which is one of major interests to the real estate community), after excluding the effects of observable housing characteristics.

Finally, to utilise decision tree algorithms housing market researchers should acquire knowledge of pattern recognition tools and methodologies. Without sufficient statistical pattern recognition knowledge, it is difficult for researchers to choose appropriate approaches (such as decision trees) and apply them to academic research in the housing area—although this is also equally true for the application of other statistical models.

## Notes

1. Pattern recognition techniques are also often known as data mining techniques in the commercial application.
2. A recent exception is Feldman and Gross (2005), who use a decision tree algorithm in analysing mortgage default.
3. Decision trees are also an important data mining tool.
4. The sample dataset is usually grouped into two subsets: a training dataset used for preliminary tree constructing and a test dataset (also called the validation dataset) used for evaluating the performance of the constructed tree. Breiman *et al.* (1984) proposed to use the resubstitution estimation of the misclassification rate—i.e. estimated from the training set—due to the difficulty of estimating the true misclassification rate.
5. If the misclassification rate of a sub-tree is estimated using the test set, the estimated misclassification rate is known as its test sample estimate. Alternatively, we can use a more complicated method—the cross-validation method—to estimate the misclassification rate. In order to carry out this test, we need randomly partition the cases in a dataset  $\Omega$  into  $M$  almost equal-sized subsets  $\Omega^1, \Omega^2, \dots, \Omega^M$ . For every  $m, m = 1, 2, \dots, M$ ,

use  $\Omega - \Omega^m$  as the training set and  $\Omega^m$  as the test set. Then the cross-validation estimation can be derived by averaging the test sample estimates over  $M$  (see Breiman *et al.*, 1984, for more details).

6. The MUP is part of the government's efforts to enhance public housing estates by providing substantial subsidies for improvement works on the selected old housing estates, such as upgrading of services, landscaping, facade enhancement and other external works (Ong *et al.*, 2003).
7. Although the Tree Node integrates the CHAID, CART and C4.5/C5.0 algorithms in order better to satisfy the tree-building needs of different situations, we can also use it to approximate the algorithms by choosing the modelling features used in the algorithms. The Tree Node also extends these algorithms in many aspects. For example, it allows the CART algorithm to use more model assessment measures for selecting the best tree and it extends the CHAID algorithm with multiway splits on interval variables.
8. In multiple regression models,  $R^2$  is frequently utilised to measure the proportion of variation of the dependent variables interpreted by the specified models. In the decision tree context, however,  $R^2$  is a concept related to the measure of model accuracy.

Let  $\bar{Y}$  and  $\hat{Y}_i$  represent the mean and the predicted values of the dependent variable  $Y_i$ , respectively. The relative error, measuring model accuracy, is defined as

$$RE = \sum_{i=1}^N (Y_i - \hat{Y}_i)^2 / \sum_{i=1}^N (Y_i - \bar{Y})^2$$

where,  $N$  is the sample size. Then,  $R^2$  can be specified as  $R^2 = 1 - RE$  (see Breiman *et al.*, 1984).

## References

- APPS, P. F. (1973) An approach to urban modeling and evaluation. A residential model: 1. theory, *Environment and Planning A*, 5, pp. 619–632.
- BEN-BASSAT, M. (1987) Use of distance measures, information measures and error bounds on feature evaluation, in: P. R. KRISHNAIAH and L. N. KANAL (Eds) *Handbook of Statistics, Vol. 2. Classification, Pattern Recognition and Reduction of Dimensionality*, pp. 773–791. Amsterdam: North-Holland.
- BENSON, E. D., HANSEN, J. L., SCHWARTZ, A. L. and SMERSH, G. T. (1998) Pricing residential amenities: the value of a view, *Journal of Real Estate Finance and Economics*, 16(1), pp. 55–73.
- BERRY, J., MCGREAL, S., STEVENSON, S. ET AL. (2003) Estimation of apartment submarkets in Dublin, Ireland, *Journal of Real Estate Research*, 25(2), pp. 159–170.
- BERRY, M. J. A. and LINOFF, G. (1997) *Data Mining Techniques: For Marketing, Sales, and Customer Support*. New York: John Wiley & Sons.
- BERRY, M. J. A. and LINOFF, G. (2000) *Mastering Data Mining: The Art and Science of Customer Relationship Management*. New York: Wiley Computer Publishing.
- BOURASSA, S. C. and PENG, V. S. (1999) Hedonic prices and house numbers: the influence of feng shui, *International Real Estate Review*, 2(1), pp. 79–93.
- BREIMAN, L., FREIDMAN, J. H., OLSHEN, R. A. and STONE, C. J. (1984) *Classification and Regression Trees*. California: Wadsworth.
- CAN, A. (1992) Specification and estimation of hedonic housing price models, *Regional Science and Urban Economics*, 22, pp. 453–474.
- CHAU, K. W., MA, V. S. M. and HO, D. C. W. (2001) The pricing of 'luckiness' in the apartment market, *Journal of Real Estate Literature*, 9(1), pp. 31–40.
- CHAU, K. W., NG, F. F. and HUNG, E. C. T. (2001) Developer's good will as significant influence on apartment unit prices, *Appraisal Journal*, 69, pp. 26–34.
- CRAWFORD, S. L. (1989) Extensions to the CART algorithm, *International Journal of Man-Machine Studies*, 31(2), pp. 197–217.
- DO, Q. and GRUDNITSKI, G. (1992) A neural network approach to residential property appraisal, *The Real Estate Appraiser*, 58, pp. 38–45.
- FELDMAN, D. and GROSS, S. (2005) Mortgage default: classification trees analysis, *Journal of Real Estate Finance and Economics*, 30(4), pp. 369–396.
- FLETCHER, M., GALLIMORE, P. and MANGAN, J. (2000a) Heteroskedasticity in hedonic house price models, *Journal of Property Research*, 17(2), pp. 93–108.
- FLETCHER, M., GALLIMORE, P. and MANGAN, J. (2000b) The modeling of housing submarkets, *Journal of Property Investment and Finance*, 18(4), pp. 473–487.
- FRYDMAN, H., ALTMAN, E. I. and KAO, D. L. (1985) Introducing recursive partitioning for financial classification: the case of financial distress, *Journal of Finance*, 40(1), pp. 269–291.
- GALLAGHER, M. and MANSOUR, A. (2000) An analysis of hotel real estate market dynamics, *Journal of Real Estate Research*, 19(1/2), pp. 133–164.
- GOETZMANN, W. N. and WACHTER, S. M. (1995) Clustering methods for real estate portfolios, *Real Estate Economics*, 23(3), pp. 271–310.

- GOODMAN, A. C. (1978) Hedonic prices, price indices, and housing markets, *Journal of Urban Economics*, 5(4), pp. 471–484.
- GOODMAN, R. M. and SMYTH, P. J. (1988) Decision tree design from a communication theory standpoint, *IEEE Transactions on Information Theory*, 34(5), pp. 979–994.
- JAIN, A. K., DUIN, R. P. W. and MAO, J. (2000) Statistical pattern recognition: a review, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1), pp. 4–37.
- KAO, D. L. and SHUMAKER, R. D. (1999) Equity style timing, *Financial Analysts Journal*, (January/February), pp. 37–48.
- LI, M. M. and BROWN, H. J. (1980) Micro-neighborhood externalities and hedonic housing prices, *Land Economics*, 56(2), pp. 125–141.
- MALPEZZI, S. (2003) Hedonic pricing models: a selective and applied review, in: T. O'SULLIVAN and K. GIBB (Eds) *Housing Economics and Public Policy*, pp. 67–89. Malden, MA: Blackwell Science.
- MCCLUSKEY, W. and ANAND, S. (1999) The application of intelligent hybrid techniques for the mass appraisal of residential properties, *Journal of Property Investment & Finance*, 17(3), pp. 218–238.
- MICHIE, D. (1986) The superarticulatory phenomenon in the context of software manufacture, *Process of the Royal Society of London*, 405A, pp. 185–212.
- MURTHY, S. K. (1998) Automatic construction of decision trees from data: a multi-disciplinary survey, *Data Mining and Knowledge Discovery*, 2, pp. 345–389.
- MURTHY, S. K., KASIF, S., SALZBERG, S. and BEIGEL, R. (1993) OC1: randomized induction of oblique decision trees, in: *Proceedings of the Eleventh National Conference on Artificial Intelligence*, pp. 322–327. Washington, DC: AAAI Press/The MIT Press.
- NGUYEN, N. and CRIPPS, A. (2001) Predicting housing value: a comparison of multiple regression analysis and artificial neural networks, *Journal of Real Estate Research*, 22(3), pp. 313–336.
- ONG, S. E. and SING, T. F. (2002) Price discovery between private and public housing markets, *Urban Studies*, 39(1), pp. 57–67.
- ONG, S. E., HO, K. H. D. and LIM, C. H. (2003) A constant-quality price index for resale public housing flats in Singapore, *Urban Studies*, 40(13), pp. 2705–2729.
- QUINLAN, J. R. (1986) Induction of decision trees, *Machine Learning*, 1, pp. 81–106.
- QUINLAN, J. R. (1987) Simplifying decision trees, *International Journal of Man–Machine Studies*, 27, pp. 221–234.
- QUINLAN, J. R. (1993) *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann.
- RODRIGUEZ, M. and SIRMANS, C. F. (1994) Quantifying the value of a view in single-family housing markets, *Appraisal Journal*, 62, pp. 600–603.
- SAFAVIN, S. R. and LANDGREBE, D. (1991) A survey of decision tree classifier methodology, *IEEE Transactions on Systems, Man and Cybernetics*, 21(3), pp. 660–674.
- SAS INSTITUTE (1998) *From data to business advantages: data mining*. The SEMMA Methodology and SAS Software, SAS Institute, North Carolina.
- SHEPPARD, S. (1999) Hedonic analysis of housing markets, in: P. C. CHESHIRE and E. S. MILLS (Eds) *Handbook of Regional and Urban Economics*, Vol. 3, ch. 41. Amsterdam: Elsevier.
- SO, H. M., TSE, R. Y. C. and GANESAN, S. (1997) Estimating the influence of transport on house prices: evidence from Hong Kong, *Journal of Property Valuation and Investment*, 15(1), pp. 40–47.
- SORENSEN, E. H., MILLER, K. L. and OOI, C. K. (2000) The decision tree approach to stock selection, *Journal of Portfolio Management*, Fall, pp. 42–51.
- STROTZ, R. H. (1957) The empirical implications of a utility tree, *Econometrica*, 25, pp. 269–280.
- STROTZ, R. H. (1959) The utility tree: a correction and further appraisal, *Econometrica*, 27, pp. 482–488.
- TAY, D. P. H. and HO, D. K. K. (1991/92) Artificial intelligence and the mass appraisal of residential apartments, *Journal of Property Valuation and Investment*, 10, pp. 525–540.
- TEO, S. E. and KONG, L. (1997) Public housing in Singapore: interpreting 'quality' in the 1990s, *Urban Studies*, 34, pp. 441–452.
- WILKINSON, R. K. (1973) House prices and the measurement of externalities, *The Economic Journal*, 83(329), pp. 72–86.
- WOODS, E. and KYRAL, E. (1997) *Ovum evaluates: data mining*. Ovum Ltd, London.
- WORZALA, E. M., LENK, M. M. and SILVA, A. (1995) An exploration of neural networks and its application to real estate valuation, *Journal of Real Estate Research*, 10(2), pp. 185–202.
- ZHANG, H. P. and SINGER, B. (1999) *Recursive Partitioning in the Health Sciences*. New York: Springer.