Thill, Jean-Claude; Wheeler, Aaron

Conference Paper

Tree Induction of Spatial Choice Behavior

39th Congress of the European Regional Science Association: "Regional Cohesion and Competitiveness in 21st Century Europe", August 23 - 27, 1999, Dublin, Ireland

Provided in Cooperation with:
European Regional Science Association (ERSA)

Tree Induction of Spatial Choice Behavior

Jean-Claude Thill
Department of Geography
State University of New York at Buffalo
Amherst, New York 14261, USA

Aaron Wheeler
Department of Computer Science
University of New Mexico
Albuquerque, New Mexico 87131, USA

Acknowledgement. The first author wishes to acknowledge partial support of the National Science Foundation to the National Center for Geographic Information and Analysis at SUNY/Buffalo under award SBR-9600465.

1. Introduction

Machine learning, a branch of artificial intelligence, investigates the mechanisms by which knowledge is acquired through experience. A large number of machine learning methods and algorithms have been developed, including neural computing (Freeman and Skapura, 1991), case-based reasoning (Kolodner, 1993), genetic algorithms (Goldberg, 1989), and inductive learning (Quinlan, 1988). These approaches form the essential toolbox of methods to extract useful information from data sets built into the knowledge base of expert systems. It has been argued that these computational methods are not only useful for the design and implementation of effective and efficient decision support and expert systems, but also as support tools in furthering scientific knowledge discovery above and beyond what conventional methods of inquiry have so far permitted. In the domain of the Spatial Sciences, this viewpoint is forcefully advocated in the research white paper on "Spatial Analysis in a GIS Environment" of the University Consortium for Geographic Information Science (UCGIS, 1997).

In this chapter, we discuss the merit of inductive learning as an analysis tool in spatial decision making theory. We analyze the capability and applicability of Ross Quinlan's (1993) C4.5 decision tree induction algorithm to the class of problems involving the choice among travel destination within an urban area. The chapter reviews the relevant destination choice modeling literature, describes the C4.5 algorithm and its relation to other decision tree induction algorithms, and illustrates its implementation on spatial behavior data from the Minneapolis-St.Paul, MN, metropolitan area.

The chapter is organized as follows. Section 2 gives a brief overview of the analytical theory of destination choice behavior. Section 3 presents the main characteristics of the C4.5 algorithm and discusses its suitability for spatial modeling in general, and destination choice modeling in particular. Section 4 describes the choice problem that is used to illustrate the algorithm's capabilities. Results are presented in Section 5. The final section contains some conclusions.

2. Spatial Behavior Modeling

a. Heuristics and Decision Tree

The dominant paradigm of spatial choice theory is that of a two-stage decision process whereby a choice set is delineated first, and one option from the choice set is subsequently selected in accordance with some utility-based compensatory model (Timmermans and Golledge, 1990; Thill and Timmermans, 1992). The decision-making process is modeled by allowing full or partial compensation of a low score on some attribute of the choices alternatives by high scores on one or more of the remaining attributes.

Over the years, another stream of research has also been pursued on the premise that individuals adopt non-compensatory decision strategies (for instance, Recker and Golob, 1979; Timmermans, 1983; Johnson and Meyer, 1984). Justification for this alternative modeling approach can be derived from the large size of choice sets in many spatial choice situations, but also from the uneven and patchy knowledge that decision makers have of options in the universal choice set. Evidence from cognitive psychology experiments suggests that people have limited capacity to process information (Bettman, 1979) and use heuristics to cope with the complexity of choice sets and reduce the time and cost of decision-making tasks. These "short-cuts" and approximate rules guide decision making, but without guaranteeing that they will lead to the best solution (Svenson, 1979; Eagle, 1980; Timmermans, 1984). Decision heuristics are embedded in computational process models wherein knowledge structure consists of declarative knowledge (factual statements) and procedural knowledge (instructions, explanations, and logical statements). State-of-the-art reviews of computational process modeling of travel behavior are available in Gärling, Kwan, and Golledge (1994) and Kwan and Golledge (1997).

In their path-breaking work, Howard (1963), Nicosia (1966), and Howard and Seth (1969) conceptualized the decision process as a narrowing of alternatives or "funneling process" leading to a final decision. Variants of these early views have been elaborated upon by scholars across various disciplines (Manski, 1977; Fotheringham, 1988; Crompton, 1992; Thill, 1992). Experiments by Newell and Simon (1972) show that problem solving with heuristics can be represented by a computational scheme called a production system, which uses rules as only algorithmic elements. See also Davis and King (1976). Production rules are generally of the following form: IF <condition> THEN <action>, and can contain either compensatory or non-compensatory knowledge. The condition part of a rule is a concatenation of elemental terms

created with the preferences <, >, =, and the logic operators AND, OR, and NOT. Newell and Simon base their conclusions about the methods and organization of human decision-making on studies of verbal protocol of subjects engaged in problem-solving in several domains, including cryptarithmetic, logic, and chess. Their analyses, together with computer simulations, provide strong evidence that people use heuristics to solve complex problems and that they organize their problem solving strategies in a way consistent with production systems. Smith and Lundberg (1984) discuss how heuristic problem solving requires knowledge of the current state of the system and of an appropriate action to take given the observed conditions. In their work, they show that the former is equivalent to the condition part of a production rule while the latter is equivalent to the action part of the production rule.

A production system can be expressed as a decision tree where the nodes of the tree are tests on some attribute or function. A path from the root node of a decision tree to a terminal leaf is equivalent to a production rule (Quinlan, 1990). Each node in a decision tree performs a context-sensitive test on an attribute. Tests performed earlier in the path define the context in which a subsequent test is appropriate (Quinlan, 1993). Different variables and models are important in different contexts for making proper spatial decisions. The decision tree representation of the spatial choice process provides a context-sensitive evaluation of behavioral primitives and controls from variables and models.

b. Spatial Decision Trees

In addition to the studies of spatial decision processes mentioned in the previous section, a handful of contributions are noteworthy. Several studies have demonstrated the suitability of production systems and decision trees for representing the consumer search process in housing markets (Smith et al., 1982; Smith and Lundberg, 1984; Smith et al., 1984; Clark and Smith, 1985). The rule induction algorithm adds rules incrementally so the final production system represents the minimum number of rules necessary for a given level of predictive accuracy. At each iteration the algorithm adds the rule that maximizes a given criterion function. They derive production rules that are both predictive of the final result and of the actual sequence of decisions leading to the final choice.

The perceptual space framed by commodity attributes and the geographic space of spatial scientists are perfectly isomorphic. Various marketing models have a tree-like hierarchical structure, including Tversky's famous elimination-by-aspects and elimination-by-tree models (Tversky, 1972; Tversky and Sattath, 1979). In their study of the dynamics of brand switching between soft drinks, Moore et al. (1986) represent consumer decision-making using a tree structure. Preference trees are parameterized by means of hierarchical clustering algorithms, but the authors add the caveat that these trees may not reflect the actual ordering of preferences in the consumer choice process. The form of the tree must be based either on prior theory or additional analytical models.

Decision tables are very similar to decision trees, except for their tabular form. The upper portion of a decision table contains the conditions while the bottom portion contains the actions. Each row in the condition part of the table corresponds to a different variable (decision criterion) and columns in that row correspond to values or ranges of values for that variable. Each column, read from top to bottom, is a production rule. Arentze et al. (1995) describe an integrated expert and decision support system (DSS) for facility location in which expert knowledge is organized in a decision table. The DSS uses the Advanced Knowledge Transfer System (AKTS) to acquire decision tables containing expert rules.

Researchers have found artificial intelligence techniques useful for generating consumer choice rules with the form of production rules. Greene and Smith (1987) use genetic algorithms (Goldberg, 1989) to derive above average systems of production rules describing consumer choices based on a set of attributes of a hypothetical product. They compare their results to a logit model and to the Concept Learning System (CLS), which is a predecessor to the decision tree induction algorithm described in the next section. While the genetic algorithm approach is found to perform comparably to the logit model, both perform better than the Concept Learning System. The CLS is also used by Currim et al. (1988) to derive consumer choice strategies for selecting between coffee brands. These authors compare the decision tree representation to a traditional logit model and conclude that the former is superior in cases of non-compensatory decision-making. Oliver (1993, 1994) employs a genetic algorithm based system to extract decision rules from a dataset of artificial choices of a carpet cleaner. Oliver finds the rules to be

accurate predictors of choices, but not necessarily indicative of the process people would use to make similar decisions.

Decision trees usually come from automated induction algorithms applied to data sets of many decision-making events. On the contrary, the tree structure of decision nets is obtained directly from verbal protocol collected through personal interviews. Verbal protocol consists of a detailed description of every step of the choice process as described by each person interviewed. A weakness of this approach is that participants may not be fully cognizant of how they make decisions or they may not be capable of clearly articulating their reasoning to the interviewer. Timmermans and van der Heijden (1987) applied decision nets to the study of recreational choice behavior in the Netherlands. More recently, van Zwetselaar and Goetgeluk (1994) describe how to use decision nets to model consumer decision-making in house purchasing. Once generated, decision nets are processed using rules of logic to remove any inconsistencies. Oskamp (1994) develops a modeling environment called LocSim and built around decision nets to simulate individual behavior in dynamic housing markets. In this environment, consumers use decision nets to select houses according to a variety of attributes, including price and relative location. Witlox (1995) discusses research using both decision trees and decision nets, mostly in the context of housing choice.

3 Decision Tree Induction

The practical use of hierarchical and tree-structured models of choice has severely been hampered by the limitations of many methods devised to establish the tree structure that is appropriate to the choice situation. Hierarchical clustering methods (for instance, Rao and Sabavala, 1981; Moore et al., 1986), linear models (for instance, Batsell and Polking, 1985; Meyer and Eagle, 1982), and many other approaches (for instance, Gensch, 1987) commonly require that the tree structure be pre-determined or that aggregate data be used. Alternatively, machine learning algorithms are ideally suited to find the most parsimonious tree representation of the data with little on no restrictions imposed on tree structure or nature of the data. Tree induction algorithms are nonparametric classification procedures that try to discriminate the population of cases presented to it by conditions into meaningful groups (leaves). The inferred "if-then" rules relate a set of

predictor variables (attributes of alternatives, characteristics of decision makers, descriptors of spatial structure) to a discrete outcome criterion or dependent variable (the stated or revealed choice). The choice between more than two discrete alternatives can be operationalized in various ways. One approach is to represent choice by a polychotomous nominal variable. Alternatively, a set of one binary variable less than the number of choice options captures equally well the choice criterion.

Contrary to econometric approaches to spatial choice such as logit modeling, tree induction algorithms are nonparametric methods that do not require specification of a functional form, thus permitting a great variety of compensatory or noncompensatory to be revealed by the data with little interference with the analyst's a priori judgement. They may serve to calibrate or train a hypothesized choice model on sampled observations, but also to forecast spatial choices out of behavioral heuristics extracted from the training data. Tree induction algorithms are computer-intensive procedures. Their use was until recently restricted by the processing capability of computers available to most researchers. This barrier has since dissipated thanks to the tremendous leap in computer technology of the past few years.

In this framework, classical tree induction algorithms include Concept Learning System (CLS) (Hunt et al., 1966), AID (Morgan and Sonquist, 1963), CART (Breiman et al., 1984), and CHAID (Perreault and Barksdale, 1980). Subsequently, Quinlan (1979) developed a variant of the original CLS algorithm, called ID3, which became part of the C4.5 family of procedures (Quinlan, 1993).

The system used to produce spatial decision trees in this chapter is a top-down, divide-and-conquer decision tree induction strategy based on the concept of information gain. The particular method is a variation of Quinlan's C4.5 decision tree induction programs. The procedure aims to discriminate chosen and unchosen alternatives with a parsimonious tree. It is said to be top-down because all observations in the training set are members of the root node and the tree is gradually built by addition of decision nodes. The divide-and-conquer strategy classifies the observations at each node according to the value of some attribute. The procedure always terminates because each partition contains fewer observations than the node in question. If a partition contains exactly the same number of observations as the node then the algorithm tests

another attribute until it finds a suitable attribute or until there are no more attributes to support further partitioning (Quinlan, 1993). The decision tree induction program used in the present research differs from C4.5 in that it contains only the tree induction algorithm and a pruning algorithm, rather than the full suite of rule generating programs in C4.5. Furthermore, the pruning algorithm in the present research differs from that used in C4.5. These differences will be discussed in more detail later in this section.

Branching and classification of observations are controlled by information criteria. Let us first define the information content (entropy) of a set S of observations to be the average number of bits necessary to correctly classify each of its elements into k classes C_j. In Quinlan's (1993) notation, information content is

$$Info(S) = -\sum_{j=1}^{k} \frac{|C_j|}{|S|} \log_2 \left(\frac{|C_j|}{|S|} \right) \quad (1)$$

where $|C_j|$ is the cardinality of class C_j in S and |S| is the cardinality of S.

Each node in a decision tree applies a test on some attribute to the observations associated with that node. The principle of the algorithm is to select the test and attribute that "best" minimizes the information necessary to correctly classify the observations. The criterion that is maximized at each node T is the difference between the entropy of the node and the entropy after partitioning the node according to the value(s) of a given attribute X, also known as the information gain. Mathematically, the goal is to maximize

$$Gain(T) = Info(T) - Info_X(T) \quad (2)$$

One significant flaw of the information gain criterion given above is its bias toward tests with many partitions. For instance, a test that partitions N observations into N singleton categories maximizes information gain but is worthless because it generates a trivial classification of observations. This bias can be corrected by emphasizing the quality of the information contained in each particular classification scheme. For this purpose, let us define the split information of a

set S of n partitions as the potential information gained by splitting T into n partitions in absence of other prior information:

$$SplitInfo(S) = -\sum_{i=1}^{n} \frac{|T_i|}{|T|} \log_2 \left(\frac{|T_i|}{|T|} \right) \quad (3)$$

The algorithm uses split information in conjunction with information gain to provide a measure of the proportion of useful information generated by the partitions. The redefined criterion is the gain ratio given by equation (4):

$$GainRatio(X) = \frac{Gain(X)}{SplitInfo(X)} \quad (4)$$

Thus, the goal of the decision tree induction method at each node is to select the attribute that maximizes the gain ratio with an added constraint that the information gain also be at least average with respect to all attributes tested (Quinlan, 1993).

The procedure outlined above for tree induction is applicable to instances of attributes with discrete values (ordinal or cardinal scales) but also to instances of continuous attributes. The difficulty of finding appropriate partitioning thresholds on continuous attributes can be overcome by following Breiman et al.'s (1984) strategy. This strategy goes as follows. Since a set of m observations v_i takes a finite number of values of a certain continuous attribute, the set can be divided into at most m-1 different partitions to create the two subsets $\{v_1,\ldots, v_i\}$ and $\{v_{i+1},\ldots, v_m\}$. It is a simple matter then to test all m-1 potential thresholds and to select the one that maximizes the gain ratio (Quinlan, 1993).

When a decision tree is used to classify a set of observations, each observation follows a path from node to node, beginning with the root node and terminating in a leaf node. Each leaf node predicts the class of the observations that it receives. The value that each leaf node assigns to all observations it receives is the most frequent class of the observations assigned to that leaf during training. Let us illustrate this point with the hypothetical case of twenty observations reaching a particular leaf node. Of these observations, fifteen have class A and five have class B. Therefore, any observation that arrives at this node is predicted to have a class of A. The

confidence level (CL) of this prediction is the ratio of correctly classified observations to all observations in the node. In this example, the confidence level of the prediction of the leaf node on the training data is CL = 15/20 = 0.75.

The purpose of inductive learning methods such as the one for decision tree induction described above is to extract classification rules from data presented to them. However, these methods may extract rules that are useful only for the data set used to train them. In such instances, the induction procedure finds rules where it should only find noise; the system is over-trained. It is standard procedure to resort to a pruning algorithm to prevent the induction algorithm from creating decision trees that perform well on training data but poorly on unseen test data. See Kim and Koehler (1995) for theoretical and practical issues on tree pruning. We use here the pessimistic pruning algorithm proposed by Quinlan (1987). This algorithm is preferred to the error-based pruning algorithm available in C4.5 because it is faster and performs equally well to the error-based pruning (Esposito et al., 1997).

The pessimistic pruning algorithm takes as input a complete, unpruned decision tree. Beginning with the root node, the algorithm examines the subtree on each branch of the current node in turn. The algorithm calculates the number of errors in the subtree assuming that each terminal leaf classifies an observation according to the most frequently found class in that leaf. If the following inequality holds then the pruning rule replaces the subtree with a leaf corresponding to the most frequent class in the subtree. The inequality to test is

$$NodeErrors + \tfrac{1}{2} < SubTreeErrors + LeavesInSubTree/2 + StdError \quad (5)$$

where the standard error is given by

$$StdError = 2\sqrt{\frac{(SubTreeErrors + \tfrac{1}{2})(ObservationsInSubTree - SubTreeErrors)}{ObservationsInSubTree}} \quad (6)$$

and contains the (dis)continuity correction for the binomial distribution. If the inequality does not hold then the algorithm proceeds deeper into the tree until all the nodes have been examined or pruned. This method produces very good results in terms of simplifying decision trees and is also very fast.

Pessimistic pruning is used to insure that the expected confidence levels obtained for predictions on the training data are similar to actual confidence levels obtained from unseen data. Expected and actual confidence levels must be quite similar for the decision trees to be meaningful to this research. Actual confidence levels that are much worse than those predicted suggest that a better decision tree can be discovered with more or different observations, while similar expected and actual confidence levels suggests that the decision tree does in fact represent the optimal choice strategy.

4. Test Problem

a. Data

The data used as a test problem are obtained from the 1990 Minneapolis-St. Paul, MN, Travel Behavior Inventory conducted by the Minneapolis-St.Paul Metropolitan Council (Metropolitan Council, 1990). The home interview survey compiles travel activities of all participants during a 24-hour period. Detailed information on more than 100,000 trips over one block in length made by all members aged five and over in 9746 randomly selected households constitute the full data set. All participants live in the metropolitan area.

The trips considered in this study have the following characteristics:

- They are home-based;
- They are not part of a multi-stop tour;
- Their purpose is shopping (no distinction is made on the basis of the type of goods purchased on the trip);
- The trip destination is located within the metropolitan area;
- They are made by car.

A total of 667 trips meeting these conditions are extracted from the entire database and use for training the tree induction algorithm.

The origin and destination of each trip are geo-referenced by the traffic analysis zones (TAZ) in which they are located. The Minneapolis-St. Paul metropolitan area is composed of 1165 internal traffic analysis zones. All 1165 TAZs form the universal choice set for the shopping destination choice problem considered here

A total of 19 independent variables are included in the choice model. See Table 1 for a summary list of variables. Three sets of variables are used to predict the choice of a shopping destination: spatial separation between the trip origin and the potential destinations, characteristics of the potential destinations, and attributes of the individual. Variables are briefly described hereunder.

[Insert Table 1 about here]

Two related measures of spatial separation are used: the shortest distance (DISTANCE) measured on the highway network built by the Minneapolis Department of Transportation, and travel time (TIME). The TIME variable is calculated form the network distance and a speed imputed to highway link as a function of their functional type (e.g., freeway, ramp, etc.) and the area type, or geographic setting within the metropolitan area (e.g., central city, rural). These variables are generated in a geographic information system.

Destination characteristics include TAZ population counts in 1990 (POP90), TAZ employment in retail businesses (RET_EM) and in personal services (PERSERV_EM) and the presence/absence of a regional shopping mall (MALL). The form of urbanization in destination TAZs is represented by three dummy variables: developed areas (AREA_TYPED1), central city/CBDs (AREA_TYPED2) and outlying business districts (AREA_TYPED3). Rural and developing areas constitute the reference group. The same classification of trip origins is also used: developed areas (AREA_TYPEO1), central city/CBDs (AREA_TYPEO2) and outlying business districts (AREA_TYPEO3). Once gain, rural and developing areas constitute the reference group. PCOMLU is the percentage of the land area of a TAZ destination that has a commercial or service land use.

Several socio-demographic characteristics of decision makers are tested in the model. They include: the age of the individual (AGE), the gender (GENDER), the household size (HHLDSIZE), annual household income (0 for income under $35,000, 1 for income over this level) (INCOME), the number of children under the age of 5 (INFANTS), and the number of cars owned by household members (CARS).

Table 2 presents descriptive statistics on the independent variables in the model.

[Insert Table 2 about here]

5. Tree Induction Results

All 667 individuals of the shopping trip sample have a universal choice set of 1165 options. Of the 777,055 possible travel instances (1165 x 667), 67,367 are selected for training the decision tree. The training set included all 667 chosen TAZs as well as 100 destination zones selected randomly among each respondent's set of unchosen zones.

The induction code is written in C++. Training is completed in 95 minutes on a 200 MHz Wintel processor with 96 MB RAM. The pruning algorithm, also written in C++, takes 5 seconds on the same machine, including input and output. The unpruned tree comprises 1,277 nodes. After pruning, the tree is reduced to 359 nodes. We also prevent branching at nodes encompassing less than 20 instances to preserve the inferential properties of the induction tree. This post-processing shrinks the tree to 327 nodes. The first five depth levels of this tree are charted in Figure 1. The complete tree is presented in the Appendix of the chapter.

[Insert Figure 1 about here]

All 19 predictors appear in at least one production rule featured in the full decision tree. The most discriminating variables are attributes of the destinations (RET_EM, PERSERV_EM, PCOMLU, POP90, MALL, AREA_TYPED3) and measures of spatial impedance between origin and destination TAZs (TIME, DISTANCE). Respondent characteristics become significant predictors further down the tree (see Appendix). AREA_TYPEO1 is the most discriminating of all personal characteristics: it anchors the branching test at node 24 (Figure 1). These results are consistent with the conclusions of the Approximate Nested-Choice Set Destination Choice (ANCS-DC) Model --a model of constrained discrete choice-- estimated on the same data (Thill and Horowitz, 1997) and the extensive literature on shopping destination theory.

The contingency table of observed versus predicted choices is given in Table 3. The χ^2 statistic associated with this matrix is 40,821.03, a value considerably larger than the theoretical value with two degrees of freedom at $\alpha = 0.001$. Not only the trained decision tree model is statistically significant, but also it captures the essence of behavioral heuristics from the working data set with great accuracy. The model compares very favorably with the ANCS-DC model ($\chi^2 = 35.27\%$), and other conventional logit models ($\chi^2 = 34.76\%$) (Thill and Horowitz, 1997).

[Insert Table 3 about here]

The mean square error (MSE) of the trained data, calculated as 1- percent_right / 100, equals 0.8%. Closer examination reveals that 99.9% of the 66700 instances not chosen by respondents are predicted correctly, against only 25.6% of the 667 chosen instances. The lower prediction of chosen instances should not be a surprise given that the predicted value that each leaf node assigns to all observations it receives is the most frequent choice class of observations assigned to that node through training, and that chosen instances for a mere eleventh of the entire training set. In this respect, the interpretation of these statistics should be sensitive to the fact that the percent_right and MSE statistics are dependent on the size of training set. For comparison purposes, we induced a decision tree on a training set of 7,337 instances, formed of 667 chosen instances and 10 randomly selected instances for each respondents. The pessimistic pruning algorithms was applied and nodes with fewer than 10 observations were not split. The MSE associated to this tree is 2.5%; unchosen alternatives have a correct prediction rate of 99.5% while 77.2% of chosen alternatives are correctly predicted. The jump in prediction rate of chosen alternatives from 25.6% to 77.2% is no less than an artifact of the smaller size of the choice set of each respondent (11 versus 111). Consequently, the tree may have learned much of the "noise" in the data presented to it at the expense of the extraction of general heuristics of spatial decision and choice. If overfitting is present, the validity of the production system on unseen data will be downgraded and its predictive power will be seriously compromised. This reinforces the need for validation on unseen data before the results of computational procedures such as Quinlan's tree induction algorithm can be given reliable and robust interpretation.

6. Conclusions

This chapter discussed the merit of inductive learning as a set of procedures to discover knowledge in large and complex databases, such as those typically available in the context of spatial choice behavior. Inductive learning shares with other computational approaches of artificial intelligence the remarkable property of a very lean body of assumptions to enable knowledge discovery. Distributional properties are not imposed, nor is the joke of functional representations. Inductive learning does not require that specific decision structure be pre-determined. It allows for the induction of production systems, or sets of decision heuristics, of all levels of complexity: from logical statements built from factual conditions, to statements using modular components akin to processors of information (models) subsequently incorporated in the evaluation of alternative choice options.

The test problem of shopping destination choice in Minneapolis-St. Paul served to illustrate the implementation of Quinlan's C4.5 algorithm and of a pessimistic pruning algorithm on the discovery of heuristics in spatial decision making. The trained model performed satisfactorily and compared very favorably to more conventional discrete choice modeling efforts on the same data. The discussion stressed the need to validate the tree model on unseen data to prevent that noise in the data be reproduced by the knowledge discovery engine. The algorithms presented here, as well as other machine learning approaches to tree induction, offer the opportunity of a major leap forward in our ability to comprehend complex spatial behavior, but also in our ability to use this newly acquired knowledge as a cornerstone of decision support systems for facility location planning and spatial planning in general. [1]

7. References

[1] The reader is referred to Reitsma (1990), Wright (1990), Arentze et al. (1995, 1996), and a few others for pioneering work on the articulation of expert systems in decision support environments for spatial planning.

Arentze, T.A., A.W.J. Borgers, and H.J.P. Timmermans 1995. "The Integration of Expert Knowledge in Decision Support Systems for Facility Location Planning," *Computers, Environment, and Urban Systems,* 19(4), 227-247.

Arentze, , T.A., A.W.J. Borgers, and H.J.P. Timmermans 1996. "An Efficient Search Strategy for Site-selection Decisions in an Expert System," *Geographical Analysis*, 28(2), 126-146.

Batsell, R.R., and J.C. Polking 1985. "A New Class of market Share Models," *Marketing Science*, 4(Summer), 177-198.

Bettman, J.R. 1979. *An Information Processing Theory of Consumer Choice*. Reading: Addison-Wesley.

Breiman, L., J. Friedman, R. Olshen, and C. Stone 1984. *Classification and Regression Trees*. Belmont: Wadsworth International.

Clark, W.A.V., and T.R. Smith 1985. "Production System Models of Residential Search Behavior: A Comparison of Behavior in Computer-simulated and Real-world Environments," *Environment and Planning A*, 17, 555-568.

Crompton, J. 1992. "Structure of vacation destination Choice Sets," *Annals of Tourism Research*, 19, 420-434.

Currim, I.S., R.J. Meyer, and N.T. Le 1988. "Disaggregate Tree-structured Modeling of Consumer Choice Data," *Journal of Marketing Research*, 25, 253-265.

Davis, R. and J. King 1976. "An Overview of Production Systems," in E.W. Elcock and D. Michie, editors, *Machine Intelligence 8*, New York: Wiley, pp. 330-332.

Eagle, T.C. 1980. "A Review of Decision Rules and their Role in Spatial Choice," *Discussion Paper No.33, Discussion Paper Series*, University of Iowa, Department of Geography.

Esposito, F., D. Malerba, and G. Semeraro 1997. "A Comparative Analysis of Methods for Pruning Decision Trees," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(5), 476-491.

Fotheringham, A.S. 1998. "Consumer Store Choice and Choice Set Definition," *Marketing Science,* 7, 299-310.

Freeman, T.A., and D.M. Skapura 1991. *Neural Networks: Algorithms, Applications and Programming Techniques*. Reading: Addison-Wesley.

Gärling, T., M.-P. Kwan, and R.G. Golledge 1994. "Computational-Process Modelling of Household Activity Scheduling," *Transportation Research B*, 28(5), 355-364.

Gensch, D. 1987. "A Two-Stage Disaggregate Attribute Choice Model," *Marketing Science*, 6 (Summer), 223-239.

Goldberg, D.E. 1989. *Genetic Algorithms in Search, Optimization, and Machine Learning*. Reading: Addison-Wesley.

Greene, D.P., and S.F. Smith 1987. "A Genetic System for Learning Models of Consumer Choice," in J.J. Grefenstette, editor, *Genetic Algorithms and Their Applications: Proceedings of the Second International Conference on Genetic Algorithms*, Hillsdale: Lawrence Erlbaum Associates, pp. 217-223.

Howard, J.A. 1963. *Marketing Management*. Homewood: Irwin Publishing.

Howard, J.A. and J.N. Seth 1969. *The Theory of Buying Behavior*. New York: Wiley.

Hunt, E.B., J. Marin, and P.J. Stone 1966. *Experiments in Induction*. New York: Academic Press.

Johnson, E. and R. Meyer 1984. "Compensatory Choice Models of Noncompensatory Processes: The Effect of Varying Context," *Journal of Consumer Research*, 11, 528-541.

Kim, H., and G.J. Koehler 1995. "Theory and Practice of Decision Tree," *Omega*, 23(6), 637-652.

Kolodner, J. 1993. *Case-based Reasoning*. San Mateo: Morgan Kaufmann Publishers.

Kwan, M.-P. and R.G. Golledge 1997. "Computational Process Modelling of Disaggregate Travel Behaviour," in M.M. Fischer and A. Getis, editors, *Recent Developments in Spatial Analysis*, Berlin: Springer, 171-185.

Manski, C.F. 1977. "The Structure of Random Utility Models," *Theory and Decision*, 8, 229-254.

Metropolitan Council. 1992. *Home Interview Survey. Methodology and Results*. Publication No. 550-92-061, Metropolitan Council, St. Paul, MN.

Meyer, R.J., and T.C. Eagle 1982. "Context-Induced Parameter Instability in a Disaggregate-Stochastic Model of Store Choice," *Journal of Marketing Research*, 19 (February), 62-71.

Moore, W.L., D.R. Lehmann, and E.A. Pessemier 1986. "Hierarchical Representations of Market Structures and Choice Processes through Preference Trees," *Journal of Business Research*, 14, 371-386.

Morgan, J.N., and J.A. Sonquist 1963. "Problems in the Analysis of Survey Data, and a Proposal," *Journal of the American Statistical Association*, 58 (September), 415-434.

Newell, A., and H.A. Simon 1972. *Human Problem Solving*. Englewood Cliffs: Prentice-Hall.

Nicosia, F.M. 1966. *Consumer Decision Processes: Marketing and Advertising Implications*. Englewood Cliffs: Prentice-Hall.

Oliver, J.R. 1993. "Discovering Individual Decision Rules: An Application of Genetic Algorithms," in S. Forrest, editor, *Proceedings of the Fifth International Conference on Genetic Algorithms*, San Mateo: Morgan Kauffman, pp. 216-222.

Oliver, J.R. 1994. "Finding Decision Rules with Genetic Algorithms," *AI Expert*, 9(3), 33-39.

Oskamp, A. 1994. "LocSim: A Probabilistic Model of Choice Heuristics," *Netherlands Journal of Housing and the Built Environment*, 9(3), 285-309.

Perreault, W.D., and H.C. Barksdale 1980. "A Model-free Approach to Analysis of Complex Contingency Data in Marketing Research," *Journal of Marketing Research*, 18 (November), 503-515.

Quinlan, J.R. 1979. "Discovering Rules by Induction from Large Collection of examples," in D. Michie, editor, *Expert Systems in the Micro Electronic Age*. Edingurgh, UK: Edinburgh University Press.

Quinlan, J.R. 1986. "Induction of Decision Trees," *Machine Learning*, 1, 81-106.

Quinlan, J.R. 1987. "Simplifying Decision Trees," *International Journal of Man-Machine Studies*, 27, 221-234.

Quinlan, J.R. 1990. "Decision Trees and Decision Making," *IEEE Transactions on Systems, Man, and Cybernetics*, 20(2), 339-346.

Quinlan, J.R. 1993. *C4.5: Programs for Machine Learning*. San Mateo: Morgan Kaufmann Publishers.

Rao, V.R., and D.J. Sabavala 1981. "Inference of Hierarchical Choice Processes from Panel Data," *Journal of Consumer Research*, 8 (June), 85-96.

Recker, W.W., and T. Golob 1979. "A Noncompensatory Model of Transportation Behavior Based on a Sequential Consideration of Attributes," *Economic Geography*, 57, 373-383.

Reitsma, R.F. *Functional Classification of Space: Aspects of Site Suitability Assessment in a Decision Support Environment*. Laxenburg: International Institute for Applied Systems Analysis.

Smith, T.R., W.A.V. Clark, and J.W. Cotton 1984. "Deriving and Testing Production System Models of Sequential Decision-making Behavior," *Geographical Analysis*, 16(3), 191-222.

Smith, T.R., and C.G. Lundberg 1984. "Psychological Foundations of Individual Choice Behaviour and a New Class of Decision Making Models," in G. Bahrenberg, M.M. Fischer, and P. Nijkamp, editors, *Recent Developments in Spatial Data Analysis*, Aldershot: Gower.

Smith, T.R., J.W. Pellegrino, and R.G. Golledge 1982. "Computational Process Modeling of Spatial Cognition and Behavior," *Geographical Analysis*, 14(4), 305-325.

Svenson, O. 1979. "Process Descriptions of Decision Making," *Organizational Behavior and Human Performance*, 23, 86-112.

Thill, J.-C. 1992. "Choice Set Formation for Destination Choice Modelling," *Progress in Human Geography*, 16, 361-382.

Thill, J.-C., and J.L. Horowitz 1997 "Modeling Non-Work Destination Choices with Choice Sets Defined by Travel-Time Constraints," in Getis A., and M.M. Fischer, editors, *Recent Developments in Spatial Analysis- Spatial Statistics, Behavioural Modelling and Neurocomputing*, Springer, Nuremberg, 1997, 186-208.

Thill, J.-C., and H.J.P. Timmermans 1992. "Analyse des Décisions Spatiales et du Processus de Choix des Consommateurs: Théories, Méthodes et Exemples d'Applications," *L'Espace Géographique*, 21(2), 144-166.

Timmermans, H.J.P. 1983. "Noncompensatory Decision Rules and Consumer Spatial Choice Behavior: A Test of Predictive Ability," *Professional Geographer*, 35, 449-485.

Timmermans, H.J.P. 1984. "Decision Models for Predicting Preferences among Multiattribute Choice Alternatives," in Bahrenberg, G., M.M. Fischer, and P. Nijkamp, editors, *Recent*

Developments in Spatial Data Analysis: Methodology, Measurement, Models, Aldershot: Gower, 337-354.

Timmermans, H.J.P., and R.G. Golledge 1990. "Applications of Behavioural Research on Spatial Problems II: Preference and Choice," *Progress in Human Geography*, 14(3), 311-354.

Timmermans, H.J.P., and R. Van der Heijden 1987. "Uncovering Spatial Decision-making Processes: A Decision Nets Approach Applied to Recreational Choice Behaviour," *Tijdschrift voor Economische en Sociale Geografie*, 78, 297-304.

Tversky, A. 1972. "Elimination by Aspects: A Theory of Choice," *Psychological Review*, 79, 281-299.

Tversky, A., and S. Sattath 1979. "Preference Trees," *Psychological Review*, 86, 542-573.

UCGIS 1997. *Spatial Analysis in a GIS Environment*. **Error! Bookmark not defined.**, University Consortium for Geographic Information Science, Washington, DC.

van Zwetselaar, M., and R. Goetgeluk 1994. "Decision Plan Nets of Housing Choice: A Critical Evaluation of the Reliability and Validity of this Technique," *Netherlands Journal of Housing and the Built Environment*, 9(3), 247-264.

Witlox, F. 1995. "Qualitative Housing Choice Modelling: Decision Plan Nets versus Decision Tables," *Netherlands Journal of Housing and the Built Environment*, 10(3), 209-237.

Wright, J.E. 1990. "Towards an Integrated Spatial Information Systems," in T.J. Kim, L. Wiggins, and J.R. Wright, editors, *Expert Systems: Applications to Urban Planning*, Berlin: Springer Verlag, 43-66.

Name	Definition
DISTANCE	Travel distance
TIME	Travel time
POP90	1990 population count
RET_EM	1990 employment in retail activities
PERSERV_EM	1990 employment in personal services
MALL	1 if regional mall; 0 otherwise
AREA_TYPED1	1 if trip destination is in a developed area; 0 otherwise
AREA_TYPED2	1 if trip destination is in central city/CBDs; 0 otherwise
AREA_TYPED3	1 if trip destination is in an outlying business district; 0 otherwise
AREA_TYPEO1	1 if trip origin is a developed area; 0 otherwise
AREA_TYPEO2	1 if trip origin is in central city/CBDs; 0 otherwise
AREA_TYPEO3	1 if trip origin is in an outlying business district; 0 otherwise
PCOMLU	Percent of area occupied by commercial/service land use
AGE	Age of the decision maker
GENDER	1 if decision maker is male; 0 if female
INCOME	1 if annual household income is over $35,000; 0 otherwise
HHLDSIZE	Number of members in the decision maker's household
INFANTS	Number of infants under 5 in the household
CARS	Number of cars in the household

Table 1. Definition of Independent Variables.

Variable	Mean/Proportion (N = 667)	Standard Deviation
DISTANCE (km)	4.600	4.429
TIME (minutes)	7.81	6.28
POP90	2060.470	1767.973
RET_EM	454.583	550.556
PERSERV_EM	44.822	47.283
MALL (%)		
1	9.6	
0	90.4	
AREA_TYPED (%)		
1	32.4	
2	15.0	
3	23.8	
AREA_TYPEO (%)		
1	27.4	
2	20.3	
3	6.0	
PCOMLU (%)	7.42	6.62
AGE	46.7	15.3
GENDER (%)		
Male	37.9	
Female	62.1	
INCOME (%)		
< $35,000	25.3	
$35,000	74.7	
HHLDSIZE	3.015	1.304
INFANTS	0.253	0.569
CARS	2.244	1.012

Table 2. Descriptive Statistics

	True Prediction	False Prediction
Chosen	171	496
Unchosen	66641	60

Table 3. Contingency Table of Observed versus Predicted Choices.

Figure Caption

Figure 1. Tree structure down to the fifth depth level. At each node, the following information is reported: the node number (first row), the number of instances at the node (second row), the number of unchosen and chosen instances respectively (third row), and the test applied to the next branching, in any (fourth row). Nodes 3, 11, 13, 15, 17, 18, and 21 are terminal leaves. Nodes at depth over 5 can be found in the Appendix.

Appendix. Complete pruned decision tree trained on 67,367 observations. Nodes with less than 20 observations have been removed. Each line represents a node. For each node, we report the depth in the tree, the number of chosen and unchosen instances, and the condition met by observations in the node. The indentation increases with the depth of the tree.

Depth=0, Root #Chosen=66700 #Unchosen=667
Depth=1, Test: RET_EM=<1183 #Chosen=66700 #Unchosen=596
Depth=2, Test: TIME=<8.48min #Chosen=2642 #Unchosen=413
Depth=3, Test: PCOMLU<4.93% #Chosen=2193 #Unchosen=205
Depth=4, Test: PERSERV_EM<216 #Chosen=2193 #Unchosen=202
Depth=5, Test: DISTANCE=<1.88km #Chosen=331 #Unchosen=100
Depth=5, Test: DISTANCE=[1.88,6.14] #Chosen=1862 #Unchosen=102
Depth=6, Test: AREA_TYPED3=0 #Chosen=1764 #Unchosen=83
Depth=7, Test: AGE<17 #Chosen=10 #Unchosen=2
Depth=7, Test: AGE≥17 #Chosen=1754 #Unchosen=81
Depth=8, Test: POP90<1107 #Chosen=662 #Unchosen=16
Depth=9, Test: AREA_TYPEO2=0 #Chosen=489 #Unchosen=16
Depth=10, Test: AREA_TYPED2=0 #Chosen=448 #Unchosen=16
Depth=11, Test: HHLDSIZE=<6 #Chosen=432 #Unchosen=16
Depth=12, Test: AREA_TYPEO1=0 #Chosen=228 #Unchosen=12
Depth=13, Test: CARS<4 #Chosen=201 #Unchosen=9
Depth=14, Test: GENDER=0 #Chosen=134 #Unchosen=5
Depth=14, Test: GENDER=1 #Chosen=67 #Unchosen=4
Depth=13, Test: CARS≥4 #Chosen=27 #Unchosen=3
Depth=14, Test: GENDER=0 #Chosen=19 #Unchosen=1
Depth=14, Test: GENDER=1 #Chosen=8 #Unchosen=2
Depth=12, Test: AREA_TYPEO1=1 #Chosen=204 #Unchosen=4
Depth=13, Test: INCOME=0 #Chosen=91 #Unchosen=3
Depth=13, Test: INCOME=1 #Chosen=113 #Unchosen=1
Depth=11, Test: HHLDSIZE≥6 #Chosen=16 #Unchosen=0
Depth=10, Test: AREA_TYPED2=1 #Chosen=41 #Unchosen=0
Depth=9, Test: AREA_TYPEO2=1 #Chosen=173 #Unchosen=0
Depth=8, Test: POP90≥1107 #Chosen=1092 #Unchosen=65
Depth=9, Test: AREA_TYPEO3=0 #Chosen=1042 #Unchosen=64
Depth=10, Test: AREA_TYPEO1=0 #Chosen=618 #Unchosen=46
Depth=11, Test: AREA_TYPED2=0 #Chosen=443 #Unchosen=31
Depth=12, Test: INCOME=0 #Chosen=200 #Unchosen=16
Depth=13, Test: AREA_TYPED1=0 #Chosen=109 #Unchosen=7
Depth=13, Test: AREA_TYPED1=1 #Chosen=91 #Unchosen=9
Depth=14, Test: HHLDSIZE<5 #Chosen=83 #Unchosen=7
Depth=14, Test: HHLDSIZE≥5 #Chosen=8 #Unchosen=2
Depth=12, Test: INCOME=1 #Chosen=243 #Unchosen=15
Depth=13, Test: AREA_TYPEO2=0 #Chosen=207 #Unchosen=14
Depth=13, Test: AREA_TYPEO2=1 #Chosen=36 #Unchosen=1
Depth=11, Test: AREA_TYPED2=1 #Chosen=175 #Unchosen=15
Depth=12, Test: AREA_TYPEO2=0 #Chosen=6 #Unchosen=0
Depth=12, Test: AREA_TYPEO2=1 #Chosen=169 #Unchosen=15
Depth=13, Test: INCOME=0 #Chosen=102 #Unchosen=6
Depth=13, Test: INCOME=1 #Chosen=67 #Unchosen=9

Depth=10, Test: AREA_TYPEO1=1 #Chosen=424 #Unchosen=18
Depth=11, Test: AREA_TYPED1=0 #Chosen=185 #Unchosen=1
Depth=11, Test: AREA_TYPED1=1 #Chosen=239 #Unchosen=17
Depth=12, Test: HHLDSIZE<6 #Chosen=229 #Unchosen=15
Depth=12, Test: HHLDSIZE≥6 #Chosen=10 #Unchosen=2
Depth=9, Test: AREA_TYPEO3=1 #Chosen=50 #Unchosen=1
Depth=10, Test: AREA_TYPED1=0 #Chosen=24 #Unchosen=1
Depth=10, Test: AREA_TYPED1=1 #Chosen=26 #Unchosen=0
Depth=6, Test: AREA_TYPED3=1 #Chosen=98 #Unchosen=19
Depth=4, Test: PERSERV_EM≥216 #Chosen=3 #Unchosen=0
Depth=3, Test: PCOMLU≥0.0493 #Chosen=449 #Unchosen=208
Depth=4, Test: PERSERV_EM=[0,251) #Chosen=401 #Unchosen=208
Depth=5, Test: POP90<6721 #Chosen=401 #Unchosen=202
Depth=6, Test: DISTANCE<1.93km #Chosen=82 #Unchosen=76
Depth=7, Test: MALL=0 #Chosen=82 #Unchosen=75
Depth=8, Test: AGE<71 #Chosen=80 #Unchosen=66
Depth=9, Test: AREA_TYPEO2=0 #Chosen=75 #Unchosen=53
Depth=9, Test: AREA_TYPEO2=1 #Chosen=13 #Unchosen=5
Depth=8, Test: AGE≥71 #Chosen=9 #Unchosen=2
Depth=7, Test: MALL=1 #Chosen=1 #Unchosen=0
Depth=6, Test: DISTANCE≥1.93km #Chosen=325 #Unchosen=120
Depth=7, Test: AGE<75 #Chosen=318 #Unchosen=112
Depth=8, Test: AREA_TYPED2=0 #Chosen=280 #Unchosen=108
Depth=9, Test: AREA_TYPEO3=0 #Chosen=250 #Unchosen=102
Depth=10, Test: AREA_TYPEO1=0 #Chosen=136 #Unchosen=70
Depth=10, Test: AREA_TYPEO1=1 #Chosen=114 #Unchosen=32
Depth=11, Test: AREA_TYPED3=0 #Chosen=73 #Unchosen=27
Depth=12, Test: INFANTS<2 #Chosen=69 #Unchosen=27
Depth=13, Test: INCOME=0 #Chosen=34 #Unchosen=17
Depth=14, Test: AREA_TYPED1=0 #Chosen=6 #Unchosen=5
Depth=14, Test: AREA_TYPED1=1 #Chosen=28 #Unchosen=12
Depth=15, Test: HHLDSIZE<4 #Chosen=20 #Unchosen=8
Depth=15, Test: HHLDSIZE≥4 #Chosen=8 #Unchosen=4
Depth=13, Test: INCOME=1 #Chosen=35 #Unchosen=10
Depth=12, Test: INFANTS≥2 #Chosen=4 #Unchosen=0
Depth=11, Test: AREA_TYPED3=1 #Chosen=41 #Unchosen=5
Depth=12, Test: CARS<4 #Chosen=37 #Unchosen=5
Depth=12, Test: CARS≥4 #Chosen=4 #Unchosen=0
Depth=9, Test: AREA_TYPEO3=1 #Chosen=30 #Unchosen=6
Depth=10, Test: AREA_TYPED3=0 #Chosen=11 #Unchosen=6
Depth=10, Test: AREA_TYPED3=1 #Chosen=19 #Unchosen=0
Depth=8, Test: AREA_TYPED2=1 #Chosen=38 #Unchosen=4
Depth=9, Test: AREA_TYPEO2=0 #Chosen=22 #Unchosen=0
Depth=9, Test: AREA_TYPEO2=1 #Chosen=16 #Unchosen=4
Depth=10, Test: INFANTS=0 #Chosen=11 #Unchosen=4
Depth=10, Test: INFANTS≥0 #Chosen=5 #Unchosen=0
Depth=7, Test: AGE≥75 #Chosen=8 #Unchosen=7

Depth=5, Test: POP90≥6721 #Chosen=6 #Unchosen=0
Depth=4, Test: PERSERV_EM≥251 #Chosen=48 #Unchosen=0
Depth=2, Test: TIME≥8.48min #Chosen=64058 #Unchosen=183
Depth=3, Test: DISTANCE<9.03km #Chosen=8069 #Unchosen=113
Depth=4, Test: MALL=0 #Chosen=8045 #Unchosen=111
Depth=5, Test: PCOMLU<5.77% #Chosen=7167 #Unchosen=70
Depth=6, Test: AREA_TYPED3=0 #Chosen=6712 #Unchosen=56
Depth=7, Test: PERSERV_EM<4 #Chosen=1173 #Unchosen=20
Depth=8, Test: POP90<6836 #Chosen=1173 #Unchosen=18
Depth=9, Test: AREA_TYPED2=0 #Chosen=981 #Unchosen=18
Depth=10, Test: AGE<54 #Chosen=649 #Unchosen=16
Depth=11, Test: AREA_TYPEO3=0 #Chosen=625 #Unchosen=16
Depth=12, Test: AREA_TYPEO2=0 #Chosen=417 #Unchosen=14
Depth=13, Test: AREA_TYPED1=0 #Chosen=305 #Unchosen=6
Depth=14, Test: AREA_TYPEO1=0 #Chosen=225 #Unchosen=6
Depth=15, Test: HHLDSIZE<9 #Chosen=223 #Unchosen=6
Depth=15, Test: HHLDSIZE≥9 #Chosen=2 #Unchosen=0
Depth=14, Test: AREA_TYPEO1=1 #Chosen=80 #Unchosen=0
Depth=13, Test: AREA_TYPED1=1 #Chosen=112 #Unchosen=8
Depth=14, Test: INFANTS<2 #Chosen=109 #Unchosen=7
Depth=15, Test: AREA_TYPEO1=0 #Chosen=43 #Unchosen=6
Depth=16, Test: CARS<3 #Chosen=27 #Unchosen=3
Depth=17, Test: INCOME=0 #Chosen=11 #Unchosen=2
Depth=17, Test: INCOME=1 #Chosen=16 #Unchosen=1
Depth=16, Test: CARS≥3 #Chosen=16 #Unchosen=3
Depth=15, Test: AREA_TYPEO1=1 #Chosen=66 #Unchosen=1
Depth=14, Test: INFANTS≥2 #Chosen=3 #Unchosen=1
Depth=12, Test: AREA_TYPEO2=1 #Chosen=208 #Unchosen=2
Depth=11, Test: AREA_TYPEO3=1 #Chosen=24 #Unchosen=0
Depth=10, Test: AGE≥54 #Chosen=332 #Unchosen=2
Depth=11, Test: AREA_TYPEO3=0 #Chosen=318 #Unchosen=1
Depth=12, Test: AREA_TYPEO1=0 #Chosen=203 #Unchosen=0
Depth=12, Test: AREA_TYPEO1=1 #Chosen=115 #Unchosen=1
Depth=13, Test: INCOME=0 #Chosen=67 #Unchosen=0
Depth=13, Test: INCOME=1 #Chosen=48 #Unchosen=1
Depth=11, Test: AREA_TYPEO3=1 #Chosen=14 #Unchosen=1
Depth=9, Test: AREA_TYPED2=1 #Chosen=192 #Unchosen=0
Depth=8, Test: POP90≥6836 #Chosen=2 #Unchosen=0
Depth=7, Test: PERSERV_EM≥4 #Chosen=5539 #Unchosen=36
Depth=8, Test: POP90<3565 #Chosen=4494 #Unchosen=21
Depth=9, Test: GENDER=0 #Chosen=2718 #Unchosen=10
Depth=10, Test: AREA_TYPED2=0 #Chosen=1283 #Unchosen=8
Depth=11, Test: AREA_TYPEO3=0 #Chosen=1216 #Unchosen=7
Depth=12, Test: AREA_TYPEO1=0 #Chosen=801 #Unchosen=6
Depth=13, Test: AREA_TYPEO2=0 #Chosen=487 #Unchosen=5
Depth=14, Test: AREA_TYPED1=0 #Chosen=272 #Unchosen=3
Depth=14, Test: AREA_TYPED1=1 #Chosen=215 #Unchosen=2

Depth=15, Test: INCOME=0 #Chosen=74 #Unchosen=1
Depth=15, Test: INCOME=1 #Chosen=141 #Unchosen=1
Depth=13, Test: AREA_TYPEO2=1 #Chosen=314 #Unchosen=1
Depth=14, Test: INCOME=0 #Chosen=177 #Unchosen=0
Depth=14, Test: INCOME=1 #Chosen=137 #Unchosen=1
Depth=15, Test: AREA_TYPED1=0 #Chosen=31 #Unchosen=0
Depth=15, Test: AREA_TYPED1=1 #Chosen=106 #Unchosen=1
Depth=12, Test: AREA_TYPEO1=1 #Chosen=415 #Unchosen=1
Depth=11, Test: AREA_TYPEO3=1 #Chosen=67 #Unchosen=1
Depth=10, Test: AREA_TYPED2=1 #Chosen=1435 #Unchosen=2
Depth=9, Test: GENDER=1 #Chosen=1776 #Unchosen=11
Depth=10, Test: AREA_TYPEO1=0 #Chosen=1172 #Unchosen=4
Depth=11, Test: AREA_TYPED1=0 #Chosen=843 #Unchosen=4
Depth=12, Test: AREA_TYPEO3=0 #Chosen=767 #Unchosen=4
Depth=13, Test: AGE<63 #Chosen=516 #Unchosen=3
Depth=14, Test: HHLDSIZE<3 #Chosen=226 #Unchosen=3
Depth=15, Test: AREA_TYPED2=0 #Chosen=48 #Unchosen=1
Depth=15, Test: AREA_TYPED2=1 #Chosen=178 #Unchosen=2
Depth=14, Test: HHLDSIZE□3 #Chosen=290 #Unchosen=0
Depth=13, Test: AGE□63 #Chosen=251 #Unchosen=1
Depth=12, Test: AREA_TYPEO3=1 #Chosen=76 #Unchosen=0
Depth=11, Test: AREA_TYPED1=1 #Chosen=329 #Unchosen=0
Depth=10, Test: AREA_TYPEO1=1 #Chosen=604 #Unchosen=7
Depth=8, Test: POP90□3565 #Chosen=1045 #Unchosen=15
Depth=9, Test: AREA_TYPEO3=0 #Chosen=1003 #Unchosen=15
Depth=10, Test: AREA_TYPED1=0 #Chosen=885 #Unchosen=11
Depth=11, Test: AREA_TYPEO1=0 #Chosen=509 #Unchosen=9
Depth=12, Test: AGE<22 #Chosen=17 #Unchosen=2
Depth=12, Test: AGE□22 #Chosen=492 #Unchosen=7
Depth=13, Test: GENDER=0 #Chosen=309 #Unchosen=3
Depth=14, Test: INCOME=0 #Chosen=154 #Unchosen=1
Depth=14, Test: INCOME=1 #Chosen=155 #Unchosen=2
Depth=13, Test: GENDER=1 #Chosen=183 #Unchosen=4
Depth=11, Test: AREA_TYPEO1=1 #Chosen=376 #Unchosen=2
Depth=12, Test: AGE<60 #Chosen=272 #Unchosen=0
Depth=12, Test: AGE□60 #Chosen=104 #Unchosen=2
Depth=10, Test: AREA_TYPED1=1 #Chosen=118 #Unchosen=4
Depth=11, Test: AGE<36 #Chosen=32 #Unchosen=0
Depth=11, Test: AGE□36 #Chosen=86 #Unchosen=4
Depth=12, Test: INFANTS=0 #Chosen=82 #Unchosen=2
Depth=13, Test: GENDER=0 #Chosen=50 #Unchosen=2
Depth=14, Test: HHLDSIZE<4 #Chosen=34 #Unchosen=2
Depth=15, Test: AREA_TYPEO2=0 #Chosen=27 #Unchosen=2
Depth=15, Test: AREA_TYPEO2=1 #Chosen=7 #Unchosen=0
Depth=14, Test: HHLDSIZE□4 #Chosen=16 #Unchosen=0
Depth=13, Test: GENDER=1 #Chosen=32 #Unchosen=0
Depth=12, Test: INFANTS=1 #Chosen=4 #Unchosen=2

Depth=9, Test: AREA_TYPEO3=1 #Chosen=42 #Unchosen=0
Depth=6, Test: AREA_TYPED3=1 #Chosen=455 #Unchosen=14
Depth=7, Test: PERSERV_EM<49 #Chosen=455 #Unchosen=10
Depth=8, Test: CARS<6 #Chosen=452 #Unchosen=9
Depth=9, Test: INCOME=0 #Chosen=211 #Unchosen=8
Depth=10, Test: AREA_TYPEO3=0 #Chosen=203 #Unchosen=8
Depth=11, Test: POP90<1191 #Chosen=125 #Unchosen=3
Depth=12, Test: AREA_TYPEO2=0 #Chosen=61 #Unchosen=2
Depth=13, Test: AGE<51 #Chosen=34 #Unchosen=2
Depth=14, Test: HHLDSIZE<6 #Chosen=30 #Unchosen=2
Depth=15, Test: INFANTS=0 #Chosen=21 #Unchosen=2
Depth=15, Test: INFANTS□0 #Chosen=9 #Unchosen=0
Depth=14, Test: HHLDSIZE□6 #Chosen=4 #Unchosen=0
Depth=13, Test: AGE□51 #Chosen=27 #Unchosen=0
Depth=12, Test: AREA_TYPEO2=1 #Chosen=64 #Unchosen=1
Depth=11, Test: POP90□1191 #Chosen=78 #Unchosen=5
Depth=12, Test: AGE<60 #Chosen=53 #Unchosen=5
Depth=13, Test: AREA_TYPEO1=0 #Chosen=46 #Unchosen=5
Depth=14, Test: GENDER=0 #Chosen=27 #Unchosen=4
Depth=14, Test: GENDER=1 #Chosen=19 #Unchosen=1
Depth=13, Test: AREA_TYPEO1=1 #Chosen=7 #Unchosen=0
Depth=12, Test: AGE□60 #Chosen=25 #Unchosen=0
Depth=10, Test: AREA_TYPEO3=1 #Chosen=8 #Unchosen=0
Depth=9, Test: INCOME=1 #Chosen=241 #Unchosen=1
Depth=8, Test: CARS□6 #Chosen=3 #Unchosen=1
Depth=7, Test: PERSERV_EM□49 #Chosen=4 #Unchosen=0
Depth=5, Test: PCOMLU□5.77% #Chosen=878 #Unchosen=41
Depth=6, Test: PERSERV_EM<54 #Chosen=380 #Unchosen=37
Depth=7, Test: AREA_TYPEO3=0 #Chosen=363 #Unchosen=37
Depth=8, Test: POP90<1220 #Chosen=274 #Unchosen=19
Depth=9, Test: AREA_TYPED3=0 #Chosen=66 #Unchosen=9
Depth=10, Test: AREA_TYPEO1=0 #Chosen=64 #Unchosen=7
Depth=11, Test: AGE<74 #Chosen=62 #Unchosen=6
Depth=12, Test: HHLDSIZE<4 #Chosen=37 #Unchosen=5
Depth=13, Test: GENDER=0 #Chosen=18 #Unchosen=4
Depth=13, Test: GENDER=1 #Chosen=19 #Unchosen=1
Depth=14, AREA_TYPED1=0 #Chosen=4 #Unchosen=1
Depth=14, Test: AREA_TYPED1=1 #Chosen=15 #Unchosen=0
Depth=12, Test: HHLDSIZE□4 #Chosen=25 #Unchosen=1
Depth=11, Test: AGE□74 #Chosen=2 #Unchosen=1
Depth=10, Test: AREA_TYPEO1=1 #Chosen=2 #Unchosen=2
Depth=9, Test: AREA_TYPED3=1 #Chosen=208 #Unchosen=10
Depth=10, Test: AGE<30 #Chosen=26 #Unchosen=0
Depth=10, Test: AGE□30 #Chosen=182 #Unchosen=10
Depth=11, Test: HHLDSIZE<2 #Chosen=20 #Unchosen=0
Depth=11, Test: HHLDSIZE□2 #Chosen=162 #Unchosen=10
Depth=12, Test: INFANTS<2 #Chosen=153 #Unchosen=10

Depth=13, Test: CARS<5 #Chosen=145 #Unchosen=10
Depth=14, Test: AREA_TYPEO1=0 #Chosen=112 #Unchosen=9
Depth=14, Test: AREA_TYPEO1=1 #Chosen=33 #Unchosen=1
Depth=13, Test: CARS≥5 #Chosen=8 #Unchosen=0
Depth=12, Test: INFANTS≥2 #Chosen=9 #Unchosen=0
Depth=8, Test: POP90≥1220 #Chosen=89 #Unchosen=18
Depth=9, Test: HHLDSIZE<2 #Chosen=2 #Unchosen=2
Depth=9, Test: HHLDSIZE≥2 #Chosen=87 #Unchosen=16
Depth=7, Test: AREA_TYPEO3=1 #Chosen=17 #Unchosen=0
Depth=6, Test: PERSERV_EM≥54 #Chosen=498 #Unchosen=4
Depth=7, Test: AGE<77 #Chosen=473 #Unchosen=3
Depth=8, Test: INCOME=0 #Chosen=234 #Unchosen=0
Depth=8, Test: INCOME=1 #Chosen=239 #Unchosen=3
Depth=9, Test: POP90<2515 #Chosen=101 #Unchosen=0
Depth=9, Test: POP90≥2515 #Chosen=138 #Unchosen=3
Depth=10, Test: INFANTS<2 #Chosen=127 #Unchosen=2
Depth=11, Test: HHLDSIZE<5 #Chosen=109 #Unchosen=1
Depth=11, Test: HHLDSIZE≥5 #Chosen=18 #Unchosen=1
Depth=10, Test: INFANTS≥2 #Chosen=11 #Unchosen=1
Depth=7, Test: AGE≥77 #Chosen=25 #Unchosen=1
Depth=4, Test: MALL=1 #Chosen=24 #Unchosen=2
Depth=3, Test: DISTANCE≥9.03km #Chosen=55989 #Unchosen=70
Depth=4, Test: PCOMLU<21.62% #Chosen=55989 #Unchosen=68
Depth=5, Test: AREA_TYPED3=0 #Chosen=50473 #Unchosen=52
Depth=6, Test: AREA_TYPEO1=0 #Chosen=37259 #Unchosen=49
Depth=7, Test: POP90<5843 #Chosen=35946 #Unchosen=49
Depth=8, Test: AREA_TYPED1=0 #Chosen=27573 #Unchosen=27
Depth=9, Test: AGE=<71 #Chosen=25823 #Unchosen=27
Depth=10, Test: PERSERV_EM<46 #Chosen=21414 #Unchosen=26
Depth=11, Test: AREA_TYPEO2=0 #Chosen=15999 #Unchosen=22
Depth=12, Test: INCOME=0 #Chosen=6539 #Unchosen=12
Depth=12, Test: INCOME=1 #Chosen=9460 #Unchosen=10
Depth=11, Test: AREA_TYPEO2=1 #Chosen=5415 #Unchosen=4
Depth=10, Test: PERSERV_EM=≥46 #Chosen=4409 #Unchosen=1
Depth=11, Test: CARS<2 #Chosen=487 #Unchosen=1
Depth=12, Test: AREA_TYPED2=0 #Chosen=203 #Unchosen=1
Depth=13, Test: HHLDSIZE<2 #Chosen=97 #Unchosen=1
Depth=13, Test: HHLDSIZE≥2 #Chosen=106 #Unchosen=0
Depth=12, Test: AREA_TYPED2=1 #Chosen=284 #Unchosen=0
Depth=11, Test: CARS≥2 #Chosen=3922 #Unchosen=0
Depth=9, Test: AGE≥71 #Chosen=1750 #Unchosen=0
Depth=8, Test: AREA_TYPED1=1 #Chosen=8373 #Unchosen=22
Depth=9, Test: PERSERV_EM<52 #Chosen=6525 #Unchosen=13
Depth=10, Test: AREA_TYPEO2=0 #Chosen=5064 #Unchosen=12
Depth=10, Test: AREA_TYPEO2=1 #Chosen=1461 #Unchosen=1
Depth=9, Test: PERSERV_EM≥52 #Chosen=1848 #Unchosen=9
Depth=10, Test: AREA_TYPEO3=0 #Chosen=1692 #Unchosen=9

Depth=11, Test: AGE<24 #Chosen=94 #Unchosen=0
Depth=11, Test: AGE≥24 #Chosen=1598 #Unchosen=9
Depth=12, Test: HHLDSIZE<2 #Chosen=105 #Unchosen=3
Depth=13, Test: GENDER=0 #Chosen=79 #Unchosen=1
Depth=13, Test: GENDER=1 #Chosen=26 #Unchosen=2
Depth=12, Test: HHLDSIZE=≥2 #Chosen=1493 #Unchosen=6
Depth=10, Test: AREA_TYPEO3=1 #Chosen=156 #Unchosen=0
Depth=7, Test: POP90≥5843 #Chosen=1313 #Unchosen=0
Depth=6, Test: AREA_TYPEO1=1 #Chosen=13214 #Unchosen=3
Depth=7, Test: INCOME=0 #Chosen=6475 #Unchosen=0
Depth=7, Test: INCOME=1 #Chosen=6739 #Unchosen=3
Depth=8, Test: GENDER=0 #Chosen=4243 #Unchosen=3
Depth=9, Test: PERSERV_EM<72 #Chosen=3893 #Unchosen=2
Depth=10, Test: AREA_TYPED1=0 #Chosen=3115 #Unchosen=1
Depth=10, Test: AREA_TYPED1=1 #Chosen=778 #Unchosen=1
Depth=9, Test: PERSERV_EM≥72 #Chosen=350 #Unchosen=1
Depth=10, Test: INFANTS=0 #Chosen=278 #Unchosen=0
Depth=10, Test: INFANTS≥0 #Chosen=72 #Unchosen=1
Depth=8, Test: GENDER=1 #Chosen=2496 #Unchosen=0
Depth=5, Test: AREA_TYPED3=1 #Chosen=5516 #Unchosen=16
Depth=6, Test: POP90<183 #Chosen=2639 #Unchosen=1
Depth=7, Test: GENDER=0 #Chosen=1633 #Unchosen=0
Depth=7, Test: GENDER=1 #Chosen=1006 #Unchosen=1
Depth=8, Test: INCOME=0 #Chosen=522 #Unchosen=1
Depth=9, Test: AREA_TYPEO1=0 #Chosen=357 #Unchosen=1
Depth=9, Test: AREA_TYPEO1=1 #Chosen=165 #Unchosen=0
Depth=8, Test: INCOME=1 #Chosen=484 #Unchosen=0
Depth=6, Test: POP90≥183 #Chosen=2877 #Unchosen=15
Depth=7, Test: PERSERV_EM<14 #Chosen=310 #Unchosen=8
Depth=8, Test: AREA_TYPEO1=0 #Chosen=267 #Unchosen=5
Depth=9, Test: AGE<20 #Chosen=10 #Unchosen=1
Depth=9, Test: AGE≥20 #Chosen=257 #Unchosen=4
Depth=8, Test: AREA_TYPEO1=1 #Chosen=43 #Unchosen=3
Depth=9, Test: AGE<62 #Chosen=33 #Unchosen=3
Depth=9, Test: AGE≥62 #Chosen=10 #Unchosen=0
Depth=7, Test: PERSERV_EM≥14 #Chosen=2567 #Unchosen=7
Depth=8, Test: AREA_TYPEO3=0 #Chosen=2417 #Unchosen=7
Depth=9, Test: INCOME=0 #Chosen=1177 #Unchosen=5
Depth=10, Test: AREA_TYPEO1=0 #Chosen=866 #Unchosen=5
Depth=11, Test: GENDER=0 #Chosen=530 #Unchosen=4
Depth=11, Test: GENDER=1 #Chosen=336 #Unchosen=1
Depth=12, Test: AGE<39 #Chosen=84 #Unchosen=1
Depth=12, Test: AGE≥39 #Chosen=252 #Unchosen=0
Depth=10, Test: AREA_TYPEO1=1 #Chosen=311 #Unchosen=0
Depth=9, Test: INCOME=1 #Chosen=1240 #Unchosen=2
Depth=8, Test: AREA_TYPEO3=1 #Chosen=150 #Unchosen=0
Depth=4, Test: PCOMLU≥21.62% #Chosen=2 #Unchosen=0

Depth=1, Test: RET_EM□2543 #Chosen=71 #Unchosen=0