Uncovering Complex Relationships in System Dynamics Modeling: Exploring the Use of CART, CHAID and SEM

Alexandra Medina-Borja, Ph.D
Department of Industrial Engineering
University of Puerto Rico at Mayagüez

II-205 Industrial Engineering Building
PO Box 9043
Mayagüez, PR 00681
Email: amedina@uprm.edu

Kalyan Sunder Pasupathy, Ph.D.
Department of Health Management & Informatics
School of Medicine

University of Missouri
311 Clark Hall, Columbia, MO 65211
Phone: (573) 882 – 7683
Email: pasupathyk@missouri.edu

Abstract

One of the premises of system dynamics is that the modeler would make relationship assumptions with enough precision to make the model useful. A common validation method is to consult with field experts, but with the advent of the internet, and automated data collection methods, knowledge is diluted as companies store abundant information without time to process it. Customers' dislikes, perceptions, intentions, opinions, and service characteristics reside in data warehouses (e.g. survey data is stored as categorical, nominal, ordinal or qualitative without further analysis). Without experts, companies are data rich but not necessarily knowledge rich. We present an application of known nonparametric predictive methodologies to uncover/confirm significant variable relationships and build the equations to feed the model: Classification and Regression Trees (CART), Chi-Square Automatic Interaction Detection (CHAID) and Structural Equation Modeling (SEM). A developing application of CHAID/SEM to explore restructuring decisions in a large service organization will be briefly discussed.

Key words: System dynamics, service quality, loyalty, relationships, data mining, CART, CHAID, structural equation modeling.

Overview

Modeling strategies vary from person to person and from problem to problem. One of the premises of system dynamics is that the modeler would be able to make assumptions about the relationships among variables with enough precision to make the model useful. The system dynamics approach should facilitate understanding into the analysis of the model and suggest behavior, but the modeler is not always sure of the validity of the structure. In many circumstances, the modeler is able to identify the important variables but the identification does not provide enough information for the modeler to make valid assumptions about mathematical relationships. The most common method to solve the relationships puzzle is to gain support from the literature or from a specific set of data collected for that purpose as to the direction of the relationship. The modeler can also ask a group of experts in the field to clarify the same. Then, an iterative modeling process begins when a simulation is run, and in many cases, adjusted after more data is collected. When data is not readily available to confirm expert knowledge, the equations representing the relationships may lack understanding as to why they are put together in a certain way.

The founders of the discipline believed that most of the information available to the modeler comes from the “actor’s heads” —their mental models or what Forrester (1994) called “mental databases”. Forrester recognized the mental database as the most important and significant source of information, placing the written database in the mid-range and giving to the numerical database the least importance both in magnitude and information about structures and policies. As the mental and written database contains mostly qualitative data, other authors have presented methods and models to deal with qualitative sources (see Luna-Reyes and Andersen, 2003). We argue, however, that all of the above presumes that knowledge about the system resides in some place. Yet, for many twenty first century business cases, organizational knowledge is diluted in several functions and departments. Contrary to measurement dilemmas of the past, when knowledge was kept in a group of senior experts who had experienced the organization in different capacities over the years, but had little or no data to support their expert knowledge, we live in an era of rapidly changing work environments, with specialized areas but very few knowledge integrators. Organizations today, contrary to the past, have an enormous amount of hard data collected through automated means, such as internet-based customer relationship management (CRM) systems, e-commerce, automated financial and service delivery systems, scanable and on-line customer satisfaction surveys, etc. With abundant information collected at a reduced cost, business analysts perform specific tasks, such as checking correlations for a specific project and, in many cases, millions of data points are stored without major exploration. Companies are data rich, but not necessarily knowledge rich.

Data collected every day is stored without any particular person concentrating on the analysis of changes and trends. If no one has the holistic approach of the organizational authorities of the past, finding an expert that will clarify the relationships is a challenging task. Even finding enough organizational documentation to point to the right direction can be difficult. We are living what Rygielski *et al.* (2002) call “the network economy” that has transformed business practices. Nonetheless, data today has always a “story to tell”.

We are concentrating this paper on one special case in which the modeler does not find readily available support to his/her theories, hunches and/or mental models and has data available to confirm these relationships.

Purpose

Intuitively, when data is abundant and no other sources of expert knowledge exist, one could expect that mathematics can settle the issue. Given the abundant information on customers likes, dislikes, perceptions, behaviors and opinions, and the multitude of options, including diversity of products, and services, data mining techniques are needed for decision-making. These techniques extract hidden predictive information from large databases, so that organizations are able to identify important patterns, predict future behaviors, and allow firms to make proactive, knowledge driven decisions. This is especially the case when the empirical evidence of the direction of the relationship resides in data collected from customers in the form of surveys, and stored as categorical, ordinal or qualitative in large data warehouses.

In some cases, we do not even know whether a relationship exists, such as the case of a new introduced technology or gadget and the number of returning customers. Automated systems may collect enough data but it might require the intervention of the market research department to uncover the outcome of such new product features. However, even when the existence of a relationship is known, uncovering the exact mathematical form of the relationship of intangible concepts described by survey data is not easy. For instance, how specific service quality characteristics (e.g. timeliness, empathy, knowledge) relate to customer retention and loyalty is likely to be a modeling challenge. Knowing that customer retention has the same direction as that of timeliness is intuitive. Nevertheless, knowing how exactly a stock variable, *number of customers*, is affected by a concept named "service quality" which in turn is composed by the customer reactions to a number of attributes or service dimensions, of which one of them is being more or less timely, is a very difficult question to answer. Statistics (a branch of mathematics) can also determine a formula for each relationship, which can be used, updated, refined, and reused over again. However, a methodology is needed to uncover "the formula" that relates an easily quantifiable variable to an abstract quality perception of the customer.

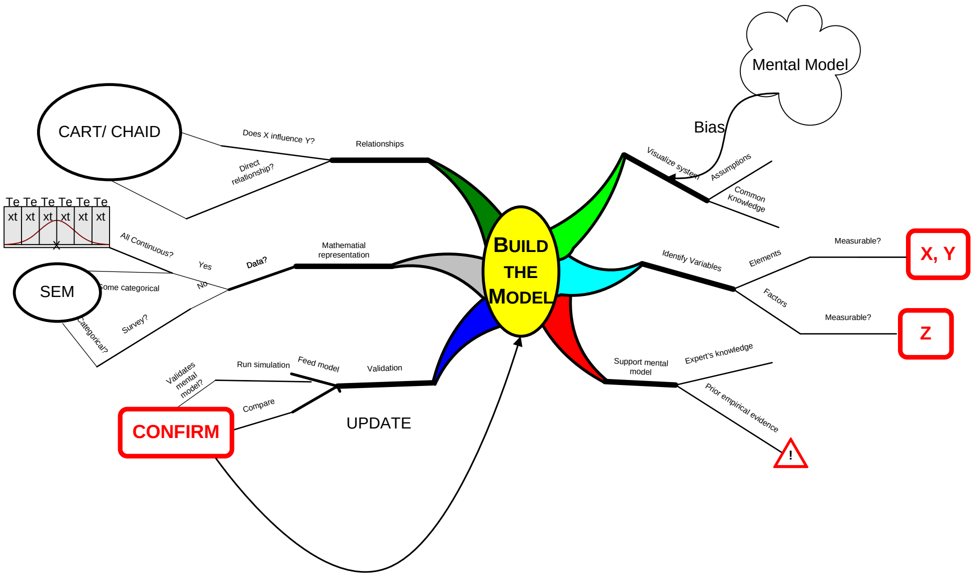
In this paper, we want to present a combination of three methodologies to uncover/confirm the significant relationships and build the equation to feed the model: Classification and Regression Trees (CART), Chi-Square Automatic Interaction Detection (CHAID) and Structural Equation Modeling (SEM). We offer the above as a potential solution for the problem of finding an adequate methodology to extract the relationships from a large data warehouse.

A Possible Modeling Strategy

While the system dynamics community is right in that the mental model of the researcher plays the most important role in the modeling process, it is undeniable that the "network society" has had an effect in the way modelers put together their systems. When abundant data resides in data warehouses, including customer data, the modeler needs to determine the relationships among variables that are sometimes abstract such as concepts, perceptions and opinions, collected through surveys. Figure 1 depicts what could be a mind map for such a situation. Here, the researcher has the main task of building a model. The first mental action is to retrieve a mental model with his/her biases, identify the issues, variables and factors and go to expert sources in search for support of his/her theories. This is a common process regardless of the nature of the problem, or the nature of the variables. Next, if expert knowledge is not available and data is, a potential modeling strategy would be to confirm the relationships through some statistical method. We propose that CART and CHAID can be used with this purpose.

However, once some of the variables can be classified as "constructs" of otherwise ambiguous perceptual concepts, such as marital happiness, service quality, customer satisfaction or political support, one would develop theories on how to measure such concepts creating a series of question items that are measured on some sort of ordinal scale (the most common one the Likert scale). But then, how these items are related to other variables in the problem, such as investment dollars, and ultimately, how can we uncover the exact mathematical representation of such a relationship so that a model is built, run, and gives useful and valid results. This is the scope of this paper.

Figure 1. Mind Map of the System Dynamics Modeling Process: Modeling Categorical Variables



Basic Algorithm for Decision Trees

Automatic tree classification methods are a family of methods that use recursive partitioning to find patterns in large data sets. As other nonparametric methods created to find patterns in the data, automatic decision trees try to overcome the limitations of parametric methods that assume linearity and therefore, can be used in a wider array of applications. Basically, all automatic tree methods follow the same algorithm:

1. Split into nodes
2. Grow branches
3. Terminate growth

Starting with the whole population in the data concentrated in a starting node (dependent or response variable), the algorithm looks for the best way to split the cases into a series of "parent" nodes and these cases into a series of "children" nodes. A pre-determined splitting criterion is followed systematically. In that way, cases are classified into branches and leaves. Through a series of termination rules, a node is declared either "undetermined" meaning that there is potential for growth and further classification, or "terminal" node, meaning that there isn't any further value in continuing the splitting.

When continuous or integer variables are part of the data set, there is potential for a huge number of data split interactions. Basically, any point can split the data. Because of this, splitting rules are developed that partition continuous data in categorical sub-sets.

The following sections discuss two of these methods, CART and CHAID in the context of SD modeling.

CART

CART is a binary decision tree whose proponents claim that it can automatically uncover the hidden structures in the data. CART was introduced originally by Freidman in 1977 and for those interested in details, an extensive methodological discussion is presented in Breiman *et al.* (1983).

In the literature, the main use of CART is that of identifying variables that are predictors of certain customer behavior. A set of rules or a profile is built based on the results, and whenever a new case arises the behavior is predicted based on the CART profile. The most common is of course that of credit decisions based on past customer data. While LOGIT and other parametric methods are also used, CART has been proved to be as or more efficient in cases where there is no assumption on the distribution of the variables (e.g. Galindo and Tamayo, 2000).

The algorithm divides the data in exactly two branches from each nonterminal node. The objective is to decrease heterogeneity. The response variable (dependent variable) can be quantitative or nominal (e.g. returned/ did not return, was satisfied with service or was dissatisfied, etc.) and the predictor variables can be nominal, ordinal, or continuous. Cross-validation and pruning are used to determine the size of the tree. Therefore, to build one such tree the modeler has to first grow the tree and then prune it.

In short, the algorithm divides the objects (data cases) in k different groups. The greatest amount of heterogeneity (or impurity) resides therefore at the top node. Then the data is split into sub-nodes that are significantly different. Each split contributes to the purity of the classification (i.e. to homogeneity of groups). Through this process, a set of important independent variables is revealed.

The validity of the model built through CART is done by cross-validating with another data set. There are issues around CART regarding the depth of the tree and pruning, but they are less worrisome than other assumptions in other methods. We are proposing that the same can be used to identify or confirm important predictors of any given variable in system dynamics.

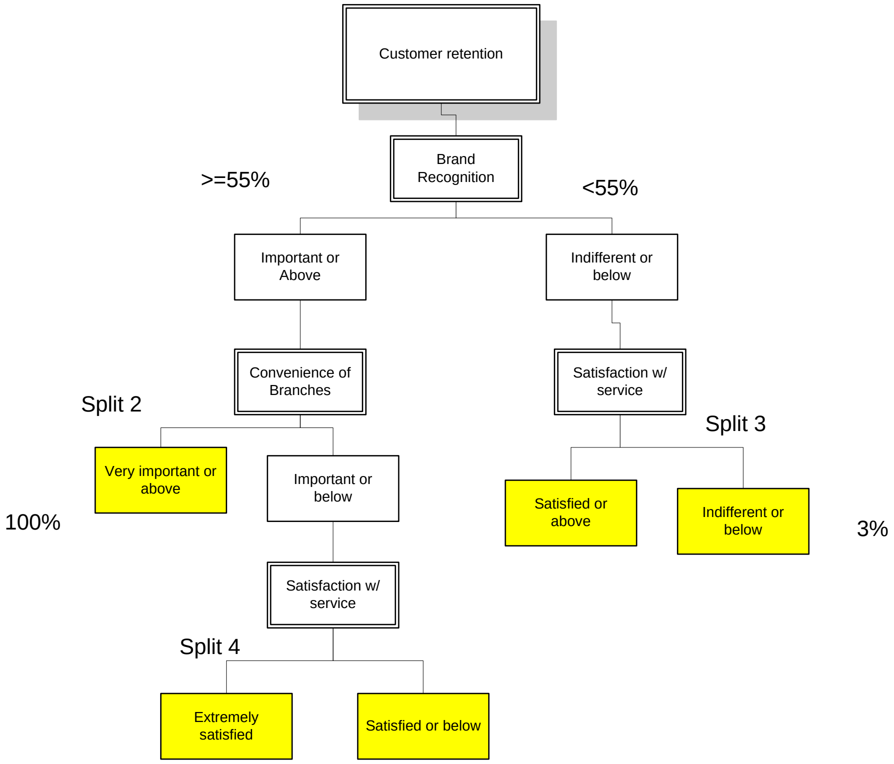
For example, let’s assume that an organization wants to re-engineer its operations by closing some of its branches in small towns, where apparently the presence of the company has no impact in the overall business. However, having wide presence might influence public opinion and brand image value. There are some not so obvious effects of having the branches in small places that are beyond pure financial numbers. It is hypothesized that having wide nation coverage would positively affect brand recognition, which in turn will positively affect both, customer retention and new customers. This is just one small piece of the system, as having more branches does have a financial effect, higher cost and perhaps not proportionately higher revenue generation. Customers from big branches were surveyed in the past and asked if nationwide coverage was important in their decision to do business with the company. They were also asked about their intention to continue. They answered Likert-scale type of questions from “strongly agree” to “strongly disagree” of the type: *I am satisfied with the number of branches, I do business in other cities, I feel a sense of security when I see a branch in a neighborhood other than mine.* Historic data on past closings and financial results were also available. Other satisfaction items were included in the survey.

CART was used to confirm the relationships and identify the most important predictors of customer retention. The tree in Figure 2 was created. The terminal nodes are shaded. According to the fictitious tree generated, brand recognition is the most important independent variable that affects customer retention. Customers who said that brand recognition is important or above will likely remain with the company in 55% of the cases. Affecting Brand recognition is convenience of branches. Those customers that said that branch convenience was very important or extremely important and for whom brand recognition was important have higher likelihood of continuing with the company. In fact, there is a probability close to 100% of that happening. Table 1 shows some of the potential rules associated with each terminal node.

For those who stated that convenience of branch availability was not as important (i.e. at the *important* or below rating in the Likert scale) but who also stated that they were extremely satisfied, the retention rate was high as well. This is confirmed by the other side of the coin, in which customers indifferent or dissatisfied had the lowest retention rate.

There are circumstances in which Chi-Square Automatic Interaction Detection, or CHAID, another nonparametric method, is more appropriate, such as when nominal variables are part of the data set or when the modeler wants to know how the independent variables interact with each level of the dependent variable (i.e. the researcher is interested in more than a dichotomous response). Like in the case when the dependent variable is customer satisfaction measured in a 5-point Likert scale and the modeler wants to see how Branch availability and Branch recognition interact to produce each of the five levels of measured satisfaction (from extremely dissatisfied to extremely satisfied). The next section explains the use of CHAID.

Figure 2. CART Results Identifying the Important Variables that Affect Customer



Retention

Table 1. Potential Rules Associated with each Terminal Node

Rules	*Potential response or outcome*	*Likelihood*
If the customer’s rating for satisfaction with service is indifferent or less; and brand recognition is indifferent or below	Losing customer Retaining customer	97% 3%
If the customer’s rating for satisfaction with service is very important or above; and brand recognition is important or above	Retaining customer	100%

CHAID

Morgan and Sonquist (1963) proposed a simple method for fitting trees to predict a quantitative variable. They called their original algorithm AID because it naturally incorporates interaction among predictors. Talking about Interaction, Wilkinson says:

> *“Interaction is not correlation. It has to do instead with conditional discrepancies. In the analysis of variance, interaction means that a trend within one level of a variable is not parallel to a trend within another level of the same variable. In the ANOVA model, interaction is represented by cross-products between predictors. In the tree model, it is represented by branches from the same node which have different splitting predictors further down the tree.” p.4*

The algorithm performs stepwise splitting by computing the within-cluster sum of squares about the mean of the cluster on the dependent variable.

CHAID is another type of decision tree method originally proposed by Kass (1980). According to Ripley (1996), the CHAID algorithm is a descendent of THAID developed after AID and discussed by Morgan and Messenger, (1973). CHAID is a combinatorial algorithm since it goes over all possible variable combinations in the data to partition the node. It is also an exploratory method used to study the relationship between a dependent variable and a series of predictor variables.

Categorical predictors that are not ordinal —such as ethnicity or race classification, or nominal options of the type of service provided— require a different approach. Since these types of nominal categories are unordered, all possible splits between categories must be considered. For deciding on one split of k categories into two groups, this means that 2 k -1 possible splits must be considered (Wilkinson, 1992). CHAID modeling selects a set of predictors and their interactions that optimally predict the dependent measure. The developed model shows how major "types" formed from the independent (predictor or splitter) variables differentially predict a criterion or dependent variable. The main difference between CHAID and CART is that

CHAID partitions the data in more than two groups, therefore, it discriminates more among categorical variables that are not necessarily binary. Any given node in CHAID can be partitioned in more than two groups.

The CHAID algorithm is particularly well suited for the analysis of larger datasets because the CHAID algorithm will often effectively yield many multi-way frequency tables (e.g., when classifying a categorical response variable with many categories, based on categorical predictors with many classes). One of the most common uses of CHAID has been in market segmentation to uncover customer characteristics for response modeling (see for example MacLennan and MacKenzie, 2000)

CHAID facilitates the development of predictive models, screen out extraneous predictor variables, and produce easy-to-read population segmentation subgroups. The splitting criteria are given by the non-parametric Chi-square test of independence, entropy measures and cross-validation differences. A larger Chi-square statistic suggests a more significant partition. Adjusted p-value measures of significance (using Bonferroni) are used to determine the best value of the partition, or the best split. Further, measures of entropy within the groups (a measure of information content within the split) are also used. An extensive explanation of how the CHAID algorithm works can be found in Wilkinson (1992). Other than the differences pointed out above, the logic behind CHAID and CART are very similar. Both clarify relationships among variables.

From Relationships to Mathematical Equations

Let us assume that the problem in our example of how CART works is part of a more comprehensive modeling endeavor. Any of the decision tree methods would help us determine whether a relationship exists, and the direction of the relationship. In the example given, we were able to determine that brand value affects customer retention in a positive way, the same as customer satisfaction. We also determined interactions with other variables. However, if we were to run the simulation model, we would be facing a problem since we do not know exactly how these categorical variables included in the survey (as part of a larger construct, for example, service quality) interact to affect customer retention, specially since they are measured in a Likert-type of scale.

In the next section, we briefly review survey design methodologies to help readers that are not familiar with the subject to understand how survey data are processed and why SEM works better for this type of problem.

Designing a Customer Survey

One of our favorite ways to explain survey design is the example of measuring people’s happiness. Happiness is a concept as abstract as service quality. Those who see it or experience it know it is there, but it is invisible, intangible and therefore subjective and difficult to measure. One cannot ask, “Are you happy?” “Yes or no.” Happiness has different degrees, and different nuances, and to be truly objective it is better to rely on the symptoms of happiness than on the simple self-evaluation of it.

If the researcher would rely on personal observations about happy people, she would probably include questions related to things she observed every time she was happy, or others around her seemed to be happy. Perhaps her mother used to wear a red dress every time she was happy, and used to wear a smile, and soften her voice, also presenting a joyful demeanor. Therefore, if she decides not to check the literature on the construct "happiness", but to rely on her mental model, she would include the following items in the initial pilot survey:

I am wearing a bright color today
My voice is soft
I feel joyful
I am smiling

In an attempt to measure the degree of the respondents' happiness, the researcher would include a scale in the survey to allow the respondent to choose among nuances of each question item above. The scale could include the following potential answers:

Describes me totally
Describes me
Does not describe me
Does not describe me at all

A numerical value would be assigned to each possible answer, 4 being assigned to "Describes me totally" and 1 assigned to "Does not describe me at all".

A pilot test with at least 50 customers would provide enough data to test the reliability of the construct "happiness" as it was built by the researcher (i.e. whether the four items above truly measure one's happiness or not). A statistic to measure the reliability of the construct would be calculated (generally Alpha Cronchbach) and if the Alpha statistic is close to 1, the items in the construct are correlated and therefore, assumed to be consistent and measuring the same thing. One can also test what would happen with the statistic if each of the items were to be removed from the survey, one at a time. If the Alpha coefficient increases, then the construct is better off without the question item. That is, the item is not consistent with the underlying concept being measured.

Let's assume that 20 persons stated that the sentences:

My voice is soft
I feel joyful
I am smiling

describe them totally. Of those, only 3 were actually wearing a bright color. Of those stating the opposite (i.e. that the three items above does not describe them at all) at least 3 stated that they were wearing a bright color. Therefore, the dress color does not seem to be consistent with the other 3 items, or in other words, it does not correlate with the other items (inter item correlation). Therefore, the internal consistency of the construct is better off if that item is not included since most likely the Alpha Cronchbach coefficient will increase if the statement about the color of cloth is deleted. The final questionnaire to elicit one's happiness will only include voice, demeanor and smile.

Now that we have explained how researchers build survey questions to measure "constructs", we are ready to move to the use of SEM to uncover the hidden mathematical relationships among variables. To make the link between the survey data and the hard data such as number of customers and service quality, Structural Equation Modeling or SEM could be used in conjunction with CHAID or CART. SEM is more of a confirmatory technique than an exploratory one. In fact, the two previous techniques discussed explore the potential relationships among variables in the data while SEM is more appropriate to confirm the relationships and build the mathematical model.

Structural Equation Modeling

Structural equation modeling (SEM) is a methodology used to model interactions, and nonlinearities among multiple latent independents measured by multiple indicators, and one or more latent dependents each with multiple indicators as well. SEM is a major component of applied multivariate statistical analysis and is used by biologists, economists, market researchers, and other social and behavioral scientists to study complex dependencies among variables in a causal framework. See for instance Hayduk, 1985; Bollen, 1989; Schumacker and Lomax, 1996; Pugesek *et al.* 2003.

Contrary to CHAID and CART, a causal model based on theory is first proposed and then tested for the data set. The model is used to test how well a model fits the data only to be accepted as a not-invalidated model. Alternatively, several proposed models can be compared against each other and based on the goodness-of-fit measures, the best model is chosen. We are proposing that the causal model could be based on exploratory methods such as CHAID and CART.

We now explain the main elements of SEM to familiarize the reader: indicators, latent variables, error terms and structural coefficients. *Indicators* are variables that are measured. They are also called as manifest variables or reference variables, such as items in a survey instrument. These indicators are used to measure unobserved variables or constructs or factors that represent an abstract concept, which are called *latent variables*. Error terms are associated with indicators and are explicitly modeled in SEM to capture the measurement error. *Structural coefficients* are the cause-and-effect sizes calculated by SEM and used to formulate the structural equations. In practice, most researchers use a hybrid approach, where a proposed theoretical model is tested with data. Then the modeler goes back to make changes in the model based on the SEM indexes. The problem of generalizability of the model (because it was modified based on a specific data set) to any data set can be overcome by a cross-validation strategy. Here the model is developed using a calibration or training data sample and then confirmed with a validation or testing sample.

Latent variable models are appropriate for continuous and discrete observed variables. Thus, SEM is especially well suited for discrete and categorical survey data. One can understand this if the concept of *latent variable* is understood. Normally, survey researchers use accepted statistical artifacts to get to the overall evaluation of the abstract construct under study. In our 'happiness" example, to evaluate how happy a person is, the researcher could either calculate the average of the responses to all the three proven items in the construct *happiness,* or find the best item to represent it. If a respondent answered “Describes me totally” to *I feel joyful* and *I am smiling* and answered “Does not describe me at all” to *My voice is soft*, the overall “happiness rating” would be the average of the numerical values (i.e. (4 + 4 + 1)/3 = 3. If another respondent answered “Does not describe at all” to all three items his/her rating would be one (1). The first respondent would be considered happier than the second one.

The same researcher using SEM, will approach the evaluation of the construct “happiness” in a different way. “Happiness” will be deemed a *latent variable*. The survey items, qualifying voice, demeanor, and smile will be the indicators. In Table 2, the observed variables are the indicators. These indicators are used to measure the latent variable *happiness*.

Table 2. Latent and Indicator Variables

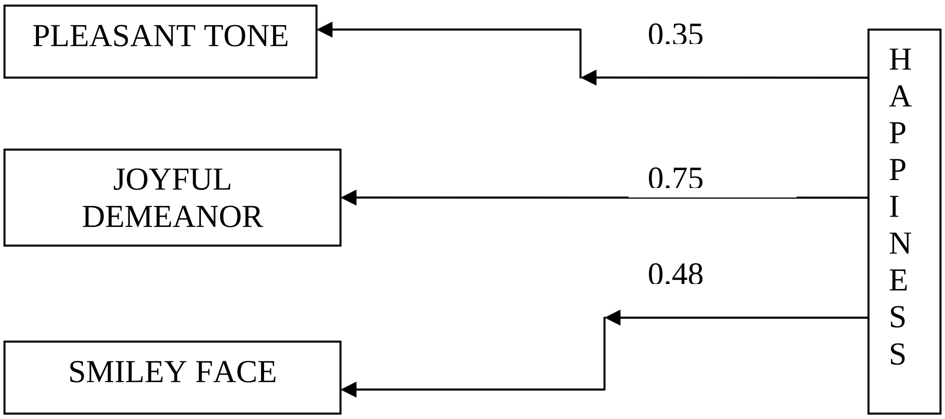
Dimension or Latent Variable	***Indicator or Observed Variables (usually measured by the item questions in a survey)***
HAPPINESS	**Voice tone** **Joyful demeanor** **Smiley face**

Each observed variable is measured with error, yet we would obtain unbiased measures of happiness. This can be done if we assume that the correlations across the observed variables arise from their common relation to the latent variable (local independence).
Similarly, we would like to obtain unbiased coefficients for the relation of happiness to other observed or latent variables (associations or causal effects).

The resulting model would be something like this:
Voice tone = 0.35 * Happiness
Joyful demeanor = 0.75 * Happiness
Smiley face = 0.48 * Happiness

Meaning that an increase in the Happiness of a survey respondent by one unit is shown by an increase in the Voice tone, Joyful demeanor and Smiley face respectively by 0.35, 0.75 and 0.48 (see Figure 3).

Figure 3. Happiness Construct

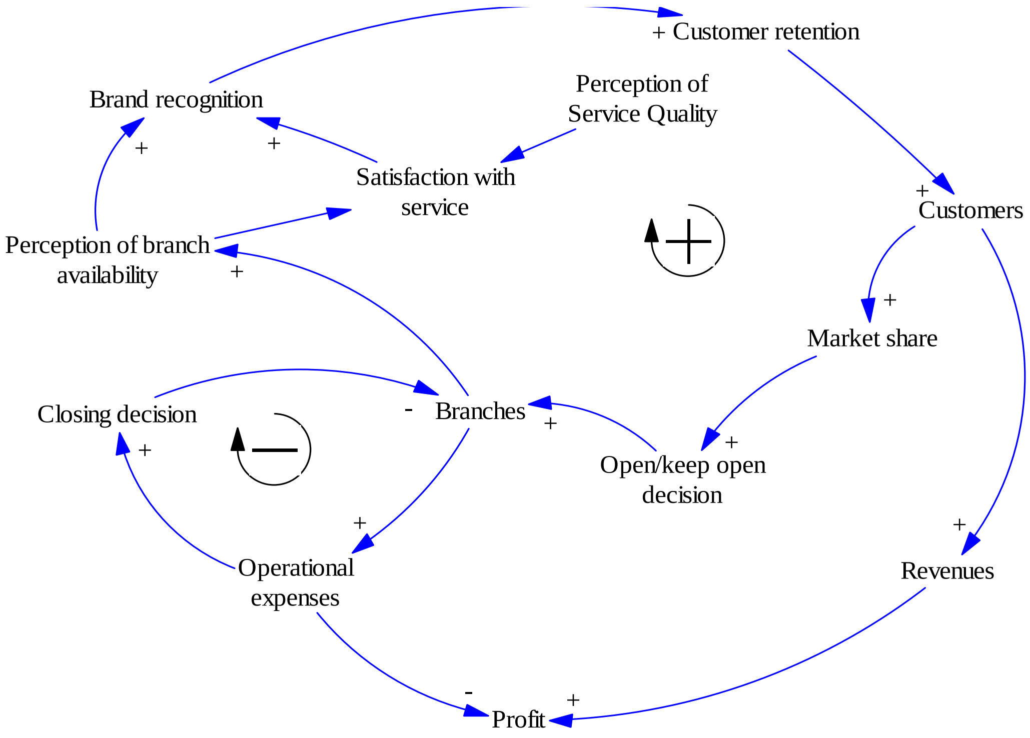


This is a radically different approach than the traditional average of the survey items. In fact, “Happiness” becomes a function that can be calculated and then related to other variables in the problem.

An illustrative example

In our example shown in Figure 2, we know that a level of satisfaction less than indifferent relates to the brand recognition, but we do not know in what magnitude, and how the same vary if the customer is “satisfied” instead. To clarify these, let us assume that the service company in our example wants to explore downsizing the number of branches. In this situation, customers perceive the quality of the services and tend to have certain levels of satisfaction. The extent to such satisfaction increases the loyalty of such customers and thus the recommendation of the service to their family and friends. Based on theory, the modeler could consider that an increase in number of branches increases the perceived service quality, all else being equal. As service quality increases, satisfaction also increases, more than otherwise would have been without enough branches. Again, with higher satisfaction, customer loyalty tends to increase. Let us assume that historical data shows a correlation between number of customers and the opening of new branches, suggesting that when the number of customers in one branch exceeds a certain range, management tends to make the decision of opening a new branch in the vicinity area. Branch availability in turn requires more operational expenses. For a re-structuring program, more expenses seem to be a negative consequence of more branches. More customers obviously will bring more revenue. Assuming, this very simple example accounts for all influencing variables, the right decision is a balance of all these interactions. The causal diagram for this situation is as shown in Figure 4.

Figure 4. Initial Casual Loop Diagram for Restructuring Decisions



If the components of service quality are measured in several survey items, then we could use SEM to uncover the actual mathematical equation of the relationship between these two variables (i.e. between the abstract concept of service quality and the hard number of returning customers). This relationship could have been previously uncovered by CHAID or CART.

Similarly, the construct of service quality in our restructuring problem is measured by ease of service procedures, knowledge of personnel, empathy/helpfulness and convenience of location. In addition, brand recognition is measured by image, logo recognition, uniqueness and bond with customer (Table 3).

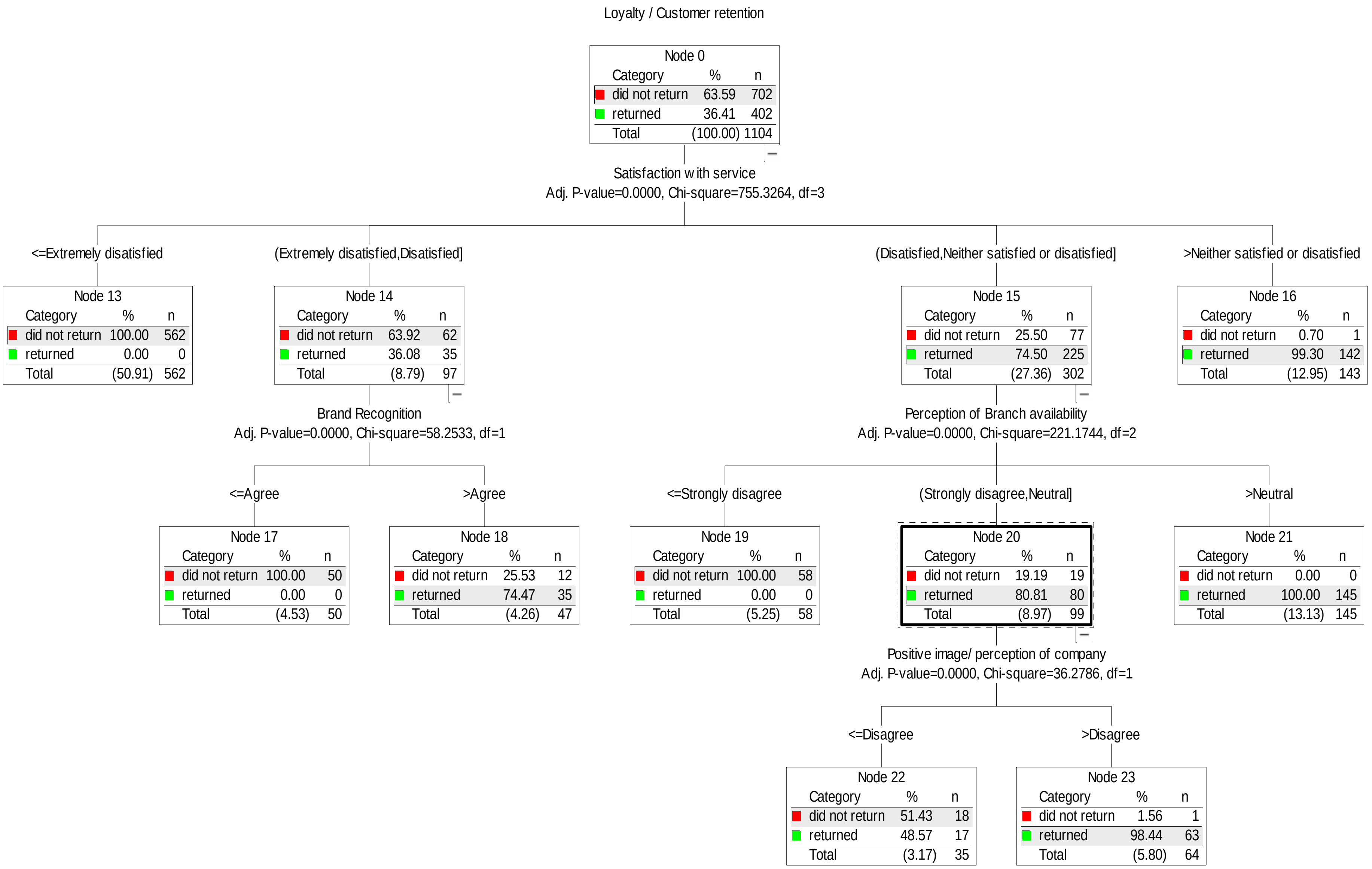
Table 3. Latent and Indicator Variables for the Restructuring Example

Latent variables	Indicators
Perceived Service Quality	Knowledge of personnel
	Ease of service procedures
	Empathy/helpfulness of personnel
	Convenience of location
	Responsiveness to customer needs
Brand/company recognition	Bond with customers – sense of security
	Positive image (setting it apart from others)
	Logo/name recognition
Customer satisfaction	Answer to “How satisfied are you with this service?
Perceived availability of branches	Answer to the question “There is a branch available whenever I need one”

All of the above indicators are part of a survey questionnaire distributed to customers. Thus, the customers’ perceived service quality (perceptions about personnel, service and location), satisfaction and brand recognition are captured this way. All other variables are hard data from company’s databases. Through customer relationship management systems, one can know whether a customer that gave a bad service quality evaluation and said to be dissatisfied, actually returned to make business in the future. By identifying the exact relationship that makes a customer return or not, it is possible to make the number of customers a stock variable and the flow is influenced by survey results. Further, it is known that a returning customer makes recommendations and referrals to friends and family, of which only 5% of the competitors are gained as new customers.

In our example, other variables measured as independent constructs are loyalty —measured as a binary variable for whether the customer returned or not — and the answer to the question measuring the customer’s perceptions of the availability of enough branches of the company. CHAID would produce the tree in Figure 5 showing that the strongest predictors of a returning customer are his/her satisfaction with the service, perception of branch availability, brand or name recognition and having a positive image of the company/service.

Figure 5. Predictors for Loyalty/Customer Retention


Loyalty / Customer retention
Satisfaction w ith service
Adj. P-value=0.0000, Chi-square=755.3264, df=3
<=Extremely disatisfied
(Extremely disatisfied,Disatisfied]
(Disatisfied,Neither satisfied or disatisfied]
>Neither satisfied or disatisfied
Brand Recognition
Adj. P-value=0.0000, Chi-square=58.2533, df=1
<=Agree
>Agree
Perception of Branch availability
Adj. P-value=0.0000, Chi-square=221.1744, df=2
<=Strongly disagree
(Strongly disagree,Neutral]
>Neutral
Positive image/ perception of company
Adj. P-value=0.0000, Chi-square=36.2786, df=1
<=Disagree
>Disagree

Node 0

Category	%	n
🟥 did not return	63.59	702
🟩 returned	36.41	402
Total	(100.00)	1104

Node 13

Category	%	n
🟥 did not return	100.00	562
🟩 returned	0.00	0
Total	(50.91)	562

Node 14

Category	%	n
🟥 did not return	63.92	62
🟩 returned	36.08	35
Total	(8.79)	97

Node 15

Category	%	n
🟥 did not return	25.50	77
🟩 returned	74.50	225
Total	(27.36)	302

Node 16

Category	%	n
🟥 did not return	0.70	1
🟩 returned	99.30	142
Total	(12.95)	143

Node 17

Category	%	n
🟥 did not return	100.00	50
🟩 returned	0.00	0
Total	(4.53)	50

Node 18

Category	%	n
🟥 did not return	25.53	12
🟩 returned	74.47	35
Total	(4.26)	47

Node 19

Category	%	n
🟥 did not return	100.00	58
🟩 returned	0.00	0
Total	(5.25)	58

Node 20

Category	%	n
🟥 did not return	19.19	19
🟩 returned	80.81	80
Total	(8.97)	99

Node 21

Category	%	n
🟥 did not return	0.00	0
🟩 returned	100.00	145
Total	(13.13)	145

Node 22

Category	%	n
🟥 did not return	51.43	18
🟩 returned	48.57	17
Total	(3.17)	35

Node 23

Category	%	n
🟥 did not return	1.56	1
🟩 returned	98.44	63
Total	(5.80)	64

We know, however that customer satisfaction is influenced also by the customer's perception of service quality dimensions. Therefore another tree was obtained having customer satisfaction as the dependent variable and all the four items in the construct service quality as potentially independent predictors. The resulting tree is shown in Figure 6. This tree shows that the most important predictor of satisfaction is the empathy or helpfulness of the service provider, followed by the perception of convenience of the location of the branches and ease of service procedures (which includes the time and paperwork involved in each transaction).

One could model these interactions among indicators or items in the survey. One method could be through IF-THEN statements that will account for the likelihood of a customer having a determined perception about the service. For example,

> IF the customer agrees or more that the service provider was helpful, exhibit empathy, and found the availability of branch locations as convenient THEN the likelihood of the customer being satisfied or extremely satisfied is 58%. However, IF the customer had the above characteristics and agreed or more than the service providers was knowledgeable about the service, THEN this likelihood of being extremely satisfied jumps to 100%.

We also ran a specific tree to uncover the relationships of the variables with the perception the customer had about the company image (Figure 7). In that we uncover that brand recognition and bond were also important predictors of a positive perception of the company's image, but so was satisfaction.

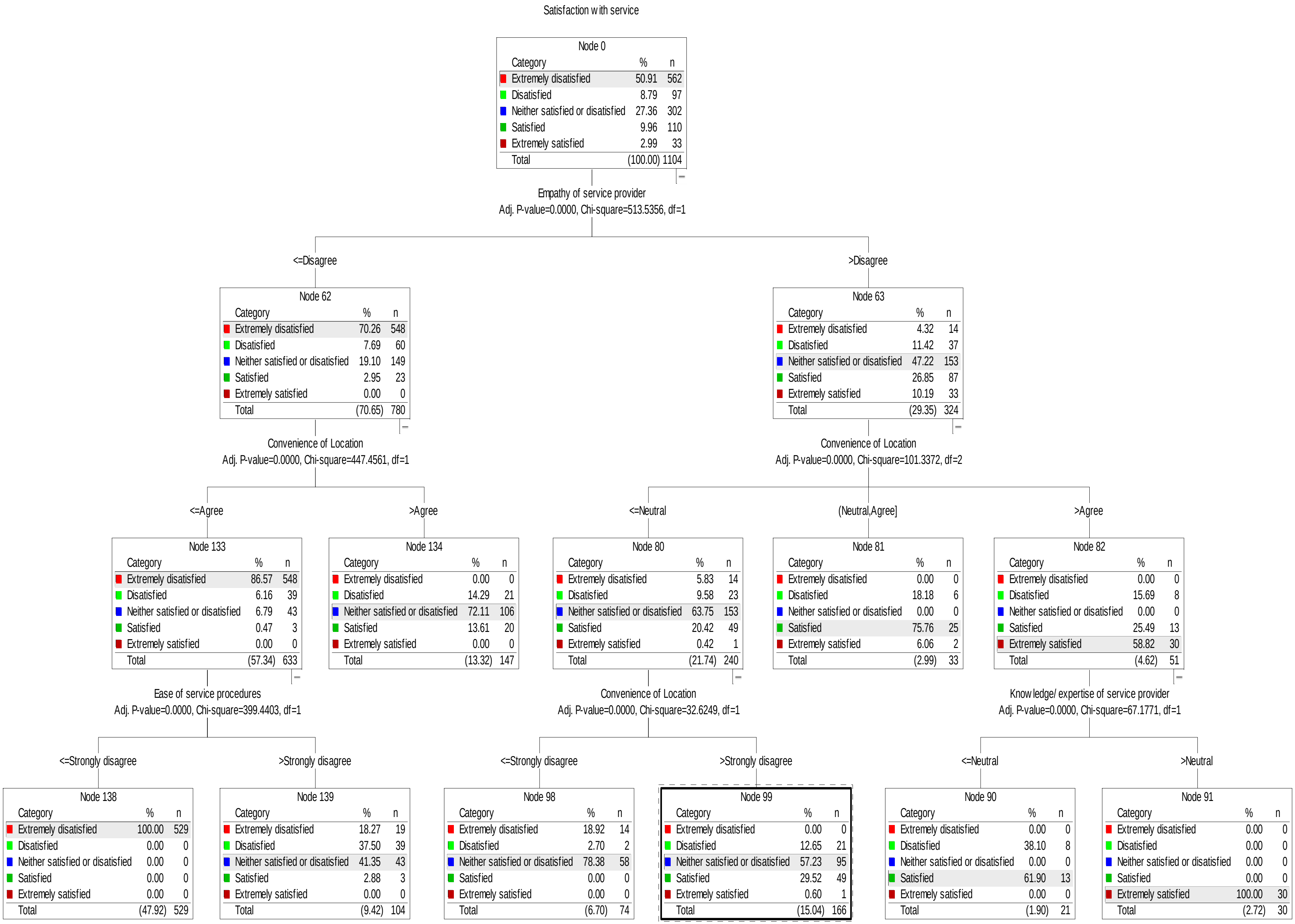
Having the same type of statements above describing the relationship of customer satisfaction with returning customers expressed in Figure 5, we could build a model including the relationships in Figure 6 and Figure 7. Therefore, one could conceive a model with survey data without using SEM. The model in Figure 8 would depict this model.

Since so far in this model we have not expanded the construct "Service Quality", the model in Figure 9 would be the expanded portion of Perceptions of Service Quality that affect customer satisfaction according to the tree in Figure 6. Here, *empathy of service provider* is the most important factor that affects a positive rating of satisfaction, and all other indicators refine this classification further, indicating they influence the perception of empathy and the final outcome of the evaluation.

However, one should observe that there is really no quantification of the increase/decrease in satisfaction and increase in retention due to the overall increase/decrease in the perception of quality. This is one disadvantage of using only CHAID or CART for modeling purposes. More importantly, SEM is necessary when we do not necessarily want to model the customer's individual reactions to items in the survey but instead we want to model the interaction of service quality as a whole concept (or latent concept) with the number of customers returning.

Using SEM the researcher would calculate the contribution to service quality of each one of its dimensions and then come up with a relationship that relates, not the survey items or dimensions, but the whole concept of service quality to satisfaction and then to loyalty.

Figure 6. Predictors for Satisfaction with Service



Satisfaction with service

Node 0

Category	%	n
🟥 Extremely disatisfied	50.91	562
🟩 Disatisfied	8.79	97
🟦 Neither satisfied or disatisfied	27.36	302
🟩 Satisfied	9.96	110
🟥 Extremely satisfied	2.99	33
Total	(100.00)	1104

Empathy of service provider
Adj. P-value=0.0000, Chi-square=513.5356, df=1
<=Disagree
>Disagree

Node 62

Category	%	n
🟥 Extremely disatisfied	70.26	548
🟩 Disatisfied	7.69	60
🟦 Neither satisfied or disatisfied	19.10	149
🟩 Satisfied	2.95	23
🟥 Extremely satisfied	0.00	0
Total	(70.65)	780

Node 63

Category	%	n
🟥 Extremely disatisfied	4.32	14
🟩 Disatisfied	11.42	37
🟦 Neither satisfied or disatisfied	47.22	153
🟩 Satisfied	26.85	87
🟥 Extremely satisfied	10.19	33
Total	(29.35)	324

Convenience of Location
Adj. P-value=0.0000, Chi-square=447.4561, df=1
Convenience of Location
Adj. P-value=0.0000, Chi-square=101.3372, df=2
<=Agree
>Agree
<=Neutral
(Neutral,Agree]
>Agree

Node 133

Category	%	n
🟥 Extremely disatisfied	86.57	548
🟩 Disatisfied	6.16	39
🟦 Neither satisfied or disatisfied	6.79	43
🟩 Satisfied	0.47	3
🟥 Extremely satisfied	0.00	0
Total	(57.34)	633

Node 134

Category	%	n
🟥 Extremely disatisfied	0.00	0
🟩 Disatisfied	14.29	21
🟦 Neither satisfied or disatisfied	72.11	106
🟩 Satisfied	13.61	20
🟥 Extremely satisfied	0.00	0
Total	(13.32)	147

Node 80

Category	%	n
🟥 Extremely disatisfied	5.83	14
🟩 Disatisfied	9.58	23
🟦 Neither satisfied or disatisfied	63.75	153
🟩 Satisfied	20.42	49
🟥 Extremely satisfied	0.42	1
Total	(21.74)	240

Node 81

Category	%	n
🟥 Extremely disatisfied	0.00	0
🟩 Disatisfied	18.18	6
🟦 Neither satisfied or disatisfied	0.00	0
🟩 Satisfied	75.76	25
🟥 Extremely satisfied	6.06	2
Total	(2.99)	33

Node 82

Category	%	n
🟥 Extremely disatisfied	0.00	0
🟩 Disatisfied	15.69	8
🟦 Neither satisfied or disatisfied	0.00	0
🟩 Satisfied	25.49	13
🟥 Extremely satisfied	58.82	30
Total	(4.62)	51

Ease of service procedures
Adj. P-value=0.0000, Chi-square=399.4403, df=1
Convenience of Location
Adj. P-value=0.0000, Chi-square=32.6249, df=1
Knowledge/ expertise of service provider
Adj. P-value=0.0000, Chi-square=67.1771, df=1
<=Strongly disagree
>Strongly disagree
<=Strongly disagree
>Strongly disagree
<=Neutral
>Neutral

Node 138

Category	%	n
🟥 Extremely disatisfied	100.00	529
🟩 Disatisfied	0.00	0
🟦 Neither satisfied or disatisfied	0.00	0
🟩 Satisfied	0.00	0
🟥 Extremely satisfied	0.00	0
Total	(47.92)	529

Node 139

Category	%	n
🟥 Extremely disatisfied	18.27	19
🟩 Disatisfied	37.50	39
🟦 Neither satisfied or disatisfied	41.35	43
🟩 Satisfied	2.88	3
🟥 Extremely satisfied	0.00	0
Total	(9.42)	104

Node 98

Category	%	n
🟥 Extremely disatisfied	18.92	14
🟩 Disatisfied	2.70	2
🟦 Neither satisfied or disatisfied	78.38	58
🟩 Satisfied	0.00	0
🟥 Extremely satisfied	0.00	0
Total	(6.70)	74

Node 99

Category	%	n
🟥 Extremely disatisfied	0.00	0
🟩 Disatisfied	12.65	21
🟦 Neither satisfied or disatisfied	57.23	95
🟩 Satisfied	29.52	49
🟥 Extremely satisfied	0.60	1
Total	(15.04)	166

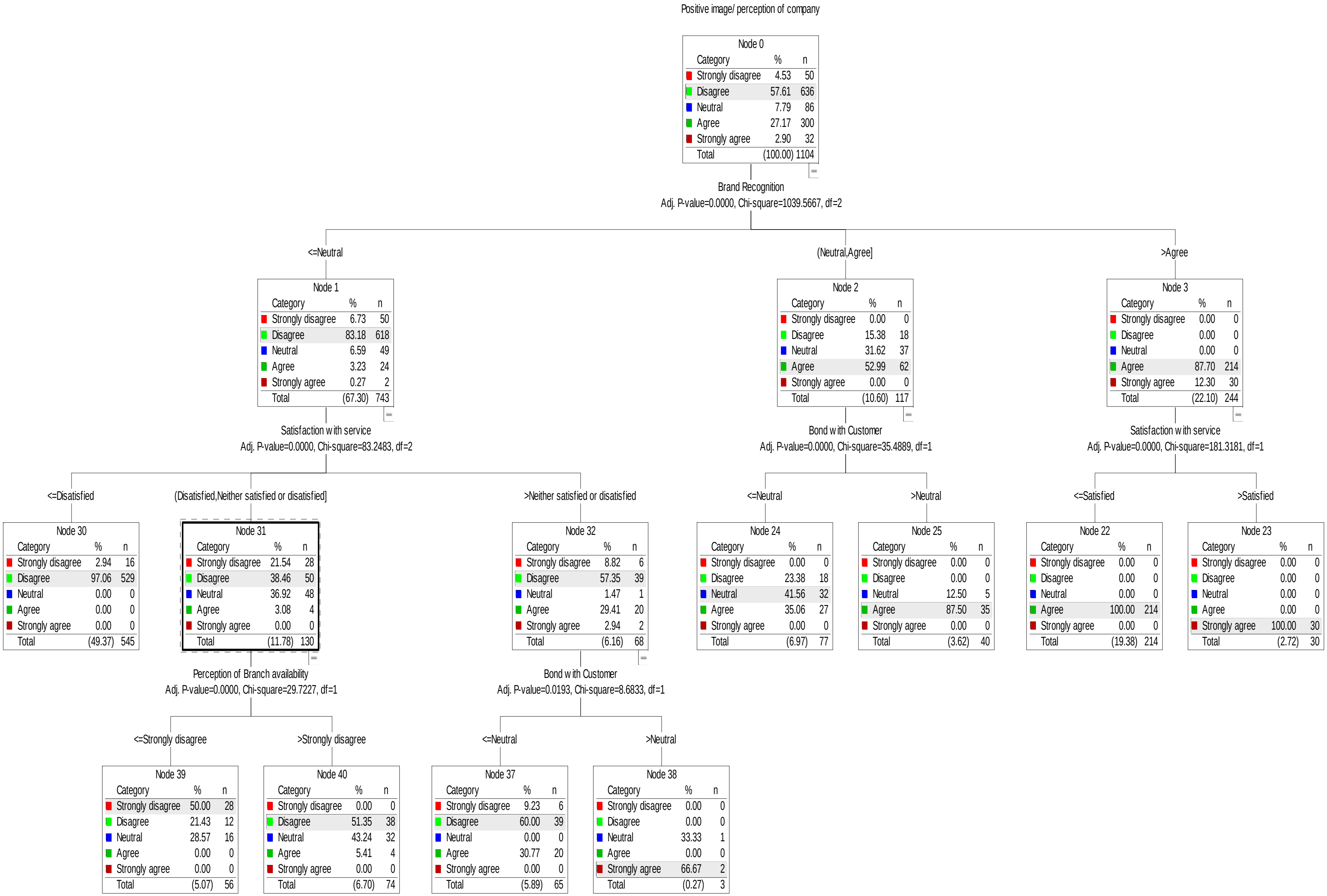
Node 90

Category	%	n
🟥 Extremely disatisfied	0.00	0
🟩 Disatisfied	38.10	8
🟦 Neither satisfied or disatisfied	0.00	0
🟩 Satisfied	61.90	13
🟥 Extremely satisfied	0.00	0
Total	(1.90)	21

Node 91

Category	%	n
🟥 Extremely disatisfied	0.00	0
🟩 Disatisfied	0.00	0
🟦 Neither satisfied or disatisfied	0.00	0
🟩 Satisfied	0.00	0
🟥 Extremely satisfied	100.00	30
Total	(2.72)	30

Figure 7. Predictors for Positive Image/Perception


Positive image/ perception of company

Node 0

Category	%	n
🟥 Strongly disagree	4.53	50
🟩 Disagree	57.61	636
🟦 Neutral	7.79	86
🟩 Agree	27.17	300
🟥 Strongly agree	2.90	32
Total	(100.00)	1104

Brand Recognition
Adj. P-value=0.0000, Chi-square=1039.5667, df=2
<=Neutral
(Neutral,Agree]
>Agree

Node 1

Category	%	n
🟥 Strongly disagree	6.73	50
🟩 Disagree	83.18	618
🟦 Neutral	6.59	49
🟩 Agree	3.23	24
🟥 Strongly agree	0.27	2
Total	(67.30)	743

Node 2

Category	%	n
🟥 Strongly disagree	0.00	0
🟩 Disagree	15.38	18
🟦 Neutral	31.62	37
🟩 Agree	52.99	62
🟥 Strongly agree	0.00	0
Total	(10.60)	117

Node 3

Category	%	n
🟥 Strongly disagree	0.00	0
🟩 Disagree	0.00	0
🟦 Neutral	0.00	0
🟩 Agree	87.70	214
🟥 Strongly agree	12.30	30
Total	(22.10)	244

Satisfaction with service
Adj. P-value=0.0000, Chi-square=83.2483, df=2
Bond with Customer
Adj. P-value=0.0000, Chi-square=35.4889, df=1
Satisfaction with service
Adj. P-value=0.0000, Chi-square=181.3181, df=1
<=Disatisfied
(Disatisfied,Neither satisfied or disatisfied]
>Neither satisfied or disatisfied
<=Neutral
>Neutral
<=Satisfied
>Satisfied

Node 30

Category	%	n
🟥 Strongly disagree	2.94	16
🟩 Disagree	97.06	529
🟦 Neutral	0.00	0
🟩 Agree	0.00	0
🟥 Strongly agree	0.00	0
Total	(49.37)	545

Node 31

Category	%	n
🟥 Strongly disagree	21.54	28
🟩 Disagree	38.46	50
🟦 Neutral	36.92	48
🟩 Agree	3.08	4
🟥 Strongly agree	0.00	0
Total	(11.78)	130

Node 32

Category	%	n
🟥 Strongly disagree	8.82	6
🟩 Disagree	57.35	39
🟦 Neutral	1.47	1
🟩 Agree	29.41	20
🟥 Strongly agree	2.94	2
Total	(6.16)	68

Node 24

Category	%	n
🟥 Strongly disagree	0.00	0
🟩 Disagree	23.38	18
🟦 Neutral	41.56	32
🟩 Agree	35.06	27
🟥 Strongly agree	0.00	0
Total	(6.97)	77

Node 25

Category	%	n
🟥 Strongly disagree	0.00	0
🟩 Disagree	0.00	0
🟦 Neutral	12.50	5
🟩 Agree	87.50	35
🟥 Strongly agree	0.00	0
Total	(3.62)	40

Node 22

Category	%	n
🟥 Strongly disagree	0.00	0
🟩 Disagree	0.00	0
🟦 Neutral	0.00	0
🟩 Agree	100.00	214
🟥 Strongly agree	0.00	0
Total	(19.38)	214

Node 23

Category	%	n
🟥 Strongly disagree	0.00	0
🟩 Disagree	0.00	0
🟦 Neutral	0.00	0
🟩 Agree	0.00	0
🟥 Strongly agree	100.00	30
Total	(2.72)	30

Perception of Branch availability
Adj. P-value=0.0000, Chi-square=29.7227, df=1
Bond with Customer
Adj. P-value=0.0193, Chi-square=8.6833, df=1
<=Strongly disagree
>Strongly disagree
<=Neutral
>Neutral

Node 39

Category	%	n
🟥 Strongly disagree	50.00	28
🟩 Disagree	21.43	12
🟦 Neutral	28.57	16
🟩 Agree	0.00	0
🟥 Strongly agree	0.00	0
Total	(5.07)	56

Node 40

Category	%	n
🟥 Strongly disagree	0.00	0
🟩 Disagree	51.35	38
🟦 Neutral	43.24	32
🟩 Agree	5.41	4
🟥 Strongly agree	0.00	0
Total	(6.70)	74

Node 37

Category	%	n
🟥 Strongly disagree	9.23	6
🟩 Disagree	60.00	39
🟦 Neutral	0.00	0
🟩 Agree	30.77	20
🟥 Strongly agree	0.00	0
Total	(5.89)	65

Node 38

Category	%	n
🟥 Strongly disagree	0.00	0
🟩 Disagree	0.00	0
🟦 Neutral	33.33	1
🟩 Agree	0.00	0
🟥 Strongly agree	66.67	2
Total	(0.27)	3

Figure 8. Resulting Causal Loop Diagram after CHAID Exploration

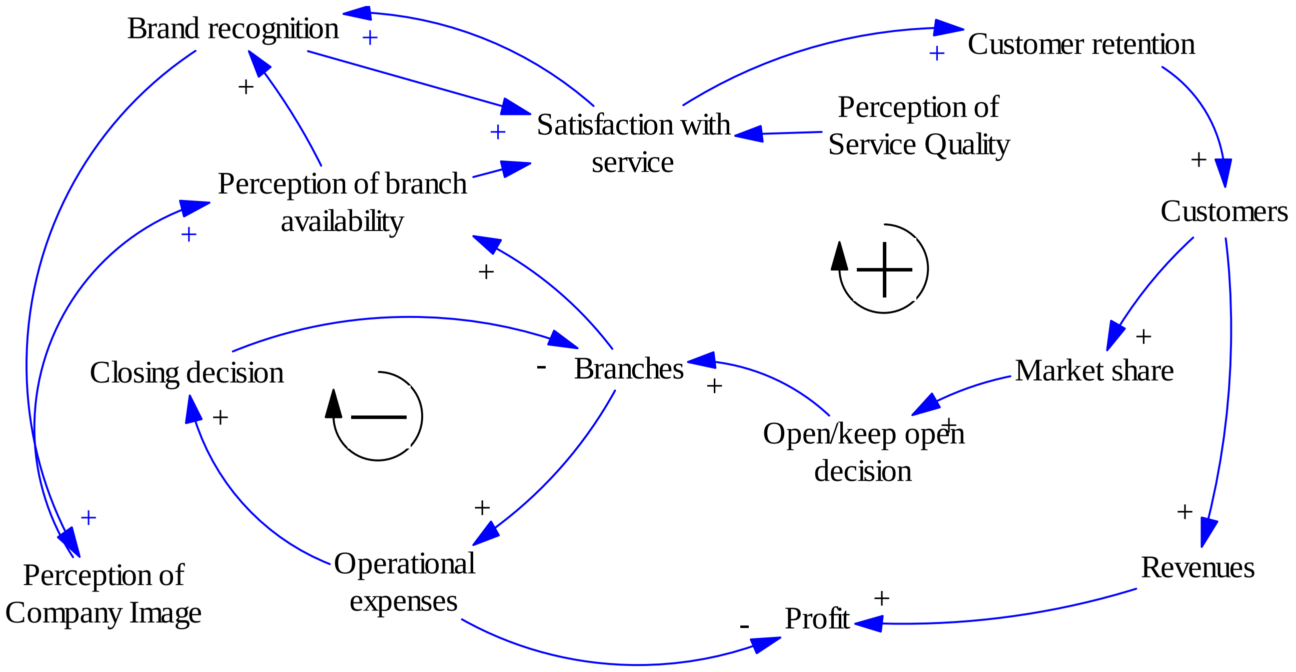


Figure 9. Expanding the Construct Service Quality

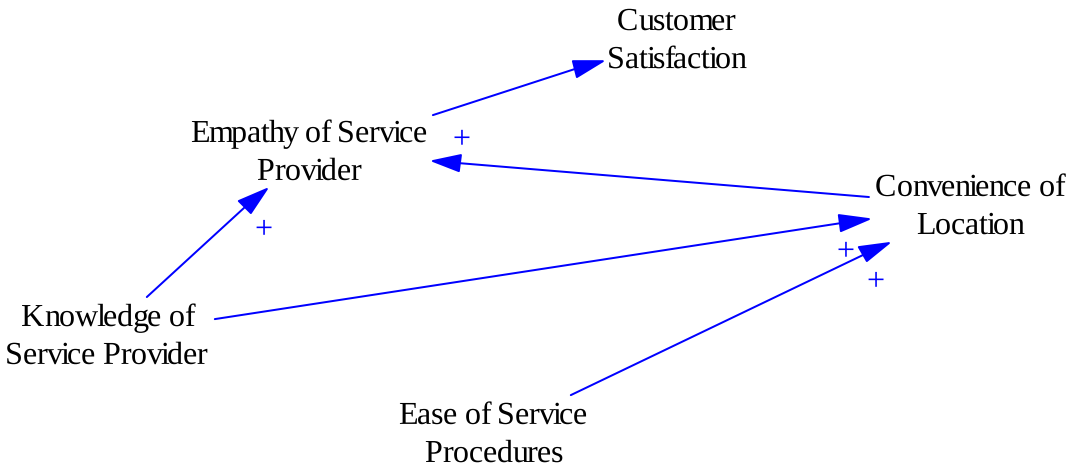
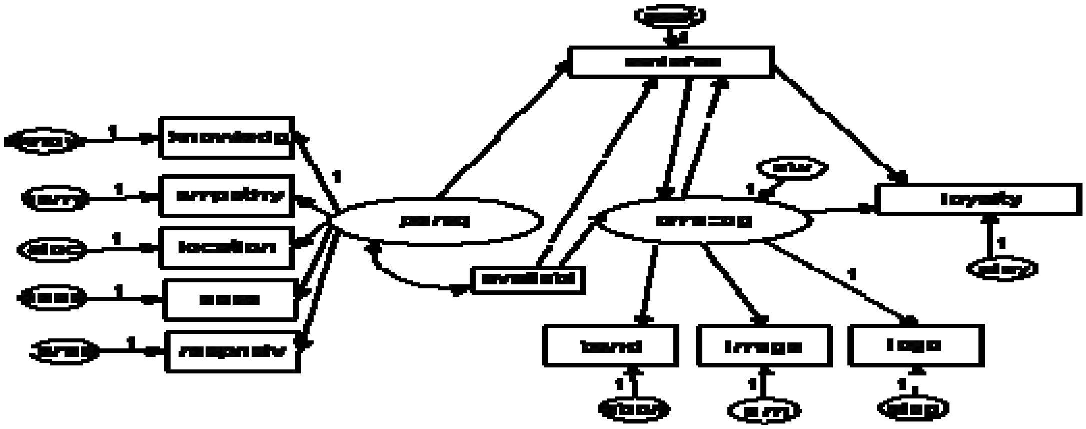

Customer Satisfaction
Empathy of Service Provider
Convenience of Location
Knowledge of Service Provider
Ease of Service Procedures
+

Figure 8 shows how one would model with SEM to represent the diagram in Figure 10. Latent variables are represented by the ellipses and indicators or measured variables by the rectangles. Each measured variable has an error term associated with it represented by the small circles.

While the contribution of each indicator (item in the survey) to the overall service quality will be estimated, the linkage that we will use in SD would be the one from the latent variable service quality to customer satisfaction, which estimates the increase in satisfaction for each unit of service quality. The researcher can then estimate the evaluation provided based on the contributions of all the variables affecting satisfaction.

Figure 10. Modeling the Restructuring Problem in Structural Equation Modeling (SEM)

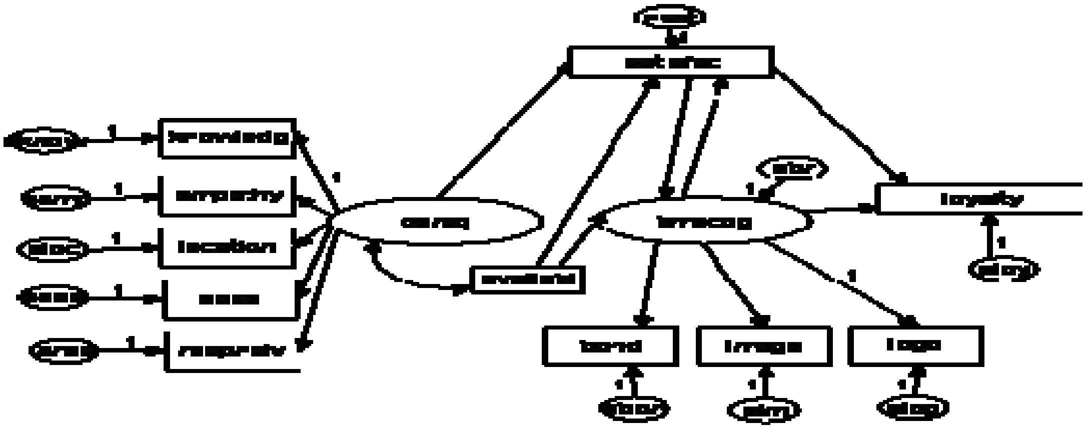


Another good feature of SEM is that it allows the researcher to explore variations to the model and see which one seems to be more appropriate. The commercially available software allows the user to explore alternative models and select the best fit one.

After running the above model, we obtained a series of parameters: estimate coefficients for the equations that link each variable, the covariances, variances, and the correlations estimates for the relationships among variables. To compare models, the researcher can use a number of statistics of goodness of fit. For the above model Chi-square was 6982.9 with 39 degrees of freedom and probability level p = 0.0001. All relationships were proven to be significant at the 0.0001 level except for the brand recognition influencing satisfaction which was significant at the 0.004 level with a negative coefficient (meaning that a higher level of recognition of the brand actually reduces the satisfaction, perhaps due to higher expectations) while a higher level of satisfaction increases brand recognition.

We then explored eliminating the two-way relationship from satisfaction to brand recognition, by first eliminating the link from brand recognition to satisfaction, leaving the link from satisfaction to brand recognition. The new model turned out to have a slightly lower Chi-square statistic (6954.5) and 40 degrees of freedom, showing that the two way model might have a better fit. A third model was also explored reversing the direction of the one-way relationship, under the theory that brand recognition influences satisfaction more than satisfaction influences recognition, even though both relationships were significant. This model had a higher fit than the other two with Chi-square of 7016.8. The resulting model with the coefficients is shown in Figure 11.

Figure 11. SEM Model with Parameter Estimates



In summary, for each unit of increase in service quality, the level of satisfaction would increase by 0.69, the latent Brand Recognition will influence satisfaction by 0.22 while the perception of branch availability will influence the evaluation of level of satisfaction by 0.16, and so forth.

Likewise, the latent variables Service quality and Brand Recognition were able to have a dimension, therefore, being linked to other variables in the model. By solving multiple equations, we can show that Service Quality can range between 0.2 and 1.3, 1.3 being high quality and 0.2 low quality. One could interpret this number as the most likely magnitude of service quality (a concept similar to the average of all the customer evaluations) given the other conditions in the model. This is useful when as said at the beginning, a relationship among variables is hidden in large amounts of data.

Table 4. Equations in the System Dynamics Restructuring Model

Relationship	Equation
Service Quality	0.279*Perception of branch availability
Satisfaction	0.683+(Perception of Service Quality*0.694+0.22*Brand recognition+0.16*Perception of branch availability)
Brand recognition	0.83*Perception of branch availability
Perception of branch availability	IF,THEN,ELSE statement based on CHAID results relating distance driven to branch for customers
Loyalty/ customer retention	IF THEN ELSE(Satisfaction with service>2.8,0.993 , IF THEN ELSE(Satisfaction with service<2.8 :AND: Brand recognition>=2.5,0.98,IF THEN ELSE(Satisfaction with service>2.8 :AND: Brand recognition<2.5, 0.423,0.053)))

Table 4 shows the equations input into Vensim for the SD modeling. All of the equations illustrated are linear in nature. Certain models will end up having non-linear relationships that are for example, products of linear combinations or combinations of piece-wise linear functions. Modeling latent variables is a mechanism to parse out measurement error by combining across observed variables (using correlations among variables) and allow for the estimation of complex causal models. In this paper, we show how one can use SEM to establish and quantify causal relationships that can be used later in system dynamics. Other decision variables, such as the Management Open/Close decision were based on pre-determined profitability and market share goals. Another introduced decision rule was that management would not open a new branch unless it had at least 6,500 customers per branch and that it would close a branch any time it had less than 2,000 customers per branch. The model then could be used to evaluate those policies. Data in this example turned out to drive radical retention rules that may not be as realistic, whenever satisfaction and brand recognition was low, the retention rate was minimal (around 5%) therefore the behavior of the level variable *customers* was not very realistic. The SD model created is shown in Figure 12 and Figure 13 shows the behavior of key variables over time.

Figure 12. Resulting System Dynamics Model for the Restructuring Problem

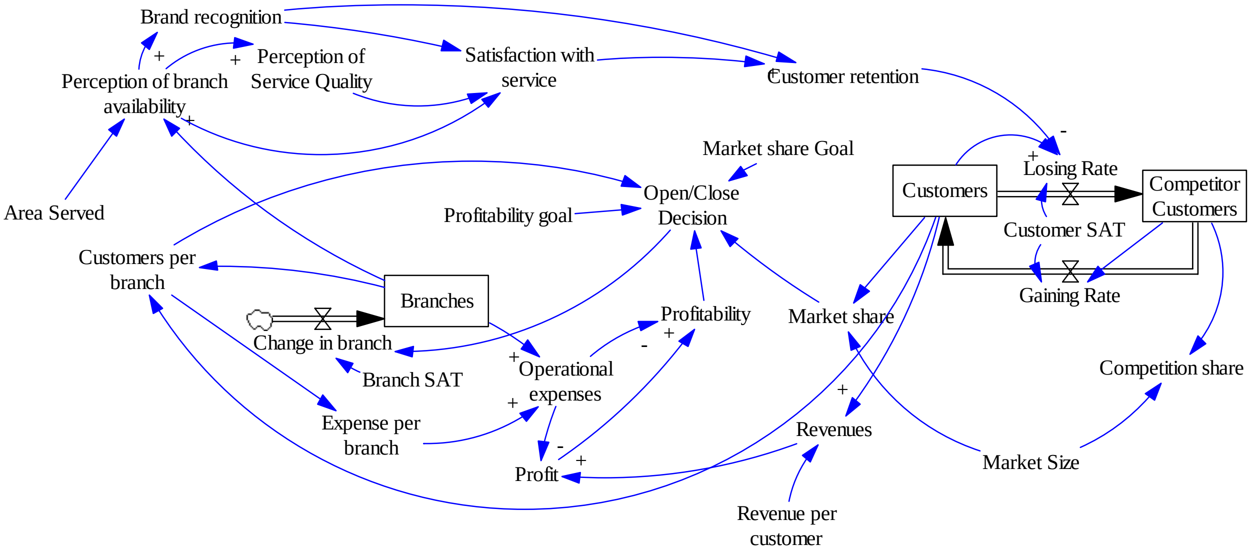
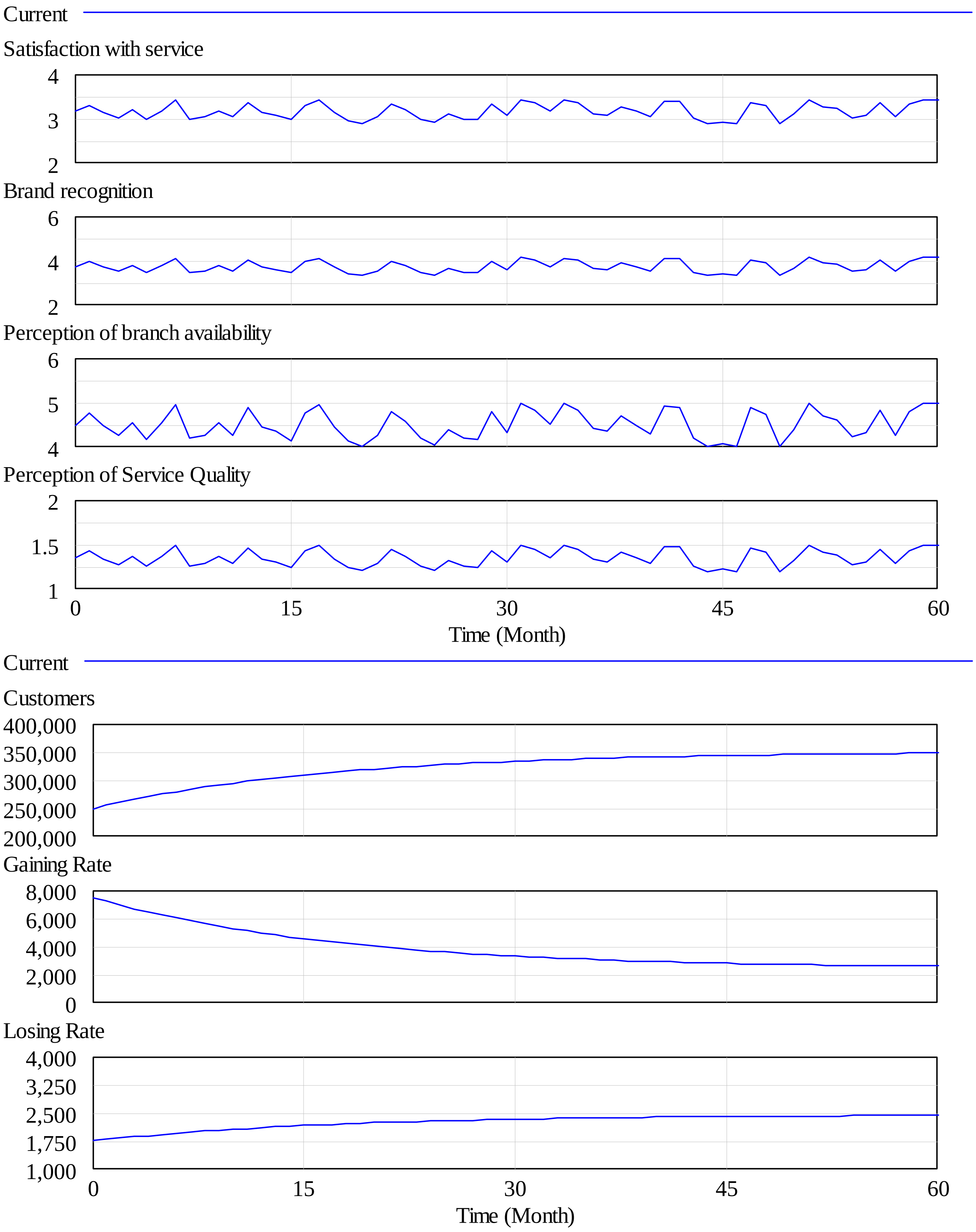


Figure 13. Behavior over time of key variables



Advantages of using Tree Pruning Methods in combination with SEM

The justification for using Tree Pruning Methods such as CART or CHAID arises primarily when knowledge needs to be extracted from large amounts of data sitting in large databases. The most likely behavior or outcome would then be uncovered regarding variable relationships. SEM can then be used to quantify the impact that a variation in the way respondents answered to constructs measured in survey items would affect the overall system.

In particular, using SEM as a resource for formulating relationships from survey data can prove to be advantageous.

- SEM can be used to either reinforce or challenge preconceived notions about relationships.
- SEM helps to draw associations between abstract concepts and constructs, which otherwise would have been close to impossible.

However, on the down side, there is a need for data and SEM applies linearity assumptions for each pair-wise relationship, which might bring misspecification problems. However, this can be overcome by exploring the fit of non-linear functions. Since a large data set is available, goodness of fit methods using the error term to compare the training set with the test set of data can be explored to adjust the equations.

Conclusions and further research

We have shown how tree data mining methods in conjunction with SEM can be used to explore and confirm relationships in large data sets when the nature, direction and intensity of the relationships among variables are unknown. The main application of the proposed three step process is for modeling problems where non-quantifiable concepts are used, such as the concept of customer satisfaction, or the construct service quality which in terms of data representation are characterized by several items in a survey.

In particular, the above three step process is currently being used to model the effect that proposed restructuring policies imposed purely based on financial performance will have on several variables representing customer perceptions, including customer satisfaction and customer loyalty. Eventually, customer retention will in turn affect revenue and sustainability of operations. Over 1 million customers answered a number of surveys for different branches of a service organization. Millions of data points and over 1000 variables are being explored and significant interactions are being identified. Eventually, a SD model will be run and validated within the company. Further research is needed for validation and non-linearity issues as well as on sensitivity analysis on the weight coefficients by introducing fuzzy mathematical concepts.

References

Bollen KA. 1989. *Structural Equations with Latent Variables*. John Wiley & Sons: New York, NY.

Breiman L, Friedman JH, Olshen RA, Stone CJ. 1983. *Classification and Regression Trees.* Wadsworth: Belmont, CA.

Forrester, JW. 1994. Policies, decisions and information sources for modeling, in *Modeling for Learning Organizations.* Morecroft J, Sterman J (eds), Productivity Press: Portland, OR, 51-84.

Galindo J, Tamayo P. 2000. Credit Risk Assessment Using Statistical and Machine Learning: Basic Methodology and Risk Modeling Applications. *Computational Economics*. Dordrecht: **15/1-2**: 107-143.

Hawkins DM, Kass GV. 1982. Automatic Interaction Detector. *Topics in Applied Multivariate Analysis.* Cambridge University Press: 269-302.

Hayduk LA. 1985. Personal Space: The Conceptual and Measurement Implications of Structural Equation Models. *Canadian Journal of Behavioural Science* **17**: 140-149.

Kass GV. 1980. An Exploratory Technique for Investigating Large Quantities of Categorical Data. *Applied Statistics* **29**: 119-127.

Luna-Reyes LF, Andersen DL. 2003. Collecting and analyzing qualitative data for system dynamics: methods and models. *System Dynamics Review*. **19/4**: 271-295.

MacLennan J, MacKenzie D. 2000. Strategic market segmentation: An opportunity to integrate medical and marketing activities. *International Journal of Medical Marketing* **1/1**: 40-52.

Morgan JN, Messenger RC. 1973. THAID – a sequential analysis program for the analysis of nominal scale dependent variables . Survey Research Centre, Institute for Social Research, University of Michigan.

Morgan JN, Sonquist JA. 1963. Problems in the analysis of survey data, and a proposal. *Journal of the American Statistical Association* **58**: 415-434.

Pugesek BH, Tomer A, Von Eye A. 2003. *Structural Equation Modeling*, Cambridge University Press: Cambridge, UK.

Ripley BD. 1996. *Pattern Recognition and Neural Networks.* Cambridge University Press: Cambridge, UK.

Rygielski C, Yen DC, Wang J. 2002. Customer relationship management in the network economy. *International Journal of Services Technology and Management* **3/3**: 297.

Schumacker RE, Lomax RG. 1996. *A Beginner's Guide to Structural Equation Modeling.* Lawrence Erlbaum: Mahwah, NJ.

Sterman JD. 2000. *Business Dynamics – Systems Thinking and Modeling for a Complex World.* Irwin McGraw-Hill: Boston, MA.

Wilkinson L. 1992. Tree Structured Data Analysis: AID, CHAID and CART. *Proceedings of the Sawtooth/SYSTAT Joint Software Conference.* Sun Valley, ID.