Utilização de árvores de decisão (CHAID) para alinhamento de atributos no desenvolvimento de novo produto

Manoel Silveira [a] (manoel@mat.ufrgs.br); Márcia Elisa Soares Echeveste [b] (echeveste@producao.ufrgs.br)
[a] Instituto de Matemática/Departamento de Estatística-UFRGS, RS – BRASIL
[b] Laboratório de Otimização de Produtos e Processos/GEDEPRO, Engenharia de Produção-UFRGS, RS – BRASIL

Resumo
Técnicas estatísticas são aplicáveis como suporte nas análises de informações que alimentam o Processo de Desenvolvimento de Produto. Nas fases iniciais auxiliam na segmentação da população e na determinação dos requisitos do produto identificando aqueles que agregam maior valor para o consumidor. O objetivo deste artigo é apresentar uma análise de árvore de decisão, inserido nas fases iniciais do Processo de Desenvolvimento de Produto por meio de uma aplicação prática para determinar os requisitos de um produto que estão associados a determinados segmentos de consumidores. Para tanto, é construída uma árvore de decisão que utiliza como critério o desmembramento de sucessivas tabelas cruzadas considerando os resultados obtidos da aplicação do teste estatístico qui-quadrado. Neste trabalho o método CHAID é aplicado a um caso que utiliza uma variável dependente, na qual os níveis representam dois segmentos populacionais (eco-orientado e não eco-orientado). Adotou-se como variáveis preditoras os requisitos de um produto com características sustentáveis. O resultado é a definição dos requisitos associados aos dois segmentos definidos.

Palavras-chave: CHAID; árvore de decisão; segmentação de mercado; requisitos do produto

1 Introdução

Em razão da crescente competição e das constantes mudanças nos padrões de consumo, as empresas têm a necessidade de desenvolver produtos com base em informações provenientes do ambiente mercadológico (YAMAN; SHAW, 1998). Uma forma de aprofundar o conhecimento sobre o mercado é encontrar quais são os segmentos com características similares e aprofundar os estudos em grupos específicos.

Para as empresas disporem de certa vantagem competitiva devem se adaptar às tendências de fragmentação do mercado, na identificação e atendimento a requisitos customizados a cada segmento. Segmentação de mercado tende a oferecer suporte aos negócios uma vez que sua detecção pode auxiliar no posicionamento quanto a promoções, atributos ou estratégias de serviços para seus clientes (CHEN, 2003).

Um exemplo de produto destinado a um público específico são os produtos eco-orientados. Entende-se por produto eco-orientado aquele que, desenvolvido de forma manual ou industrializada, não seja poluente, não seja tóxico, não acarrete prejuízos à saúde e ao meio-ambiente e ao mesmo tempo contribua para o desenvolvimento de um modelo social e economicamente sustentável (ARAUJO, 2009). Esse tipo de produto é preferencialmente utilizado por segmentos de consumidores que valorizam requisitos que minimizem prejuízos â natureza. Neste sentido, empresas que atendam princípios de sustentabilidade devem conscientizar o consumidor apresentando-lhe produtos atrativos a custos acessíveis.

O objetivo deste artigo é apresentar o método CHAID (*Chi-square Automatic Indentificator Detector*) inserido nas fases iniciais do Processo de Desenvolvimento de Produto por meio de uma aplicação prática para determinar os requisitos de um produto que estão associados a determinados segmentos de consumidores.

Esse artigo é organizado da seguinte forma: são apresentadas uma revisão teórica sobre o método CHAID, vantagens, limitações e validação do método. A seguir, é apresentado o método de pesquisa e uma estratégia de aplicação que servirá como um guia para uso do método. Posteriormente, é apresentada uma aplicação no desenvolvimento de um produto produzido com características de sustentabilidade. Finalmente, são feitas algumas considerações que encerram o trabalho..

2 Método CHAID

O CHAID (*Chi-Square Automatic Interaction Detection*) é um método utilizado para segmentação de uma população de interesse. Esta árvore é, geralmente, utilizada quando a segmentação é definida em termos de características demográficas ou variáveis categóricas com poder de predição (MAGIDSON, 1993).

Alguns aspectos a respeito dos dados coletados ou do universo no qual esses provêm devem ser considerados, como, por exemplo: (i) existe uma grande variedade de informações a respeito de cada indivíduo na pesquisa; (ii) na maior parte das vezes não é tratada diretamente a variável e sim a sua classificação; (iii) os dados são oriundos de uma amostra, geralmente coletada através de um delineamento experimental; (iv) muitas vezes os fatores exploratórios utilizados na análise podem estar correlacionados; (v) pode existir interação entre os efeitos; (vi) na realidade, existem propriedades lógicas e relação de causa e efeito entre as variáveis (MORGAN; SONQUIST, 1963). O método CHAID é baseado nos testes de associação qui-quadrado e particiona o conjunto de dados em subconjuntos mutuamente exclusivos que melhor descrevem a variável resposta exaustivamente (TURE et al., 2006).

2.1 **Procedimento**

O método CHAID opera em uma variável dependente de escala nominal ou ordinal e maximiza a significância da estatística qui-quadrado em cada partição, caracterizando o CHAID como uma estrutura de testes de significância (SPSS 18®). Devido aos sucessivos testes de comparações aplicados nessa técnica, é calculado um fator de correção na desigualdade de Bonferroni utilizado para obter-se um nível de significância ajustado.

A proposta de KASS (1980) é pesquisar por um $T^{(*)}_{(j)}$ (estatística qui-quadrado) máximo utilizando o método *stepwise*, avaliando a entrada de cada variável no modelo e verificando se sua contribuição é significativa ou não, entre as variáveis preditoras. A proposta pode ser resumida em 5 passos, como segue: (i) para cada preditor, fazer uma tabela cruzada das categorias do preditor com as categorias da variável dependente. (ii) encontrar os pares de categorias dos preditores (somente considerando pares determinados pelos diferentes tipos de preditores) para os quais 2xd tem diferença menos significativa. Caso essa significância não tenha um valor crítico alto, unir as duas categorias, e repetir esse passo; (iii) para cada categoria constituída a partir das três ou mais categorias originais, encontrar a partição binária mais significativa para os quais a mescla das categorias pode ser resolvida. Caso a significância esteja além de um valor crítico, implementar a divisão e repetir (ii); (iv) calcular a significância de cada preditor considerado e isolar o mais significante de todos. Caso a significância seja maior que um valor crítico, subdividir os dados de acordo com o número de categorias do preditor seguinte. Esse passo requer um teste de significância da tabela de contingência reduzida; (v) para cada partição dos dados que ainda não foi testada, retornar ao passo (i).

2.2 Vantagens e Limitações

Os resultados obtidos utilizando o CHAID são apresentados de forma gráfica sendo de fácil interpretação e leitura (HOARE, 2004). Uma importante consideração dos resultados do CHAID é que este pode ser usado para gerar escores individuais de probabilidade dos indivíduos da amostra pertencerem a determinado nódulo. Como os segmentos ou a resposta de interesse são definidos pelas combinações de variáveis preditoras, novos casos podem ser classificados para certo segmento pelos valores dessas variáveis, assim, as probabilidades para novos casos podem ser estimadas. Ainda, este

método pode fazer estimação para toda a população considerada ou somente parte dela (DIEPEN; FRANSES, 2005).

As desvantagens do método são que as variáveis independentes (preditoras) são consideradas de modo seqüencial e não simultâneo e que o CHAID não garante uma única solução ótima (PERREAUL e BARKSDALE, 1980). Ainda, Diepen e Franses (2005) indicam dois problemas em relação ao método: i) instabilidade da árvore CHAID, quando a árvore pode ajustar um conjunto de dados de maneira aceitável, mas se a tabela original de dados sofre alteração, uma nova árvore completamente diferente é criada; ii) *over-fitting*, ocorre quando a variância entre o valor médio gerado por um estimador e os valores observados é muito grande.

2.3 Validação

Três critérios podem ser considerados para validação do modelo adotado: (i) avaliação gráfica pela representação do ganho acumulado (*gain chart*). Este gráfico se caracteriza por ter forma de arco sobre uma reta diagonal. O eixo da abcissa do gráfico varia de 0 (zero) a 100%; (ii) risco estimado, indica o risco associado à classificação errada da categoria de referência da variável dependente; (iii) porcentagem de classificação correta que o modelo confere à categoria tomada como referência. Tanto para (ii) e (iii), os valores aceitáveis fazem parte das decisões a serem tomadas pelos pesquisadores levando em conta a categoria utilizada como referência na sua pesquisa.

3. Diretrizes de aplicação do CHAID

Os itens seguintes se referem à aplicação do método CHAID baseado na sequência de passos utilizados para o desenvolvimento deste trabalho. As diretrizes de aplicação podem resumir-se em oito passos: (i) definição do problema de pesquisa; (ii) caracterização amostral; (iii) determinação da variável dependente; (iv) determinação das variáveis preditoras; (v) avaliação descritiva das variáveis; (vi) representação gráfica do CHAID; (vii) avaliação da tabela do CHAID; (viii) representação gráfica do ganho. Como recurso computacional, foi utilizado o pacote estatístico SPSS 18®.
Na etapa de **definição do problema de pesquisa** para aplicação do método CHAID, o problema é determinado de maneira que evidencie e caracterize as variáveis necessárias ao modelo. Na sequência, procede-se a **caracterização amostral**. O tamanho da amostra tem que ser suficientemente grande para garantir a aplicação do teste estatístico qui-quadrado. Para a **determinação das variáveis do modelo** a única exigência é que sejam categóricas.

Uma das grandes vantagens da aplicação do método é que seu resultado pode ser interpretado através de uma **representação gráfica do CHAID** de fácil entendimento. A leitura da árvore é *bottom up*, e inicia no último nódulo subdividido chamado nódulo final e segue pelos seus nódulos precursores até chegar ao nódulo inicial. Na representação em tabela do CHAID, as colunas indicam as proporções de freqüências das categorias em cada nódulo, os percentuais da categoria referência e as demais categorias em relação à subdivisão que está sendo realizada e em relação ao total da amostra.

4. Aplicação no desenvolvimento produto limpeza eco-orientado

A aplicação do método CHAID é apresentada através do estudo realizado no desenvolvimento de um produto limpeza com características de sustentabilidade. Conforme mencionado, a apresentação do desenvolvimento é realizada de acordo com as etapas descritas na seção 3.

4.1 Definição do Problema de Pesquisa

O problema de pesquisa define-se na importância de identificar as características do consumidor voltado para a temática contemporânea relativa à preservação do meio ambiente no que diz respeito à aquisição, utilização e descarte de produtos. É necessário, então, buscar o entendimento de quais características referentes ao produto fariam um consumidor migrar para um produto eco-orientado. Desta maneira, os desenvolvedores de produtos podem produzir produtos com características mais atrativas ao mercado e ao mesmo tempo atender questões de sustentabilidade.

Para associar os requisitos do produto aos segmentos de interesse utilizou-se o método CHAID. Este método é capaz de fornecer suporte para responder a questão de pesquisa descrita identificando os requisitos que um produto de limpeza pode agregar para satisfazer as expectativas dos consumidores eco-orientados.

O instrumento de pesquisa utilizado foi elaborado a partir de cinco requisitos de um produto de limpeza denominados "certificação", "marca do produto", "praticidade", "rendimento" e "estabelecimento". Os requisitos foram definidos pela aplicação da técnica estatística Análise Fatorial sobre um conjunto que contava com 37 requisitos do produto. Os requisitos selecionados são aqueles negociáveis em relação ao produto (aqueles não obrigatórios, não normativos e que podem ser ajustados para atender a um segmento específico) com maior carga fatorial. Os requisitos certificação e estabelecimento foram inseridos no modelo devido ao interesse da equipe pesquisadora nos seus resultados. Os requisitos foram divididos em dois níveis, (+) representa a presença e (-) a ausência da característica.

A partir da combinação destes níveis foram construídos oito cenários ou perfis conforme Tabela 1. Os cenários foram apresentados aos respondentes que ordenaram os cenários de acordo com sua preferência.

Tabela 1. Cenários utilizados na pesquisa de preferência

Cenário	Marca do produto	Praticidade	Certificação	Estabelecimento	Rendimento	Valor
1	Marca Conhecida	Exige Preparo para uso	Selo Verde	Loja Física	Rendimento 30% Menor	R$7,85
2	Marca Conhecida	Pronto para uso	Sem Selo Verde	Via Internet	Rendimento Igual	R$3,80
3	Marca Conhecida	Pronto para uso	Selo Verde	Via Internet	Rendimento 30% Menor	R$7,15
4	Marca Não Conhecida	Exige Preparo para uso	Selo Verde	Via Internet	Rendimento Igual	R$6,30
5	Marca Não Conhecida	Pronto para uso	Selo Verde	Loja Física	Rendimento Igual	R$8,00
6	Marca Não Conhecida	Exige Preparo para uso	Sem Selo Verde	Via Internet	Rendimento 30% Menor	R$2,65
7	Marca Conhecida	Exige Preparo para uso	Sem Selo Verde	Loja Física	Rendimento Igual	R$4,50
8	Marca Não Conhecida	Pronto para uso	Sem Selo Verde	Loja Física	Rendimento 30% Menor	R$4,35

A cada cenário, estimou-se o valor monetário baseado no valor comercial comparativo do produto. Esta estimativa não representa necessariamente o valor real do mercado, mas sim a base para relativizar a escolha dos respondentes, apontando o preço que o consumidor estaria disposto a pagar pelo produto representado em determinado cenário.

4.2 Caracterização amostral

A amostragem foi realizada no período de 04/10/2010 a 30/10/2010. Considerou-se como consumidor eco-orientado aquele constituído pelos alunos da Entidade de ensino UNIPAZ-Sul, uma Instituição que tem uma proposta holística de atuar na educação, saúde, organizações e meio ambiente. O consumidor não eco-orientado é formado por consumidores comuns sem preocupações ambientais declaradas.

4.3 Determinação das variáveis do modelo

A variável dependente neste trabalho é denominada "segmento" e possui duas categorias que definem os segmentos em estudo que são os consumidores definidos como eco-orientados e não eco-orientados. A categoria de interesse (referência), neste estudo, é os consumidores considerados eco-

orientados. As variáveis preditoras, neste estudo, são os requisitos do produto (certificação, marca do produto, praticidade, rendimento e estabelecimento).

4.4 Avaliação descritiva dos dados

A Figura 1 apresenta os cenários 5, 4 e 3 com maior frequência de escolha com 55,60%, 20,00% e 15,60% respectivamente. Em geral, o consumidor não considera importante o fato do produto ser oferecido por uma marca reconhecida no mercado.

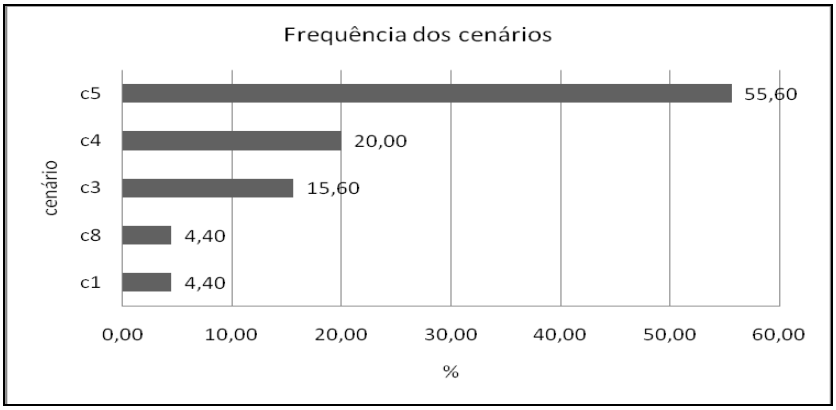

Figura 1. Cenários com maior preferência entre os entrevistados

Assim, houve predisposição para pagar um valor maior pelo produto quando este apresenta requisitos como possuir um selo verde que o certifique como sustentável, esteja pronto para uso, tenha rendimento igual a outro produto oferecido no mercado e a compra ser feita em loja física.

4.5 Representação gráfica do CHAID

O nódulo zero traz um resumo de toda a amostra em relação à variável dependente. A partir das categorias deste nódulo, as variáveis preditoras são testadas através de tabelas cruzadas seguindo a metodologia do CHAID. O diagrama do CHAID é apresentado na Figura 2.

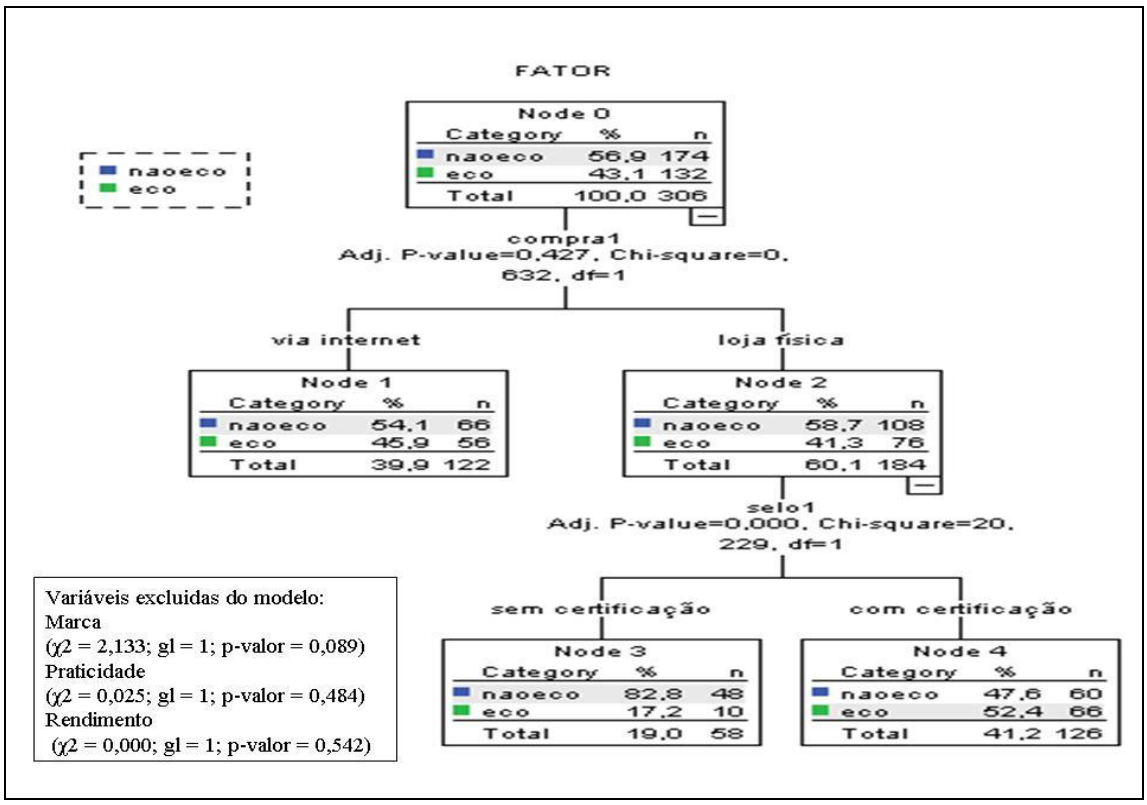

Figura 2 - Representação do diagrama CHAID

Utilizando o método CHAID, evidenciou-se que os requisitos estabelecimento e certificação são os mais relevantes para o produto. A variável independente, estabelecimento, foi inserida no modelo por ser uma variável de interesse de investigação para desenvolvimento de trabalhos futuros. Dessa maneira, pela Figura 2 observa-se que o nódulo quatro está associado aos entrevistados declarados eco-orientados e esses, por sua vez, evidenciam preferência por um produto que apresente um certificado de garantia quanto ao seu caráter ecológico e ainda preferem fazer a compra do produto em loja física.

4.6 Representação em tabela do CHAID

O diagrama de árvore pode ser representado através de uma tabela com o resumo dos resultados. Nessa tabela são apresentadas as informações relevantes disponíveis no diagrama CHAID. Baseada na árvore de decisão, a tabela apresenta para cada nódulo, a categoria de maior freqüência e sua porcentagem, conforme a Tabela 2, evidenciando a categoria predita em cada nódulo.

Tabela 2. Tabela da árvore CHAID - I

	N	Porcentagem não-eco	N	Porcentagem eco	N	Porcentagem total	categoria predita	nódulos prévios
0	174	56,9%	132	43,1%	306	100,0%	não-eco	
1	66	54,1%	56	45,9%	122	39,9%	não-eco	0
2	108	58,7%	76	41,3%	184	60,1%	não-eco	0
3	48	82,8%	10	17,2%	58	19,0%	não-eco	2
4	60	47,6%	66	52,4%	126	41,2%	eco	2

A categoria predita é aquela que apresenta mais de 50% da frequência entre as duas categorias da variável dependente no nódulo final. O nódulo 4, por exemplo, apresenta os consumidores eco-orientados como categoria predita. Para este nódulo, a coluna eco (N) indica uma frequência de 66 casos que representa 52,4% da frequência total do nódulo. O nódulo 4 possui no total 126 casos que

representam 41% do número total dos casos analisados. A leitura o nódulo 4 está associado a consumidores eco-orientados que têm preferência de compra por produtos que apresentem selo de certificação e que o local de compra do produto seja realizada em loja física.

Na tabela de classificação, Tabela 3, as linhas correspondem às categorias observadas pelos respondentes e as colunas representam as categorias preditas utilizando o modelo CHAID. O modelo apontou aproximadamente 58,8% de exatidão total para classificar corretamente os entrevistados em relação à sua condição de ser eco-orientado ou não eco-orientado.

Tabela 3 - Tabela de classificação

	Não eco-orientado	Eco- orientado	Porcentagem correta
não eco-orientado	114	60	65,5%
eco-orientado	66	66	50,0%
Porcentagem total	58,8%	41,2%	58,8%

O valor 0,412 indica um risco de que o critério adotado para caracterizar os grupos de consumidores eco-orientados e não eco-orientados pode não ter sido suficientemente discriminatório para detectar esta distinção. A tabela de classificação merece uma consideração: para os consumidores considerados eco-orientados é predito corretamente aproximadamente 50,00% dos casos. Isto leva a considerar que os critérios utilizados para definir os segmentos eco-orientados e não-eco-orientados não foram suficientemente discriminatórios. Desta forma, em muitos casos seus comportamentos acabam sendo tão similares que as diferenças não são captadas por testes aplicados.

5. Considerações finais

O objetivo desse artigo é apresentar o método CHAID (*Chi-square Automatic Indentificator Detector*) inserido nas fases iniciais do Processo de Desenvolvimento de Produto, teoricamente e fazer uma aplicação prática para determinar os requisitos de um produto que estão associados a determinados segmentos de consumidores. Assim, é possível evidenciar duas considerações finais importantes: (i) o método foi abordado e aplicado de maneira objetiva para que desenvolvedores de produtos possam ter mais esta opção de análise no auxílio para tomada de decisões; (ii) baseado no caso de desenvolvimento de um produto de limpeza o método é fácil de ser reproduzido nas fases iniciais de desenvolvimento de qualquer produto manufaturável.

Os resultados foram satisfatórios porque foi possível identificar grupos e requisitos do produto específicos que analisados pelo pesquisador podem auxiliá-lo a determinar onde seus esforços devem ser concentrados. Este método pode contribuir como mais um recurso para tomada de decisões no momento de definição dos requisitos de um produto nas fases iniciais do PDP. Para um futuro trabalho, uma sugestão é agregar mais requisitos ao produto e mais níveis a estes requisitos, para que a análise de trade-off tenha um caráter mais discriminatório nas opções dos respondentes.

Quanto aos segmentos estudados, avalia-se que as empresas que desenvolvem produtos ecológicos podem aplicar programas especiais para que seus produtos tenham maior aceitação. A equipe de PDP pode realizar levantamento criterioso sobre requisitos com caráter de sustentabilidade que podem ser agregados ao seu produto. A empresa desenvolvedora do produto pode traçar estratégias de vendas específicas para segmentos definidos, ressaltando os aspectos do produto de modo fidelizar o segmento que utiliza seu produto ou conquistar outro novo segmento. Ainda, sugere-se aplicar um método de classificação mais discriminatório entre consumidores eco-orientado e não eco-orientados.

Referências

ARAÚJO M. A. Instituto para o Desenvolvimento da Habitação Ecológica - IDHEA. Disponível em <http://www.idhea.com.br>. Acesso em 20/12/2009.

CHEN J.S. Market Segmetation by Tourist´s Sentiments. *Annals of Tourism Research*. vol. 30. n. 1. pp. 178-193, 2003.

HOARE R. Using CHAID for Classification Problems. *New Zealand Statistical* Association *Conference*. 2004.

KASS G.V. An Exploratory Technique for Investigating Large Quantiles of Categorical Date. University of the Witwatersand. *Appl. Statist*. 29, n. 2, pp. 119-127, 1980.

MAGIDSON J. The Use of the Neu Ordinal Algorithm in CHAID to Target Profitable Segments. *The Journal of Database Marketing*. vol 1. pp 29-48, 1993.

MAGIDSON J. SPSS® for Windows™ CHAID Release 6.0. SPSS Inc. Chicago, 1993.

MORGAN J.N.; SONQUIST J. A. Problems in the Analisys of Survey Date: and a proposal. *Journal of the American Statistical Association*. Vol. 58. N 302. pp .415-434, 1963.

PERREAUL W. D.; BARKSDALE H. C. A Model-Free Approch for Analysis of Complex Contingency Date in Survey Research. Journal of Marketing Research. 17 (4) 503-515, 1980.

TURE M.; TOKATLI F.; KURT I. Using Kaplan-Meier Analysis together with Decision Tree Methods (C&RT, CHAID, QUEST, C4.5 and ID3) in Determining Recurrence-free Survival of Breast Cancer Patients. Science Direct. Expert System with Applications 36. 2017 – 2006, 2009.

YAMAN H.; SHAW R. The Conduct of Marketing Research in Tourism. *Journal of Travel Research*. 36(4):25-32, 1998.