MODELAGEM DE DADOS CATEGÓRICOS ATRAVÉS DO PROCEDIMENTO DE CHI-SQUARE AUTOMATION INTERACTION DETECTION (CHAID) E ANÁLISE DE VARIÂNCIA – UMA APLICAÇÃO NO SETOR IMOBILIÁRIO

Ana Augusta F.de Freitas, MsC
Programa de Pós-Graduação em Engenharia de Produção - UFSC
E-mail: augusta@eps.ufsc.br
Maria Carolina Gomes de Oliveira, MsC
Pós-Graduação em Engenharia de Produção - UFSC
E-mail: mcgo@eps.ufsc.br
Luiz Fernando Mählmann Heineck, PhD
Programa de Pós-Graduação em Engenharia de Produção – UFSC
Caixa Postal 476, Campus Trindade, 88010-970, Florianópolis - SC

ABSTRACT: This research works explores the use of two different statistical techniques for the explanation of the price potential housebuilding clients are willing to pay for their homes. Data was obtained through 3000 direct interviews in eleven Brazilian cities. The interviews were performed during building sales fairs held in those cities. The questionnaires were structured with more than 100 questions dealing with the preference of homebuyers in terms of specific housing attributes and the households socio-economic characteristics. The relative importance of these attributes was obtained by preprocessing the data with CHAID- Chi-Square Automatic Interaction Detection, a technique that enables the user to get an insight on which variables to include in generalized linear models. Among these models, the analyses of variance technique was used in order to handle categorical data and the interaction between variables. Results show the importance of taking into consideration interregional differences in price determination and the overwhelming influence of monthly income as the major determinant of desired price.

KEYWORDS: housing attributes, CHAID technique, analyses of variance

RESUMO: Este trabalho tem como objetivo analisar dados de preferências habitacionais, através de duas técnicas complementares. O banco de dados utilizado na aplicação consta de cerca de 3000

entrevistas com clientes potenciais do mercado imobiliário. As entrevistas foram conduzidas em onze diferentes cidades do Brasil, durante a realização de feiras de imóveis. Um exemplo é usado para mostrar a importância da consideração das diferenças regionais na quantificação do preço do imóvel procurado, bem como a exata definição do valor da renda mensal familiar. As variáveis iniciais foram pré-processadas através da técnicas de CHAID e sua relação numérica com o preço quantificada através dos modelos de análise de variância.

1. INTRODUÇÃO

Em um mercado livre em competição tem sido cada vez mais importante estudar as características quantitativas e qualitativas do consumidor, identificar nichos de mercado de produtos específicos e descobrir oportunidades de negócios. Isto é particularmente verdade para a indústria da construção civil, que agora depende quase que exclusivamente dos clientes como principal fonte de financiamento direto das suas atividades. Com base nisto, torna-se imprescindível para os empresários e responsáveis pelo setor o conhecimento do mercado e das variáveis que influenciam a decisão de comprar um imóvel.

Para estudar as características envolvidas no processo de escolha da compra de um imóvel, várias técnicas foram pesquisadas e aplicadas dentro da área de escolha e mobilidade residencial. CLARK et al. (1988) sumarizam o conjunto destas técnicas e ilustram a sua utilização na análise da escolha habitacional.

Como regra geral, os pesquisadores demonstram o relacionamento entre as características sócio-econômicas e demográficas das famílias e a escolha da habitação (com respeito ao tipo, localização e preço). Nesta linha encontram-se os trabalhos de DEURLOO et al. (1990), CLARK et al. (1994) e BOOSTMA (1995).

Além das variáveis tradicionais como idade e renda, algumas variáveis aparecem para fazer parte do modelo. Entre elas, o modelo poderá incluir composição familiar, participação no mercado de trabalho, número de pessoas que contribuem com a formação da renda familiar e as características regionais ligadas ao tipo de mercado habitacional onde os dados foram coletados. Para esta última variável, DEURLOO (1987) alerta para a necessidade de considerá-la, especialmente quando estamos trabalhando com mercados cujas restrições na oferta são diferenciadas.

Este tipo de pesquisa no Brasil ainda não é comum, aparecendo nos últimos anos trabalhos pioneiros como o de FREITAS (1995) e OLIVEIRA (1998), que utilizaram dados de pesquisas de mercados imobiliários a fim de explicar o comportamento de mudança e escolha de uma nova habitação.No entanto, ainda se procura a melhor maneira de analisar as inúmeras variáveis envolvidas neste processo de decisão. Na maioria dos casos, as preferências sobre atributos da habitação e as características do consumidor são variáveis de natureza categórica e em grande número, o que restringe as técnicas possíveis de aplicação. FREITAS (1995), por exemplo, mostra que se quisermos trabalhar com modelos Logit é necessário uma redução drástica no número de variáveis e suas categorias.

Com o objetivo de diminuir a complexidade deste problema, o trabalho tem como objetivo testar o uso de duas técnicas combinadas (CHAID e Análise de Variância), propícias à análise de dados categóricos e menos rigorosa em termos de restrições estatísticas.

A fim de exemplificar a sua aplicação, dados de pesquisas de mercado serão utilizados para explicar a relação entre as variáveis sócio-econômicas e o preço do imóvel desejado, considerando as características do mercado local.

2. COLETA DE DADOS

Cerca de 3 mil clientes potenciais foram entrevistados em onze diferentes cidades do país: Belém, Recife, Natal, Vitória, Blumenau, Florianópolis, Porto Alegre, Caxias, Pelotas, Santa Maria e Passo Fundo. Nas últimas quatro cidades, os dados foram cedidos pelos responsáveis pelas pesquisas nestes locais.

Os questionários seguiram uma estrutura similar e eram divididos em quatro partes. A primeira era formada por perguntas relacionadas às características sócio-econômicas do indivíduo (estado civil, número de filhos, idade, condição de posse do imóvel atual, renda mensal, valor dos bens disponíveis para colocar no negócio).

A segunda parte abordava questões relativas às macro-variáveis do imóvel (número de quartos, garagens, suítes, localização) e condições de pagamento (preço do imóvel procurado, valor da prestação, valor da entrada e prazo).

A terceira parte do questionário analisava a disponibilidade por pagar a mais por vários atributos residenciais, classificados dentro das seguintes categorias:

1. área privativa (por exemplo: closet, lavabo, despensa);

2. área de lazer (por exemplo: piscina, quadra, quadra);
3. qualidade do imóvel (por exemplo: número de apartamentos por andar, número de blocos, sofisticação da fachada) e
4. equipamentos (por exemplo: bancada de granito, aterramento, box blindex nos banheiros).

A última parte testava a força de alguns atributos através de questões onde o entrevistado era submetido a avaliar a troca entre possibilidades de projeto (por exemplo: sala maior sem varanda ou sala menor com varanda; ou sala maior ou cozinha com espaço para mesa de refeições).

O presente trabalho utiliza dados relativos à primeira e segunda parte do questionário, tratando do preço expresso pelo cliente para a compra do imóvel e de algumas características sócio-econômicas do entrevistado.

As entrevistas foram realizadas em Salões de Imóveis (Belém, Recife, Natal, Vitória, Blumenau, Florianópolis, Caxias, Porto Alegre) ou Feiras de Exposição (Passo Fundo, Santa Maria e Pelotas) e os entrevistados eram convidados a responder o questionário caso expressassem o desejo de comprar um imóvel dentro dos próximos anos.

Em algumas cidades as pessoas eram chamadas indistintamente a participar da pesquisa. Nestes casos, apenas os questionários que tinham resposta afirmativa à questão sobre a pretensão de compra de um imóvel dentro dos próximos anos foram utilizados.

Os dados foram coletados entre abril de 1995 e março de 1998. Nas cidades de Belém, Florianópolis e Santa Maria as pesquisas foram feitas mais de uma vez em anos distintos.

3. FUNDAMENTAÇÃO TEÓRICA DA ANÁLISE DE CHAID

O procedimento original de automatic interaction detection (AID), desenvolvido por Sonquist e Morgan (1964), tem a sua origem na análise de variância. Por esta técnica, assume-se a utilização de uma variável dependente contínua e variáveis independentes qualitativas (ou categorizadas), onde através de um procedimento em cascata, divide-se o conjunto de variáveis em dois subgrupos, através da maximização da soma dos quadrados entre subconjuntos.

Esta técnica foi expandida para os casos onde a variável dependente é qualitativa, como propõe KASS (1980), e é conhecida como Chi-quare automatic interaction detection (CHAID). No caso do CHAID, os dados são divididos, a cada passo, em grupos otimizados (e não necessariamente em dois subgrupos), através da maximização da significância da estatística do Chi-quadrado. Alguns

autores fizeram uso desta técnica em pesquisas de mercados em geral (PERREAULT e BARKSDALE, 1980), mas raros exemplos podem ser encontrados para o caso do mercado habitacional (DEURLOO, 1988).

No caso do CHAID, as categorias das variáveis independentes são agregadas se elas mostrarem padrões de comportamento semelhantes em relação à variável dependente. Além disto, para cada uma das categorias das variáveis independentes selecionadas, a técnica escolhe a próxima variável que melhor prediz a categoria da variável anterior. Ao final, os resultados da análise são mostrados em forma de uma árvore (chamada na literatura de dendograma), onde as variáveis independentes aparecem de acordo com a capacidade de prever níveis específicos de outras variáveis independentes. O resultado final do CHAID mostra segmentos da população, que diferem segundo um determinado critério.

Os segmentos derivados do CHAID são mutuamente exclusivos e exaustivos. Isto significa que eles não se sobrepõem e cada indivíduo está contido em apenas um segmento. Além disto, pelo fato de serem definidos através de combinações de variáveis independentes, pode-se facilmente classificar cada caso dentro de um segmento.

O objetivo principal do CHAID é encontrar as principais interações entre grupos de pessoas e escolha habitacional em grandes tabelas de cross-tabulations (no caso, uma variável dependente e cinco preditores) e prover uma descrição parcimoniosa sobre o conjunto de dados.

No entanto, os resultados podem ser usados ainda para reduzir as dimensões dos problemas de modelagem (quando, por exemplo, trabalha-se com modelos Logit), através da redução do número de categorias e de variáveis (CLARK, 1991).

Em especial, o tipo de resposta obtida através da análise do dendograma é propícia para o uso de técnicas em redes (*nested approach)*, como o Nested Multinomial Logit Models, como descrito por este último autor.

Com será visto a seguir, a variável renda é o principal elemento de explicação do preço desejado, no entanto o valor dos bens influencia classes específicas de renda, enquanto o estado civil influencia outras. Ou seja, diferentes variáveis atuam em distintos níveis da estrutura em rede.

Além disto, se ao invés de aplicar modelos mais complexos e de maior exigência computacional como os Logits, estivermos interessados apenas em conhecer o relacionamento de cada variável

independente com a variável dependente, pode-se ainda utilizar os resultados do CHAID em modelos lineares gerais, como os modelos de análise de variância.

4. ANÁLISE DOS DADOS

Para exemplificar o funcionamento da técnica, o preço do imóvel desejado foi considerado a variável dependente e será explicado em função das características sócio-econômicas do indivíduo e das diferenças regionais.

Neste último caso, as cidades foram agregadas de acordo com o nível de riqueza da população. Neste caso, assumiu-se uma classificação conforme mostra a tabela abaixo, baseada no conhecimento prévio dos pesquisadores sobre os mercados locais. No entanto. uma classificação mais rigorosa deve ainda ser testada, considerando por exemplo dados coletados por institutos de pesquisas demográficas (IBGE) ou ainda dados sobre oferta de imóveis em cada cidade e preços de área construída praticados. Estes últimos dados algumas vezes são coletados por Sindicatos de Construção locais.

As variáveis envolvidas na análise foram escolhidas preliminarmente com base na literatura nacional e internacional que as aponta como importantes dentro do processo de decisão de escolha do imóvel.

A variável preço foi tomada como exemplo de variável dependente, embora alternativamente possa-se usar características físicas do apartamento, como número de quartos.

O conjunto total de pessoas entrevistadas nas feiras de imóveis reduziu-se neste exemplo de 2764 casos para 2344. Cada um destes estará alocado em uma das ramificações da análise do CHAID. Em alguns casos, não foi possível obter informações para as cinco variáveis independentes, o que caracteriza a existência de pontos faltantes (*missing values*). Para fins de simplificação da visualizacão, os *missing values* relativos às variáveis independentes foram retirados do dendograma, quando eles apareciam em uma categoria isolada. Onde esta categoria (*missing values)* teve semelhança com outras categorias, as mesmas foram automaticamente agregadas pelo CHAID e estão representadas por um ponto (por exemplo: bens 34.).

A Tabela 1, abaixo, mostra as categorias das variáveis envolvidas na análise.

Variáveis	Categorias
Preço do imóvel desejado (variável dependente)	1. Até R$ 42.500 2. De R$ 42.500 à R$ 55.000 3. De R$ 55.000 à R$ 75.000 4. De R$ 75.000 à R$ 120.000 5. Mais de R$ 120.00
Renda mensal familiar	6. Ate R$ 1.000 7. De R$ 1.000 à R$ 2.000 8. De R$ 2.000 à R$ 3.000 9. De R$ 3.000 à R$ 4.000 10. De R$ 4.000 à R$ 5.000 11. Mais de R$ 5.000
Valor dos bens	12. Até R$ 13.000 13. De R$ 13.000 a R$ 27.000 14. De R$ 27.000 a R$ 41.000 15. De R$ 41.000 a R$ 78.000 16. Mais de R$ 78.000
Condição de Posse do Imóvel	17. Próprio 18. Alugado 19. Outros
Tamanho da família	20. De 1 a 2 pessoas 21. De 3 a 4 pessoas 22. De 5 a 6 pessoas 23. Mais de 7 pessoas
Idade	24. Até 25 anos 25. De 26 a 35 anos 26. De 36 a 45 anos 27. Mais de 45 anos
Região	28. Fraca em termos econômicos (Belém, Passo Fundo, Pelotas, Santa Maria) 29. Moderada em termos econômicos (Natal, Blumenau, Recife, Vitória) 30. Forte em termos econômicos (Florianópolis, Caxias, Porto Alegre)

Tabela 1 – Variáveis dependente e independentes e suas categorias

As categorias das variáveis independentes são representadas no dendograma (Figura 1) pelo mesmo número que foram codificadas na Tabela 1.

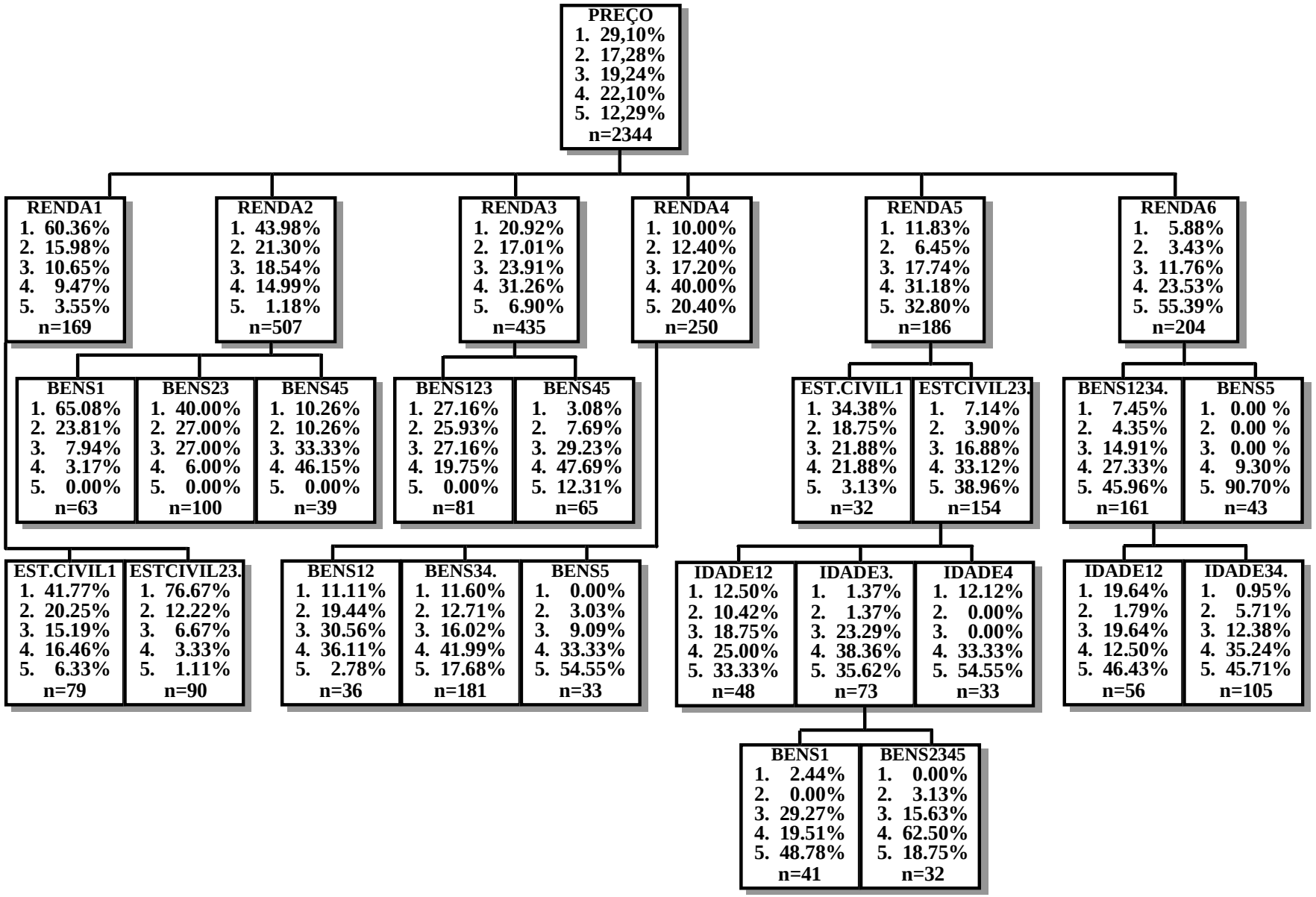

Figura 1 – CHAID Dendograma. Os valores nos quadros correspondem ao percentual em cada categoria de preço

Pela análise do dendograma é possível concluir alguns itens importantes:

1. A variável mais importante na predição do preço desejado é renda mensal familiar. O fato de não ter sido possível agregar nenhum nível de renda mostra a tamanha importância da definição desta variável, já que cada classe acima referenciada comporta-se de maneira estatisticamente diferente das outras;
2. Para a categoria de renda mais baixa (até R$ 1.000) a grande maioria dos clientes (60%) escolhe imóveis até RS 42.500, como era de se prever. No entanto, em regiões ricas a média do preço do imóvel desejado para a compra é maior;
3. Para indivíduos com renda mensal familiar entre R$ 1.000 e R$ 3.000, a principal característica diferenciadora é o mercado onde os mesmos se encontram. Nos três casos, quanto mais rica a região, maior a propensão por pagar a mais pelo futuro imóvel. Nestas duas classes de renda, aparece também a condição do posse do imóvel atual como variável importante. Aqueles que já possuem imóvel próprio pagam a mais pelo futuro imóvel, reflexo provável da disposição de colocar este imóvel no negócio usando, por exemplo, para amortizar uma parcela do pagamento inicial;
4. Para os clientes com renda acima R$ 3.000, as características do mercado local diminui em importância, surgindo os aspectos mais demográficos como idade e composição familiar. Em ambos os casos o aumento da idade e do tamanho da família é acompanhado pela aumento do preço a pagar no imóvel.

Em linhas gerais, pode-se perceber que todas as variáveis aparecem em algum momento do processo, embora diferenciada dentro de cada segmento de mercado. Nestes casos, pode-se escolher em prosseguir o estudo com modelos que expliquem o comportamento de cada faixa de mercado (envolvendo apenas as variáveis importantes para determinada faixa) ou construir um modelo geral envolvendo as variáveis que apareceram no dendograma com mais freqüência.

No caso em questão, resolveu-se por utilizar um modelo único, aplicando os modelos de análise de variância.

5. UTILIZAÇÃO DOS RESULTADOS DO CHAID EM MODELOS DE ANÁLISE DE VARIÂNCIA

A análise geral dos resultados do dendograma nos permite gerar algumas equações através dos modelos de análise de variância em relação a quais variáveis são importantes e em que nível de agregação elas devem ser analisadas. Com base nisto, um modelo único será desenvolvido a fim de quantificar a relação entre as variáveis independentes escolhidas e o preço do imóvel desejado. Neste caso, o valor do imóvel desejado foi tomado na sua forma contínua para facilitar a interpretação dos resultados.

A Tabela 2 mostra as variáveis (e categorias) escolhidas para fazer parte do modelo final.

Variáveis	Categorias
Renda familiar	1. 1. Ate R$ 1.000; 2. De R$ 1.000 à R$ 2.000; 3. De R$ 2.000 à R$ 3.000; 4. De R$ 3.000 à R$ 4.000; 5. De R$ 4.000 à R$ 5.000; 6. Mais de R$ 5.000
Região	7. Fraca ou moderada em termos econômicos; 8. Forte em termos econômicos
Idade	9. Até 35 anos; 10. Mais de 35 anos
Tamanho da família	11. Até duas pessoas; 12. Mais de 2 pessoas
Valor dos bens	13. Até R$13.000; 14. Mais de R$13.000

Tabela 2 - Variáveis escolhidas para o modelo final

Inicialmente apenas um modelo considerando o efeito principal das variáveis foi considerado. Uma análise dos efeitos de interação constatou existir relação entre a renda e a região e entre a renda e o tamanho da família. Estas interações foram então consideradas também no modelo final e respondem pela não linearidade do modelo.

A Tabela 3 mostra as estimativas dos parâmetros. Apenas aquelas com nível de significância acima de 0.05% são mostradas. Para cada variável existe uma categoria chamada base, que serve como o

nível de comparação, quando se usa os parâmetros. No caso, por exemplo, da variável renda, a categoria 6 é a base e por isto não aparece na tabela pois seu efeito seria redundante.

Parâmetro	**B**	**Sig**
Intercepto	193893	0.000
Renda 1	- 143.782	0.000
Renda 2	-133.162	0.000
Renda 3	- 109.821	0.000
Renda 4	-81.083	0.000
Renda 5	-81.694	0.000
Idade 1	-5.332	0.060
Tamanho da família 1	-46.483	0.000
Valor dos Bens 1	-17.462	0.000
Região 1	-58385	0.000
(Renda 1) (Região 1)	54.213	0.000
(Renda 2) (Região 1)	59.566	0.000
(Renda 3) (Região 1)	42.179	0.000
(Renda 4) (Região 1)	44.897	0.000
(Renda 5) (Região 1)	46.853	0.000
(Renda 1) (Tamanho 1)	50.123	0.000
(Renda 2) (Tamanho 1)	43.548	0.000
(Renda 3) (Tamanho 1)	45.422	0.000
(Renda 4) (Tamanho 1)	27.508	0.007
(Renda 5) (Tamanho 1)	28.215	0.010

Tabela 3 – Parâmetros do modelo

Algumas conclusões podem ser traçadas de acordo com o valor dos parâmetros:

1. valor do intercepto pode ser interpretado como o valor médio que os clientes com renda mensal maior que R$5.000, com idade maior que 36 anos, famílias com mais de 2 pessoas, valor dos bens acima de R$13.00, vivendo em regiões ricas estariam dispostas a pagar. Neste caso, em torno de R$ 194.000;
2. Com o decréscimo da renda (*ceteris paribus* as outras variáveis) a disponibilidade de pagamento pode cair até 75%;
3. A diminuição na idade não leva a decréscimo tão elevados, apenas na ordem de R$5.000;

4. As famílias menores pagariam cerca de R$46.000 a menos pelo novo imóvel, reflexo natural da não necessidade de muito espaço;
5. As diferenças regionais podem diminuir o valor do imóvel em torno de 30%;
6. Os sinais positivos dos coeficientes de interação acontecem devido a um balanceamento geral da equação. Sendo assim, pessoas de renda baixa morando em regiões mais fracas economicamente estariam dispostas a pagar em média R$ 46.000 (resultado da equação: 193.893 - 143.782 - 58.385 + 54.213).

6. CONCLUSÕES

A análise dos resultados obtidos pela aplicação do CHAID mostrou que a principal variável influenciadora do preço que os indivíduos desejam pagar pelo novo imóvel é a renda familiar. Em geral, a região onde o indivíduo se encontra é também uma variável muito importante, principalmente para as categorias de baixa renda. Com base nestes resultados, conclui-se que a técnica do CHAID mostrou-se um bom método para descobrir estruturas principais nas tabulações cruzadas multidimensionais.

Como etapa preliminar da análise de dados, o método CHAID também propiciou a definição de variáveis importantes a serem usadas em modelos de análise de variância. Na análise de variância, obtém-se diferenças de valores em função das várias variáveis independentes, enquanto que para o CHAID indicavam-se apenas as percentagens escolhidas dentro de cada categoria.

Aproveitando-se das características do CHAID, abre-se o caminho para a utilização de técnicas mais sofisticadas como os modelos logits, sugeridos na literatura. A utilização de uma técnica que ajude a diminuir o número e as categorias das variáveis seria de enorme ajuda neste tipo de modelos, os quais possuem como principal desvantagem as restrições impostas em relação ao número de variáveis a serem utilizadas. No entanto, deve-se verificar a verdadeira necessidade de se escolher modelos que exijam perdas consideráveis no número de variáveis, já que os modelos de análise de variância mostraram-se suficientemente adequados para fins de quantificação da relação entre variáveis independentes e dependentes, sejam elas categóricas ou contínuas.

7. REFERÊNCIAS BIBLIOGRÁFICAS

BOEHM. T. P. A Hierarchical Model of Housing Choice. *Urban Studies*. 1982, Vol. 19, p.17-31.

BOOSTMA, H. G. The Influence of a Work-Oriented Life Style on Residential Location Choice of Couples. *Netherlands Journal of Housing and the Built Environment*. 1995, Vol. 10, N^{o} 1, p.45-63.

CLARK, W. A. V.; DEURLOO, M. C.; DIELEMAN, F. M. Modeling Strategies for Categorical Data: Examples for Housing and Tenure Choice. *Geographical Analysis*. 1988, Vol. 20, p.198-219.

CLARK, W. A. V.; DEURLOO, M. C.; DIELEMAN, F. M. Categorical Data with Chi Square Automatic Interaction Detection and Correpondence Analysis. *Geographical Analysis*. 1991, Vol. 23, p.332-345.

CLARK, W. A. V.; DEURLOO, M. C.; DIELEMAN, F. M. Tenure Changes in the Context of the Micro-Level Family and the Macro-Level Economic Shifts. *Urban Studies*. 1994, Vol. 31, Nº 1, p.137-154.

DEURLOO, M. C. *A Multivariate Analysis of Residential Mobility*. Tese de Doutorado, Istituut voor Sociale Geografie, Universiteit van Amsterdam, 1987, 210 p.

DEURLOO, M. C.; DIELEMAN, F. M.; CLARK, W. A.V. Multinomial Response Models of Housing Choice. *Environment and Planning A*. 1988, Vol. 19.

DEURLOO, M. C.; CLARK, W. A. V.; DIELEMAN, F. M. Choice of Residential Environment in the Randstad. *Urban Studies*. 1990, Vol. 27, Nº 3, p.335-351.

FISCHER, M. M.; AUFHAUSER, E. Housing Choice in a Regulated Market: A Nested Multinomial Logit Analysis. *Geographical Analysis.* 1988, Vol. 29, p.47-69.

FREITAS, A. A. F. *Modelagem Comportamental dos Decisores Através de Técnicas de Preferência Declarada: Uma Aplicação no Setor Imobiliário de Florianópolis-SC*. Dissertação de Mestrado, Florianópolis, Programa de Pós-Graduação em Engenharia de Produção, UFSC, 1995.

OLIVEIRA, M. C. G. *Os Fatores Determinantes da Satisfação Pós-Ocupacional de Usuários de Ambientes Residenciais*. Dissertação de Mestrado, Florianópolis, Programa de Pós-Graduação em Engenharia de Produção, UFSC, 1998.

PERREAULT, W. D.; BARSDALE, H. C. A Model-Free Approach for Analysis of Complex Contingency Data in Survey Research. *Journal of Marketing Research.* 1980, Nº 27, p.503-515.

WRIGLEY, N. *Categorical Data Analysis for Geographers and Environmental Scientists*. Longman, 1985.