UTILIZAÇÃO DO MÉTODO DE CHI-SQUARE AUTOMATION INTERACTION DETECTION (CHAID) NA EXPLICAÇÃO DO PREÇO PROCURADO DE IMÓVEIS EM DISTINTOS SEGMENTOS DO MERCADO

FREITAS, Ana Augusta F. (1); OLIVEIRA, Maria Carolina G. (2); HEINECK, Luiz Fernando M. (3)

(1) Mestre em Engenharia de Produção, doutoranda do Programa de Engenharia de Produção, Universidade Federal de Santa Catarina, Caixa Postal 476 -CEP: 88040-900 Florianópolis-SC

(2) Mestre em Engenharia de Produção, doutoranda do Programa de Engenharia de Produção

(3) PhD, Professor Titular do Programa de Engenharia de Produção

RESUMO

Este trabalho tem como objetivo explicar o preço do imóvel procurado em diversos segmentos do mercado, através da utilização da técnica de CHAID e dos modelos de análise de variância. A primeira técnica tem como objetivo o pré-procesamento dos dados através da escolha das variáveis importantes de serem incluídas na equação de regressão desenvolvida na segunda etapa. O banco de dados utilizado na análise 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 renda mensal familiar. Com base nestes resultados, utiliza-se os modelos de análise de variância para quantificar a diferença dos preços entre os diversos segmentos, indicando também as diferenças regionais encontradas.

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 determination.

1. O PROCESSO DE ESCOLHA E OS DADOS UTILIZADOS NO ESTUDO

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) como nos 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 como composição familiar, participação no mercado de trabalho e número de pessoas que contribuem com a formação da renda familiar. No entanto, alguns trabalhos mostram que estas variáveis afetam diferentemente os diversos segmentos do mercado em relação à escolha do imóvel (CLARK *et al*., 1991).

A fim de descobrir esta estrutura, propõe-se a utilização da técnica do CHAID, com o objetivo de identificar os principais grupos (formado em função de algumas variáveis sócio-econômicas) e suas escolhas em relação ao preço do imóvel desejado para a compra. Para fins de exemplificação da técnica, um banco de dados contendo cerca de 3 mil entrevistas com clientes potenciais será utilizado. Estes dados são referentes a entrevistas conduzidas em onze diferentes cidades do país. São elas: 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 as características sócio-econômicas do indivíduo (ex. 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 as macro-variáveis do imóvel (número de quartos, garagens, suítes e localização) e condições de pagamento (ex. preço do imóvel procurado, valor da prestação). A terceira parte do questionário analisava a disponibilidade em pagar a mais por vários atributos residenciais e a última testava a força de alguns atributos através de questões onde o entrevistado era colocado em uma situação onde ele deveria avaliar a troca entre possibilidades de projeto. O presente trabalho utiliza dados relativos à primeira e segunda parte do questionário.

2. FUNDAMENTAÇÃO TEÓRICA DO CHAID E ANÁLISE DOS DADOS

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, onde através de um procedimento em cascata, divide-se o conjunto de variáveis em dois sub-grupos, através da maximização da soma dos quadrados entre sub-conjuntos. Esta técnica foi expandida para os casos onde a variável dependente é qualitativa, como propõe KASS (1980), e é conhecida como Chi-square automatic interaction detection (CHAID). Neste caso, as categorias das variáveis independentes são agregadas, se mostrarem padrões de comportamento semelhantes em

relação à variável dependente. Além disto, para cada uma das categorias das variáveis independentes selecionadas, a próxima variável que melhor prediz a categoria da variável anterior é escolhida. Ao final, os resultados da análise são mostrados em forma de uma árvore (dendograma), onde os segmentos da população são determinados.

O objetivo principal do CHAID é encontrar as principais interações entre grupos de pessoas e escolha habitacional 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). 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 em modelos lineares gerais, como os modelos de análise de variância. No presente trabalho, o preço do imóvel desejado foi considerado a variável dependente e será explicado em função das características sócio-econômicas do indivíduo. Numa etapa inicial várias variáveis foram selecionadas para compor a análise, sendo que a técnica escolheu aquelas mais importantes para a predição do preço desejado. As mesmas são mostradas na tabela 1 abaixo, juntamente com as suas categorias.

Tabela 1 – Variável dependente e independentes e suas categorias

Variáveis	Categorias
Preço do imóvel desejado (variável dependente)	**1**. Até R$ 42.500; **2**.De R$ 42.500 a R$ 55.000 **3**. De R$ 55.000 a R$ 75.000; **4**. De R$ 75.000 a R$120.000; **5**. Mais de R$ 120.000
Renda mensal familiar	**1**. Até 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
Estado civil	**1**. Solteiro; **2**. Casado; **3**.Outros
Valor dos bens	**1**. Até R$ 13.000; **2**. De R$ 13.000 a R$ 27.000 **3**. De R$ 27.000 a R$ 41.000; **4**. De R$ 41.000 a R$78.000; **5**. Mais de R$ 78.000
Posse da moradia atual	**1**. Própria; **2**. Alugada; **3**. Outras
Idade	**1**. Ate 25 anos; **2**. De 26 a 35 anos **3**. De 36 a 45 anos; **4**. Mais de 45 anos

O conjunto total de pessoas entrevistadas nas feiras de imóveis reduziu-se neste exemplo de 2764 casos para 2344. Cada um destes indivíduos esta alocado a uma das ramificações da análise. Em alguns casos, não foi possível obter informações para as cinco variáveis independentes, o que caracteriza a existência de pontos faltantes. Para fins de simplificação da visualização, os mesmos foram retirados do dendograma (figura 1), quando eles apareciam em uma categoria isolada. Onde este grupo de valores faltantes apresentaram semelhança com outras categorias, os mesmos foram automaticamente agregadas pelo CHAID e estão representadas por um ponto (ex: bens 34.). As categorias das variáveis independentes são representadas no dendograma, pelo mesmo número que foram codificadas na tabela 1.

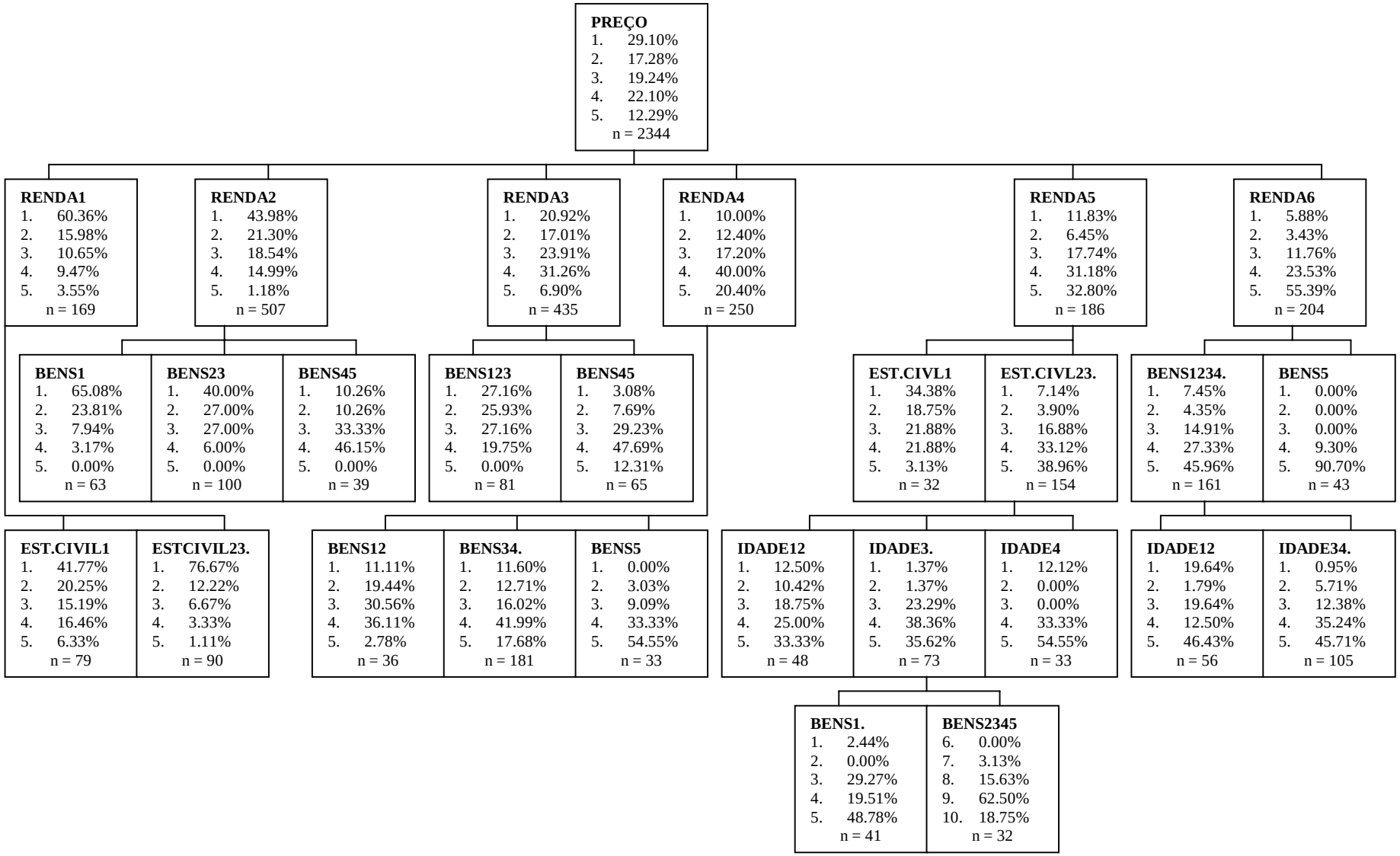

Figura 1 – CHAID Dendograma. Os valores nos quadros correspondem ao percentual em cada categoria de preço

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, os solteiros têm uma predisposição de pagar um pouco mais comparativamente aos casados e outros (divorciados e viúvos);
3. Indivíduos com renda mensal familiar entre R$ 1.000 e R$ 4.000 consideram em segundo lugar o valor dos bens que eles possuem. Em todos os casos o aumento do valor dos bens é acompanhado pelo aumento da predisposição em pagar a mais pelo valor do imóvel. No entanto, é importante perceber como as categorias destas variáveis foram agregadas, segundo os diferentes níveis de renda. Nas classes de renda até R$ 2.000, o aumento do valor dos bens leva a um ajuste mais preciso da definição do preço desejado (mais categorias da variável valor dos bens). Em classes superiores de renda, o nível de agregação das categorias é maior (em geral apenas duas categorias). Poderia-se, neste caso redividir a categoria de valor dos bens. Inicialmente ela foi considerada de maneira geométrica (até R$13.000, de R$13.000 a R$27.000, de R$27.000 a R$41.000, etc.), o que garantiu que se tivesse mais ou menos o mesmo número de casos em cada uma das categorias. Seria o caso agora de explorar uma subdivisão maior do valor dos bens para estas rendas mais altas a fim de conseguir uma precisão maior em relação ao efeito desta variável ;
4. Para os clientes com renda entre R$ 4.000 e R$ 5.000, o elemento diferenciador volta a ser estado civil. Casados e outros (divorciados e viúvos) têm uma predisposição a adquirir apartamentos de maior valor do que os solteiros. No entanto, entre estes, o aumento da idade está associada a um aumento de preço. Pode-se perceber que houve uma inversão da influência do estado civil na predisposição para pagar a mais pelo imóvel. Na primeira faixa de renda, os solteiro pagam mais que os casados, enquanto que na faixa de R$ 4.000 R$ 5.000 inverte-se esta relação;
5. Para os indivíduos na última faixa de renda, a variável diferenciadora volta a ser valor dos bens, só que a um nível muito mais agregado, e é claro separando os patrimônios elevados dos de menor valor (maior que R$ 78.000 e os outros). No caso de valor patrimoniais altos, 90.70% das pessoas procuram imóveis da mais alta faixa de preço (acima de R$120.000). Para faixa de valor de bens menores, a idade do chefe da família é importante na definição do valor a pagar, sendo os jovens mais representados nas faixas menor de valor.

A análise geral dos resultados do dendograma nos permite gerar algumas equações através dos modelos de análise de variância, para cada nível de renda. Estas equações levam em consideração as variáveis escolhidas pela análise do CHAID e a agregação das suas categorias. Para facilitar a interpretação dos resultados a variável preço foi utilizada na forma contínua e não mais na forma categórica, como apresentado anteriormente.

Renda 1 (até R$ 1.000): A principal variável para esta categoria de renda é o estado civil, sendo que este pode ser agregado em apenas dois níveis: solteiros e outros (casados, divorciados e viúvos). A equação toma a seguinte forma:

Preço = 36.602 + 16.182 (civil). Isto significa que os solteiros pagam em média 16 mil reais a mais que os casados, um valor alto considerando que o valor dos imóveis são nas maioria dos casos menores que R$ 42.500.

Renda 2 (de R$1.000 a R$ 2.000): A principal variável para esta categoria de renda é o valor dos bens, divididos em três níveis: até R$13.000, entre R$13.000 e R$ 41.000 e mais de R$ 41.000. A equação neste caso, toma a seguinte forma:

Preço = 69.871 – 28.165 (bens1) – 20.616 (bens2). Isto significa que um aumento nos valor do bens disponíveis pode significar uma disponibilidade para pagamento do imóvel na ordem de 20 a 30 mil reais.

Renda 3 (de R$ 2.000 a R$ 3.000): A principal variável para esta categoria de renda é o valor dos bens, divididos em dois níveis: até R$41.000 e mais de R$ 41.000. A equação neste caso, toma a seguinte forma:

Preço = 82.307 – 25.394 (bens1). Isto significa que os indivíduos que ganham entre 2 e 3 mil reais com um valor de bens maior que R$ 41.000 pagariam cerca de 30% a mais pelo imóvel desejado do que os indivíduos com valor de bens menor do que este limite.

Renda 4 (de R$ 3.000 a R$ 4.000): A principal variável para esta categoria de renda é o valor dos bens, divididos em três níveis: até R$ 27.000, entre R$ 27.000 e R$ 78.000 e mais de R$ 78.000. A equação neste caso, toma a seguinte forma:

Preço = 111.071 – 43.807 (bens1) - 34.600(bens2). Isto significa que um aumento nos valor do bens disponíveis pode significar uma disponibilidade para pagamento do próximo imóvel na ordem de 30 a 40 mil reais.

Renda 5 (de R$ 4.000 a R$ 5.000): As principais variáveis nesta categoria são estado civil (em dois níveis: solteiros e outros), idade (em três níveis: ate 35 anos, entre 36 e 45 anos e mais de 45 anos) e valor dos bens (em dois níveis: até R$13.000 e mais de R$13.000). A equação neste caso, toma a seguinte forma:

Preço = 110.951 – 18.205 (bens1) – 16.400 (idade1) – 8.295(idade2) – 33.927 (estado1). É interessante notar que nesta equação apenas a variável estado civil é significativa (a um nível de 5%). Isto significa que a variável idade, assim como detectado pelo CHAID só é importante para os indivíduos casados e divorciados ou viúvos. Da mesma forma, a variável valor dos bens só é significativa para os indivíduos como idade entre 36 e 45 anos.

Renda 6 (mais de R$ 5.000): As principais variáveis nesta categoria são: valor dos bens (em dois níveis: até R$ 78.000 e mais de R$ 78.000) e idade (em dois níveis: ate 35 anos e mais de 36 anos). A equação neste caso, toma a seguinte forma:

Preço = 172.786 – 69.551 (bens1) – 10.964 (idade1). Interpretando a equação, nota-se que os indivíduos nesta faixa de renda, acima de 36 anos com alto valor patrimonial estariam dispostos a pagar em média cerca de 172 mil reais pelo imóvel. A diferença para segunda categoria (valor dos bens menor que R$ 78.000 e menos de 36 anos) é de cerca de 46%.

3. 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, o valor dos bens patrimonial é também uma variável muito importante, sendo que o estado civil aparece como fator influente no valor de compra nas rendas menores que R$1.000 reais e entre R$4.000 e R$5.000. Para as categorias de

renda mais alta, a idade também mostra-se importante. Com base nestes resultados conclui-se que a técnica do CHAID apresentou-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 usados em modelos de análise de variância, indicando para segmentos específicos um conjunto diferente de fatores influenciadores do processo de escolha. Na análise de variância, obtém-se valores para os imóveis escolhidos, 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 valia neste tipo de modelos que possuem como principal desvantagem as restrições impostas em relação ao número de variáveis a serem utilizadas.

4. REFERÊNCIAS BIBLIOGRÁFICAS

BOEHM. T. P.. A Hierarchical Model of Housing Choice. **Urban Studies**, v. 19, 1982, 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**, v. 10, nº 1, 1995, 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**, v. 20, 1988, p. 198-219.

CLARK, W. A. V.; DEURLOO, M. C.; DIELEMAN, F. M.. Categorical Data with Chi-Square Automatic Interaction Detection and Correspondece Analysis. **Geographical Analysis**, v. 23, 1991, 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**, v. 31, nº 1, 1994, p. 137-154.

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

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

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

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