APLICAÇÃO DE TÉCNICAS DE DESCOBRIMENTO DE CONHECIMENTO EM BASES DE DADOS E DE INTELIGÊNCIA ARTIFICIAL EM AVALIAÇÃO DE IMÓVEIS

Marco Aurélio Stumpf González

Porto Alegre
dezembro 2002

MARCO AURÉLIO STUMPF GONZÁLEZ

APLICAÇÃO DE TÉCNICAS DE DESCOBRIMENTO DE CONHECIMENTO EM BASES DE DADOS E DE INTELIGÊNCIA ARTIFICIAL EM AVALIAÇÃO DE IMÓVEIS

Tese apresentada ao Programa de Pós-Graduação em Engenharia Civil da Universidade Federal do Rio Grande do Sul, como parte dos requisitos para obtenção do título de Doutor em Engenharia.

Porto Alegre

dezembro 2002

GONZÁLEZ, Marco Aurélio Stumpf González

Aplicação de técnicas de descobrimento de conhecimento em bases de dados e de inteligência artificial em avaliação de imóveis / Marco Aurélio Stumpf González. – Porto Alegre: PPGEC/UFRGS, 2002.

300 p.

Tese de Doutorado, Programa de Pós-Graduação em Engenharia Civil da Universidade Federal do Rio Grande do Sul; Doutor em Engenharia. Orientador: Carlos Torres Formoso.

1. Avaliação de imóveis I. Aplicação de técnicas de descobrimento de conhecimento em bases de dados e de inteligência artificial em avaliação de imóveis.

CCAA2

MARCO AURÉLIO STUMPF GONZÁLEZ

APLICAÇÃO DE TÉCNICAS DE DESCOBRIMENTO DE CONHECIMENTO EM BASES DE DADOS E DE INTELIGÊNCIA ARTIFICIAL EM AVALIAÇÃO DE IMÓVEIS

Esta tese de doutorado foi julgada adequada para a obtenção do título de DOUTOR EM ENGENHARIA e aprovada em sua forma final pelo professor orientador e pelo Programa de Pós-Graduação em Engenharia Civil da Universidade Federal do Rio Grande do Sul.

Porto Alegre, 02 de Dezembro de 2002.

Prof. Carlos Torres Formoso
PhD pela University of Salford, Grã Bretanha
Orientador

Prof. Francisco P. S. L. Gastal
PhD pela North Carolina State University, EUA
Coordenador do PPGEC/UFRGS

BANCA EXAMINADORA

Profa. Beatriz de Faria Leão (Min. da Saúde)
Dra. pela Escola Paulista de Medicina/SP

Profa. Claudia Monteiro De Cesare (PMPA)
PhD pela University of Salford, Grã Bretanha

Prof. José Luis Duarte Ribeiro (PPGEP/UFRGS)
Dr. pelo PPGEC/UFRGS

Prof. Lucio Soibelman (University of Illinois at Urbana-Champaign/USA)
PhD. pelo Massachusetts Institute of Technology, EUA

Com a dedicação ao Doutorado, ao curso de Direito, à docência, à Coordenação da Engenharia Civil e a tantas outras atividades, sei que coisas importantes foram adiadas, mas a prioridade sempre foi uma só:

Viviane.

AGRADECIMENTOS

Inicialmente, devo agradecer ao Professor Carlos Torres Formoso, meu orientador, que contribuiu decisivamente ao longo do desenvolvimento deste trabalho. Agradeço também à Professora Claudia Monteiro De Cesare, que sugeriu o tema de pesquisa e contribuiu na definição geral do trabalho, e ao Professor Lucio Soibelman, que também participou de forma fundamental no desenvolvimento da pesquisa, nas discussões e no apoio importante recebido nas semanas que passei na Universidade de Illinois, em Urbana-Champaign, EUA.

Diversas pessoas e instituições contribuíram para o desenvolvimento deste trabalho, de várias formas. Embora deva reconhecer que possivelmente existam omissões, gostaria de lembrar uma lista relevante de pessoas. Assim, agradeço:

Ao Professor Jose Antonio do Nascimento Pinto, com quem tive a honra de dividir a Coordenação do Curso de Engenharia Civil da UNISINOS durante três anos de intenso trabalho. Seu apoio incondicional foi extremamente importante. Mais que um amigo, eu o considero um irmão mais velho.

Ao Professor Diego Alfonso Erba, grande amigo, parceiro de muitos trabalhos e pesquisas, e também sócio, com quem enfrentei tantas turbulências ao longo destes oito anos que nos conhecemos, parece que agora "o ano que vem será diferente, mesmo".

Ao Professor Luiz Fernando Mählmann Heineck, que contribuiu generosamente com grande quantidade de literatura, e que é um exemplo de pesquisador a ser seguido.

À Professora Sílvia Costa Dutra, Diretora do Centro de Ciências Exatas e Tecnológicas da UNISINOS, pessoa que admiro profundamente e que me proporcionou a oportunidade de atuar na Coordenação do curso de Engenharia Civil.

Ao Professor Volnei Pereira da Costa, Vice-Diretor do Centro de Ciências Exatas e Tecnológicas da UNISINOS, pelo apoio constante.

Ao Professor André Maciel Zeni, o primeiro mestre na metodologia científica de avaliações, que despertou o interesse para esta área da Engenharia.

À Professora Andrea Parisi Kern, e aos demais colegas da UNISINOS.

Ao amigo Cristóvão Carneiro Cordeiro, e ao pessoal de Feira de Santana.

À Universidade do Vale do Rio dos Sinos, que apoiou financeiramente o desenvolvimento deste Doutorado por quatro anos e, principalmente, que me proporciona diariamente a oportunidade de crescer no convívio com a filosofia da formação integral, no exercício da docência e da pesquisa, enfim, na busca do *Magis*.

O início desta caminhada foi em 1997, no Programa de Pós-Graduação em Economia da UFRGS, à época coordenado pelo Professor Marcelo Portugal, com quem cursei disciplinas de Econometria. A Professora Otília B. K. Carrion foi a primeira orientadora. Agradeço a esses professores e também ao CNPq, que forneceu apoio financeiro nesse primeiro ano de Doutorado.

Por fim, é necessário agradecer à Universidade Federal do Rio Grande do Sul, e especialmente à sociedade brasileira, que me proporcionou ensino de qualidade em dois cursos de graduação, no Mestrado e agora neste Doutorado. A dívida é grande, e pretendo começar a retribuir brevemente.

Pelo apoio fundamental, às minhas duas famílias, a de nascimento e a família da minha querida Viviane: Alberto, Rodrigo, Cristina, Ana Lúcia, Yara e Antônio.

Sobretudo, é preciso agradecer ao Criador e dizer da saudade de minha mãe, Gladis N. S. González.

RESUMO

GONZÁLEZ, M. A. S. Aplicação de Descobrimento de Conhecimento em Bases de Dados e Inteligência Artificial em Avaliação de Imóveis. 2002. Tese (Doutorado em Engenharia Civil) – Programa de Pós-Graduação em Engenharia Civil, UFRGS, Porto Alegre.

A comparação de dados de mercado é o método mais empregado em avaliação de imóveis. Este método fundamenta-se na coleta, análise e modelagem de dados do mercado imobiliário. Porém os dados freqüentemente contêm erros e imprecisões, além das dificuldades de seleção de casos e atributos relevantes, problemas que em geral são solucionados subjetivamente. Os modelos hedônicos de preços têm sido empregados, associados com a análise de regressão múltipla, mas existem alguns problemas que afetam a precisão das estimativas.

Esta Tese investigou a utilização de técnicas alternativas para desenvolver as funções de preparação dos dados e desenvolvimento de modelos preditivos, explorando as áreas de descobrimento de conhecimento e inteligência artificial. Foi proposta uma nova abordagem para as avaliações, consistindo da formação de uma base de dados, ampla e previamente preparada, com a aplicação de um conjunto de técnicas para seleção de casos e para geração de modelos preditivos. Na fase de preparação dos dados foram utilizados as técnicas de regressão e redes neurais para a seleção de informação relevante, e o algoritmo de vizinhança próxima para estimação de valores para dados com erros ou omissões. O desenvolvimento de modelos preditivos incluiu as técnicas de regressão com superfícies de resposta, modelos aditivos generalizados ajustados com algoritmos genéticos, regras extraídas de redes neurais usando lógica difusa e sistemas de regras difusas obtidos com algoritmos genéticos, os quais foram comparados com a abordagem tradicional de regressão múltipla.

Esta abordagem foi testada através do desenvolvimento de um estudo empírico, utilizando dados fornecidos pela Prefeitura Municipal de Porto Alegre. Foram desenvolvidos três formatos de avaliação, com modelos para análise de mercado, avaliação em massa e avaliação individual. Os resultados indicaram o aperfeiçoamento da base de dados na fase de preparação e o equilíbrio das técnicas preditivas, com um pequeno incremento de precisão, em relação à regressão múltipla. Os modelos foram similares, em termos de formato e precisão, com o melhor desempenho sendo atingido com os sistemas de regras difusas.

Palavras-chave: avaliação de imóveis; inteligência artificial; descobrimento de conhecimento em bases de dados; redes neurais artificiais; inferência estatística.

ABSTRACT

GONZÁLEZ, M. A. S. Aplicação de Descobrimento de Conhecimento em Bases de Dados e Inteligência Artificial em Avaliação de Imóveis. 2002. Tese (Doutorado em Engenharia Civil) – Programa de Pós-Graduação em Engenharia Civil, UFRGS, Porto Alegre.

The most used property valuation method is the sales-comparison approach, which is based on the collection, analysis and interpretation of market data. There are some problems in current appraisals, such as problems with available data, which is frequently dirty and incomplete, and difficulties on the choice of relevant cases and attributes. These problems are regularly solved by using the property valuation expertise, which results in limitations in terms of predictive accuracy and defensibility. In recent years though, hedonic regression models have been used to valuation, but remain some problems, affecting the quality of estimates.

In order to improve valuations, this thesis looks for alternative techniques for data preparation and data modeling, exploring the knowledge discovery in databases and artificial intelligence fields. This thesis presents an application of an appraisal system, developed to support property valuation, by investigating some tools in data preparation and modeling. In the data preparation step were used neural and regression models in the selection of relevant cases and attributes, and nearest neighbors to improve wrong or missing data. A set of tools were investigated to modeling step, using response surface, generalized additive models, rules extracted from neural networks and fuzzy rule-based systems to generate predicting models.

This approach was tested using data from Porto Alegre Council. These models were compared with traditional regression models, in three approaches: market models, mass-appraisal models and singular valuation models. This multi-agent approach reduces the risk of the estimates and data preparation positively improves data quality. Final results point to a general equilibrium among the modeling techniques, in regard of predictive accuracy, with a small predominance to fuzzy rule-based systems.

Key-words: property valuation; artificial intelligence; knowledge discovery in databases (KDD); artificial neural networks; statistical inference.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

ABREVIATURAS

Termo adotado nesta Tese		Correspondente em inglês	
ARM:	Análise de Regressão Múltipla	*MRA:*	*Multiple Regression Analysis*
COD:	Coeficiente de Dispersão	*COD:*	*Coefficient of Dispersion*
DCBD:	Descobrimento de Conhecimento em Bases de Dados	*KDD:*	*Knowledge Discovery in Databases*
EA:	Erro Absoluto percentual	*MAE:*	*Mean Absolute Error*
EP:	Erro Padrão	*SE:*	*Standard Error*
IA:	Inteligência Artificial	*AI:*	*Artificial Intelligence*
IPTU:	Imposto sobre a Propriedade Territorial Urbana		*Property Tax*
ITBI:	Imposto sobre a Transmissão de Bens Imóveis		*Sales Tax*
MAG:	Modelos Aditivos Generalizados	*GAM:*	*Generalized Additive Models*
MD:	Mineração de Dados	*DM:*	*Data Mining*
PCA:	Análise de Componentes Principais	*PCA:*	*Principal Component Analysis*
RBC:	Raciocínio Baseado em Casos	*CBR:*	*Case-Based Reasoning*
RNA:	Redes Neurais Artificiais	*ANN:*	*Artificial Neural Networks*
SBRD:	Sistemas Baseados em Regras Difusas	*FRBS:*	*Fuzzy Rule-Based Systems*
SBRDE:	Sistemas Baseados em Regras Difusas Evolucionário	*GFRBS:*	*Genetic Fuzzy Rule-Based Systems*
SID:	Sistema de Inferência Difusa	*FIS:*	*Fuzzy Inference System*

1 INTRODUÇÃO

1.1 CONSIDERAÇÕES INICIAIS

A avaliação de imóveis consiste na determinação do valor de mercado de um imóvel, entendido como o preço mais provável que este imóvel atingiria em uma transação normal, de acordo com suas características e com as condições do mercado naquele momento. Embora as estimativas de valor sejam importantes - e mesmo decisivas - para diversas utilizações, as técnicas utilizadas atualmente apresentam alguns inconvenientes que resultam em diminuição da precisão destas estimativas, indicando a necessidade de aperfeiçoamento.

A pesquisa de alternativas justifica-se pela importância econômica e social do mercado imobiliário e pelas possíveis conseqüências da imprecisão nas mensurações realizadas neste campo, ou seja, os potenciais prejuízos econômicos e sociais decorrentes dos erros cometidos nas estimativas do valor de mercado dos imóveis. Podem ser identificados três tipos ou formatos de avaliação, incluindo modelos de análise geral, para avaliação em massa e para avaliação individual. Entre outras aplicações, estes modelos podem ser empregados na definição de planos diretores, em estudos de viabilidade econômica de novas construções, nas estimativas para liberação de financiamentos (estimativa do montante de garantia), nas desapropriações e na tributação imobiliária. Nos casos citados, avaliações incorretas podem causar erros no planejamento urbano ou na avaliação da viabilidade econômica, excesso de pagamento do mutuário ou falta de garantia no caso de inadimplência, evasão de recursos públicos ou ineqüidade na tributação (com injustiça social), respectivamente. Apresenta-se a seguir alguns dados que evidenciam esta relevância, e em seguida os problemas típicos nas avaliações.

As transações de financiamento imobiliário são embasadas por avaliações do valor do bem. O principal organismo atuante em empréstimos é a Caixa Econômica Federal (CEF), com aproximadamente 85% dos empréstimos para aquisição de imóveis residenciais, tendo atingido um saldo de empréstimos de R$ 75 bilhões em 1999 (Leal, 2001). Nos últimos anos,

as transações foram de mais de R$ 900 milhões anuais, em média. O número de transações está na faixa de 18 a 19 mil transações por ano, em todo o país. Percebe-se que as transações com financiamento oficial representam apenas uma pequena parte do mercado imobiliário, mas ainda assim o montante de recursos é importante. Deve-se ressalvar também que estes valores incluem apenas a parcela financiada pela CEF, existindo ainda um expressivo montante de recursos do adquirente, provenientes de economias próprias ou de liberações de contas do Fundo de Garantia por Tempo de Serviço (FGTS), e também outros financiamentos, destinados à autoconstrução[1].

Outra utilização importante das avaliações é na tributação. Os tributos imobiliários, tais como o Imposto Predial e Territorial Urbano (IPTU) e o Imposto sobre a Transmissão de Bens Imóveis (ITBI), são baseados nos valores venais dos imóveis, ou seja, nos valores de mercado. Estes valores são obtidos através da estimação em avaliações gerais, conhecidas como Plantas Genéricas de Valores (PGV)[2]. A correta estimação é fundamental para garantir a eqüidade na tributação. Estes tributos, especialmente o IPTU, representam uma importante fonte de arrecadação para os municípios brasileiros (De Cesare, 1998). Bremaeker (1994) verificou que a participação média do IPTU na receita dos municípios brasileiros era de 3% até 1988 e de cerca de 4% depois das alterações na distribuição dos tributos[3]. Por exemplo, no caso de Porto Alegre, estes tributos, somados, representam atualmente cerca de 12% da arrecadação total, embora o IPTU tenha atingido próximo de 20% no início da década de 70. Em 2000, para uma arrecadação total de cerca de R$ um bilhão, a arrecadação com o IPTU foi de R$ 83,6 milhões, e a do ITBI de R$ 42 milhões (Porto Alegre, 1970-2000). A redução ocorrida nos últimos anos deve-se basicamente à nova distribuição tributária gerada pela

---

[1] Os valores de financiamento são tabulados pelo Banco Central do Brasil (BACEN/SFH–SBPE/DINOR/DECAD/DIHAB) e foram obtidos em www.cbic.org.br. Em outros países o mercado imobiliário também é importante economicamente. Por exemplo, nos Estados Unidos, o estoque de unidades para habitação era de 115,9 milhões em 1999, sendo aproximadamente 65% delas ocupadas pelos proprietários. O valor total foi estimado em USD 11,6 trilhões, com um saldo de hipotecas sobre estes imóveis em USD 5,1 trilhões, na mesma época, contra US$ 3 trilhões em 1991 (Case, 2000; Eckert *et al.*, 1993).

[2] Na visão tradicional, as Plantas Genéricas de Valores (PGV) consistem de uma planta ou conjunto de plantas com os valores unitários dos terrenos (Lima, 1990; Möller, 1995). Mais recentemente, têm sido propostas PGV com base em modelos de regressão específicos para cada tipo de imóvel (González e Erba, 1997; Zancan, 1996).

[3] Bremaeker (1994) afirmou que o IPTU era de 20% e 27% das receitas tributárias, respectivamente antes e depois da Constituição, e que as receitas tributárias representam em média apenas 15% da receita dos municípios brasileiros, sendo o restante originado de transferência municipais e federais. Este autor também afirmou que a participação do IPTU é significativamente diferente nas capitais dos estados, em relação aos demais municípios.

Constituição brasileira de 1988, com o retorno do ITBI intervivos para os municípios, mas com um significativo incremento das receitas oriundas de transferências federais e estaduais. A participação dos tributos imobiliários ainda é importante, embora existam críticas quanto à eqüidade do IPTU e quanto à pequena exploração do potencial de arrecadação deste tributo (De Cesare, 1998). Por outro lado, o imposto sobre a transmissão geralmente envolve um grande número de avaliações, nas grandes cidades, indicando a necessidade de um sistema ágil e preciso de avaliação dos valores de mercado.

As avaliações judiciais em ações relativas a desapropriações também desempenham um papel economicamente relevante. Desde a edição da Constituição atual, que ampliou a função social da propriedade, a tendência de desapropriações para regularização fundiária de áreas irregulares vem substituindo a prática de expulsão dos invasores, adotada em décadas anteriores, sob outro contexto político. A nova legislação urbana, especialmente o Estatuto da Cidade (Lei 10.257/01), reforça esta visão, oferecendo novos mecanismos para gestão das cidades, tais como facilidades para desapropriações urbanas, regularização fundiária e impostos progressivos, os quais têm vinculação com os valores dos imóveis. Na área rural, a Reforma Agrária também vem ampliando as desapropriações, com significativos volumes de recursos. Porém, levantamentos recentes demonstram evidências de abusos em avaliações urbanas e rurais, com indicações de expressivos prejuízos para a sociedade, através da evasão de recursos públicos (Moreira, 2001). Em 2000, por exemplo, o governo federal tinha dívidas de mais de R$ 3,1 bilhões de créditos em precatórios para desapropriações (principalmente áreas rurais) e alguns municípios tinham dívidas igualmente grandes: R$ 250 milhões em Santo André, R$ 155 milhões em São Bernardo do Campo e R$ 100 milhões em Diadema, entre outros municípios citados. Esta situação sinaliza a importância de mecanismos eficientes e confiáveis de avaliação do valor do solo (Moreira, 2001).

Por fim, a eventual utilização de outras formas de tributação, como a contribuição de melhorias e a captura de mais-valias sobre a valorização de terrenos urbanos também exige estimativas corretas. Alguns autores defendem a utilização extrafiscal para auxiliar na gestão urbana, através de instrumentos como a outorga onerosa ("solo criado"), ampliando o escopo e a importância das avaliações tributárias para a sociedade (Abramo, 2001b; Biava, 1986; George, 1972; Smolka e Furtado, 2001; Wilderode, 1997). Efetivamente, alguns municípios têm utilizado instrumentos como a outorga onerosa, geralmente em forma de troca de índices construtivos por recursos destinados à produção de habitação de interesse social, tais como

Porto Alegre (solo criado), Rio de Janeiro e São Paulo (operações interligadas), mas nem sempre com sucesso. Por exemplo, Wilderode (1997) descreve a prática das operações interligadas em São Paulo, nas quais, mesmo envolvendo vultuosas somas e lucros para os incorporadores, os resultados em termos de benefício social são tímidos. De qualquer forma, a gestão da outorga onerosa é complexa e a verificação dos valores de mercado do solo e das construções é fundamental, permitindo a verificação das vantagens que podem ser obtidas pelo incorporador e a fixação de contrapartidas proporcionais.

As plantas de valores, utilizadas atualmente apenas para tributação, poderiam desempenhar importante papel nesta gestão, mas, infelizmente, sabe-se que os valores cadastrais geralmente são desatualizados, mesmo em municípios grandes. Por outro lado, há problemas decorrentes de erros nos valores estimados. Por exemplo, De Cesare (1998) demonstrou a existência de ineqüidades no imposto sobre a propriedade em Porto Alegre, e há relatos semelhantes em outros locais. Para a eficácia de qualquer destas funções é necessário obter-se boas estimativas dos valores dos imóveis. As diferentes valorizações relativas entre regiões induzem ineqüidades nos impostos imobiliários, pois a administração municipal, via de regra, não consegue captar os efeitos de localização no valor de mercado, utilizado como base de cálculo para fins tributários e a eqüidade nas avaliações é requisito fundamental para a garantia da justiça tributária (De Cesare, 1998; Gonçalves, 1988; Leal, 1990; Liporoni, 1993; Smolka, 1994a, 1994b; Varsano, 1977). Conforme Moscovitsch (1997), devem ser utilizados métodos e técnicas de avaliação menos subjetivos. Thrall (1998) e Wachs (1978), neste sentido, afirmam que as avaliações de massa deveriam utilizar procedimentos automatizados, a fim de aumentar a precisão (através da redução dos erros humanos) e diminuir o custo da reavaliação periódica das propriedades.

Além disto, é importante lembrar que, além da relevância econômica, o mercado imobiliário tem significativa importância social e política, com fortes implicações nos sonhos e ideários da população brasileira, pois envolve a moradia familiar, bem como significativa atenção e interferência governamental (Maricato, 1987, 1997; Valladares, 1981; Werna *et al.*, 2001).

1.2 HISTÓRICO E PROBLEMAS DETECTADOS NAS PRÁTICAS ATUAIS

O mercado imobiliário tem algumas características diferenciadoras, tais como forte influência da localização e da heterogeneidade dos bens, além de ter funcionamento em regime de concorrência imperfeita. Existem vários métodos para realizar a avaliação do valor de mercado, sendo que a comparação de dados de mercado é o método mais utilizado atualmente[4]. Basicamente este método realiza uma compensação das diferenças entre os imóveis de uma amostra do mercado e aquele para o qual se deseja obter o valor. Como método de avaliação, a comparação de dados tem o apelo de reproduzir o comportamento intuitivo dos agentes no mercado, sejam eles vendedores, intermediários ou compradores, os quais ponderam as diferenças entre os imóveis disponíveis para então decidir sobre o preço adequado para vender um imóvel ou sobre qual das alternativas disponíveis a escolha de compra deve recair (Abramo, 1999 e 2001a; Balchin e Kieve, 1986; Dantas, 1998; Gelbtuch *et al.*, 1997; Moreira, 1997; Robinson, 1979; Sauter, 1985).

Na realidade, o método comparativo de dados de mercado participa implicitamente nos outros processos avaliatórios. Os outros métodos são utilizados nos casos em que não existem dados sobre o tipo específico de imóvel, mas dependem da estimação de alguns parâmetros de mercado. Por exemplo, o método da renda fundamenta-se na estimação de aluguéis ou outras rendas e de taxas de rentabilidade médias de mercado, enquanto que os métodos involutivo e do custo de reprodução utilizam como base valores de terrenos (ABNT, 1989; 2000; Dantas, 1998; Fiker, 1993; Moreira, 1997).

Como os imóveis são heterogêneos e há diversas características importantes a serem consideradas simultaneamente, é necessário utilizar uma técnica ou algoritmo para realizar o ajustamento das diferenças entre os imóveis. Uma das técnicas mais utilizadas atualmente pelos avaliadores para esta função é a análise de regressão múltipla, que busca um modelo do segmento de mercado em questão (na forma de uma equação contendo as características importantes e seus pesos respectivos), validado estatisticamente e posteriormente utilizado na projeção do valor do imóvel em estudo (ABNT, 1989; Dantas, 1998; González, 1997;

[4] No capítulo 2 serão desenvolvidos em maior detalhe os conceitos sobre mercado imobiliário, valor de mercado e métodos de avaliação.

Isakson, 2001; Ramsland e Markham, 1998).

Embora utilizada há muito tempo em pesquisa e em avaliação de massa, a análise de regressão ganhou popularidade a partir da disseminação dos computadores pessoais, na década de 80, permitindo este desenvolvimento técnico também nos escritórios de avaliações (Caples *et al.*, 1997; Griliches, 1971; Jensen, 1987; Rayburn e Tosh, 1995; Wachs, 1978).

No Brasil, a regressão múltipla foi introduzida aproximadamente na mesma época, impulsionada pelos trabalhos pioneiros de Domingos de Saboya Barbosa Filho, sendo proposta como parte de uma visão mais objetiva, chamada na época de "metodologia científica", em oposição às técnicas anteriores, criticadas por serem extremamente subjetivas. A regressão foi introduzida na Norma brasileira de avaliações, após grande debate com a corrente que defendia a tradicional homogeneização de fatores (ABNT, 1989; Barbosa Filho, 1974; Dantas, 1998; González, 1997; IBAPE, 1974)[5].

Tendo já decorrido um razoável período de tempo desde a introdução da regressão múltipla no âmbito das avaliações, a técnica já foi relativamente bem absorvida pela comunidade. Existem *softwares* amigáveis, manuais de avaliação e *sites* na Internet, disponibilizando vasto conteúdo sobre o assunto, inclusive em português (Isakson, 1998; Dantas, 1998; Gelbtuch *et al.*, 1997; González, 1997).

Entretanto, há indicações recentes de que erros de 10 a 15% do valor estimado são aceitos como normais, pelos avaliadores e pelos Tribunais da Grã-Bretanha e da Austrália (Crosby *et al.*, 1998a, 1998b; Gilbertson, 2001). Geralmente não existem indicações dos erros efetivamente cometidos na avaliação, pois os imóveis avaliados não são transacionados em diversas situações, como é o caso da avaliação para tributação. Assim, os erros são medidos indiretamente pela diferença entre as estimativas e os preços dos imóveis da amostra de dados

[5] O professor Saboya é considerado o "pai" da inferência estatística na Engenharia de Avaliações brasileira. É digno de nota que os profissionais do Rio Grande do Sul também tiveram participação expressiva neste processo de mudança. Por exemplo, os professores Ibá Ilha Moreira Filho (Universidade Federal do Rio Grande do Sul - UFRGS) e André Maciel Zeni (Pontifícia Universidade Católica do Rio Grande do Sul - PUCRS) participaram ativamente da renovação da NB 502 (renumerada para NBR 5676 em 1989) e contribuiram muito para a disseminação da inferência estatística através de cursos técnicos em diversos locais e disciplinas nos cursos de Engenharia Civil das suas Universidades. Até hoje exercem liderança local e nacional nos Institutos de avaliações (IBAPE e IGEL, respectivamente) e na atuação profissional, embora ambos agora estejam afastados da academia.

(Crosby *et al.*, 1998a, 1998b; IAAO, 1990; Newell e Kishore, 1998; Parker, 1998).

1.2.1 Subjetividade no processo avaliatório

Atualmente o processo é muito subjetivo e improvisado. Em parte, as dificuldades devem-se ao desconhecimento dos profissionais sobre o comportamento do mercado, muitas vezes por atuarem simultaneamente em diversas faixas de mercado, em diferentes locais ou com diferentes tipos de imóvel. Geralmente as avaliações são realizadas de forma pontual, usando heurísticas e sem o apoio anterior nem a consolidação posterior em um modelo geral do mercado. Efetivamente, não é comum a análise quantitativa geral do mercado, investigando tendências ou comportamentos de forma consistente, mas apenas a análise qualitativa do mercado, com a utilização de uma amostra pequena (tipicamente menor do que 50 casos e algumas vezes com apenas três casos) na geração de "modelos do mercado". Outras questões importantes, tais como escolha do formato do modelo, seleção de casos e variáveis, também são resolvidas subjetivamente (Goddard, 1999; Kinnard, 1971;Lentz e Wang, 1998; Mills e Reynolds, 1999; Smith, 1995).

Ademais, foram relatadas significativas diferenças entre as avaliações realizadas por profissionais iniciantes e por profissionais experientes, o que provavelmente decorre da acumulação de conhecimento especializado (em forma de heurísticas ou habilidades pessoais), e não do conhecimento da técnica ou método de avaliação em si, embora também existam alertas sobre o desconhecimento da técnica por parte dos avaliadores (Detweiler e Radigan, 1996; Spence e Thorson, 1998). Também existem estudos que demonstram que a autoconfiança dos avaliadores experientes pode levar a erros, e há os que falam em "arte da avaliação", aceitando a subjetividade como inerente ao processo (Diaz, 1997; Gilbertson, 2001; Kinnard, 1966; Smalley, 1995).

1.2.2 Ruptura dos pressupostos da análise de regressão

Existem críticas ao uso da análise de regressão nas avaliações, ligadas à própria possibilidade de uso da regressão em avaliações e ao respeito das condições básicas desta ferramenta. Entre outros autores, podem ser citados Can (1990), Dubin (1992), Isakson (1998), Kummerow

(2000), Mark e Goldberg (1988), Newsome e Zietz (1992) e Smith (1995). Em síntese, os principais problemas detectados devem-se às dificuldades da regressão em lidar com características peculiares do mercado imobiliário, tais como:

a) distribuição espacial: provoca autocorrelação espacial dos erros (Can, 1990; Dubin, 1992; Dubin e Sung, 1987; McCluskey *et al.*, 2000);

b) suspeita de não-linearidade dos relacionamentos e desconhecimento da influência de cada atributo: há desconhecimento do formato correto para o modelo (Caples *et al.*, 1997; Isakson, 2001);

c) influência simultânea de múltiplos atributos importantes e inter-relacionados: causa multi-colinearidade (González, 1993; Kain e Quigley, 1970; Maddala, 1988; Morton, 1977; Wilkinson e Archer, 1973);

d) não-normalidade dos erros: afeta os testes estatísticos sobre o modelo e as variáveis (Kummerow, 2000; Neter *et al.*, 1990; Smith, 1995);

e) presença de observações espúrias (*outliers*): distorce os coeficientes do modelo (Barnett e Lewis, 1984; Belsley *et al.*, 1980; Caples *et al.*, 1997; Hair *et al.*, 1998).

Estes e outros problemas afetam a precisão e até a confiabilidade do modelo gerado. Violações sérias dos pressupostos podem invalidar a característica probabilística do modelo, impedindo a generalização e transformando a análise de regressão em ferramenta de ajuste de curvas, gerando, na realidade, um modelo determinístico (Malinvaud, 1967). Neste segundo caso o modelo não poderia realizar generalizações, o que inviabiliza o seu uso em plantas genéricas de valores, por exemplo.

Em grande medida, os problemas são causados por falta de tratamento adequado aos dados. O tipo de informação utilizado como base (dados de transações imobiliárias correntes) geralmente apresenta erros ou omissões que não são considerados de forma sistemática nas análises convencionais (McCluskey *et al.*, 1997). A coleta de dados é bastante informal e, adicionalmente, no Brasil raramente estão disponíveis bases de dados coletadas por

organismos de classe ou empresas[6]. Entretanto, mesmo quando há dados disponíveis em quantidade, o avaliador normalmente busca uma amostra com os dados de maior interesse, sem analisar o todo. O tratamento inicial dos dados é dificultado justamente pela falta de um modelo geral do mercado, mais estável, que permita guiar a preparação e a modelagem dos dados. Por exemplo, é comum a confusão entre erros provocados por alterações recentes do mercado ou falhas no modelo (modelo incompleto ou incorretamente especificado), e erros decorrentes de transações realmente com problemas, não representativas do mercado imobiliário normal. No primeiro caso, o modelo precisa ser aprimorado, enquanto que no segundo as observações com problemas devem ser identificadas e tratadas adequadamente.

Um elemento complicador é que duas das principais variáveis, o valor de mercado (variável dependente) e a localização (variável preditiva), não podem ser observadas diretamente no mercado. A primeira porque, em função das dificuldades de comparação, falta de conhecimento sobre o mercado e diferenças que existem entre os agentes (compradores e vendedores), entre outras razões, o mercado imobiliário pode ser classificado como de concorrência imperfeita. Neste caso, o preço pago pelos imóveis – elemento que pode ser observado – pode diferir (e freqüentemente afasta-se bastante) do valor de mercado.

Em função da imobilidade dos bens, há uma forte relação dos preços com a localização, ou seja, as características de acessibilidade e de qualidade de vizinhança são fundamentais para o entendimento das variações de preços no mercado imobiliário. Porém não há medidas padronizadas e o avaliador geralmente utiliza estimativas subjetivas, baseadas na sua experiência, com dificuldades de generalização, detalhamento e justificação[7].

Embora a obtenção do valor de mercado seja a tarefa central nas avaliações, a análise geral do mercado é fundamental para atingir este objetivo, e também pode ser necessária para embasar decisões do cliente. É importante incorporar explicitamente nas avaliações o conhecimento sobre o mercado, na forma de modelos gerais, identificação de tendências ou relações entre tipos de imóveis, o que poderá também reduzir o erro das estimativas.

---

[6] Em alguns países existem serviços de coleta e disseminação de informações, e facilmente podem ser obtidas massas de dados, para pesquisa ou construção de sistemas de avaliações (Bonissone *et al.*, 1998; Dubin, 1998; Wendt, 1974).

[7] Algumas destas alternativas foram relatadas no item sobre autocorrelação, apresentado adiante (item 2.3.4.5.2a).

1.3 QUESTÕES DE PESQUISA

A análise da situação atual permite concluir que o processo, mesmo com a adoção da regressão, ainda conta com forte influência do avaliador, e sofre também com questões relacionadas com os dados, afetando algumas condições da análise estatística e a precisão das estimativas. Pode-se concluir que os problemas apontados ainda demandam esforço para o desenvolvimento, e a questão de pesquisa formulada é a seguinte:

- Como aprimorar as avaliações de imóveis, diminuindo os erros nas estimativas do valor de mercado?

As alternativas para o aperfeiçoamento das avaliações basicamente são: (a) aperfeiçoar o processo atual, melhorando as condições para a aplicação da regressão e (b) alterar o paradigma de análise. Assim, a questão de pesquisa foi subdividida em duas partes:

- como aumentar a qualidade dos dados, visando aprimorar a modelagem?
- quais as técnicas mais adequadas para a modelagem, ou seja, quais podem diminuir o erro de estimação, em relação à regressão?

Os problemas apontados no processo atual de avaliação podem ser abordados como um problema de descobrimento de conhecimento em bases de dados (DCBD)[8], com a fase de desenvolvimento de modelos dos dados composta por técnicas das áreas de estatística e inteligência artificial. O sucesso em outras áreas do conhecimento sugere a possibilidade de aperfeiçoamento também na análise do mercado imobiliário, através do tratamento adequado dos dados e de técnicas de estimação que proporcionem modelos mais precisos (Berry e Linoff, 2000; Cordón *et al.*, 2001; Weiss e Indurkhya, 1998). Entretanto, o conjunto de técnicas disponíveis é extremamente variado, em seus pressupostos e potenciais aplicações. A revisão bibliográfica preliminar indicou que a escolha da técnica mais apropriada para cada caso em parte está vinculada ao tipo de análise e às peculiaridades dos dados, reforçando a importância de utilizar conhecimento do domínio no processo de definição das etapas e de escolha das ferramentas. Segundo alguns autores, tais como Berry e Linoff (2000), Bruha

[8] *Knowledge Discovery in Databases* (KDD). Ver capítulo 3.

(2001), Sheppard (1999) e Wolpert e Macready (1995, 1997), não existe uma técnica superior a todas as demais para qualquer tipo de aplicação, mas apenas indicações de aplicabilidade em função das características de cada uma, exigindo o teste empírico em cada área.

1.4 OBJETIVOS

Tendo em vista a importância da correta estimativa dos valores, as dificuldades encontradas presentemente e a disponibilidade de alternativas tecnológicas, este trabalho tem como objetivo geral propor uma nova abordagem para as avaliações de imóveis pelo método comparativo de dados de mercado, de forma a obter maior objetividade e precisão nas estimativas e explorando alternativas nas áreas de descobrimento de conhecimento em bases de dados e inteligência artificial. O trabalho tem como objetivos específicos:

a) identificar as técnicas potencialmente mais adequadas para preparação dos dados e para desenvolvimento de modelos de avaliação de imóveis, selecionando um subconjunto de técnicas para a análise empírica;

b) propor uma seqüência de tratamento de dados para o mercado imobiliário, gerando uma base de dados adequada para a fase de modelagem;

c) propor formas de seleção de casos e atributos para os diferentes tipos de avaliações;

d) comparar os resultados das técnicas de modelagem alternativas com os resultados obtidos através da análise de regressão e comparar os resultados obtidos com cada tipo de avaliação, verificando se há indicações de desempenho superior de uma ou mais técnicas em cada caso.

1.5 PROPOSIÇÕES

De uma forma geral, entende-se que é possível aprimorar as avaliações, ou seja, aumentar objetividade e a precisão das estimativas em relação à análise de regressão, utilizando algumas técnicas das áreas de descobrimento de conhecimento e inteligência artificial. Além

disto, tendo em vista a divisão das avaliações em três formatos básicos (análise geral, tributação e avaliação individual), supõe-se que :

a) é necessário adaptar o processo de avaliação de imóveis conforme o tipo de avaliação (segundo os requisitos de cada um dos três), mas o processo geral permanece fundamentalmente o mesmo;

b) os três tipos de avaliação de imóveis citados podem ser desenvolvidos com a mesma base de dados, através da utilização de critérios adequados para a seleção de casos;

c) os três tipos de avaliação podem ser desenvolvidos com as mesmas técnicas, sem degradação significativa de desempenho em função do aumento ou diminuição do tamanho da amostra.

1.6 ESCOPO E LIMITAÇÕES DA ANÁLISE

As técnicas examinadas fazem parte das áreas de descobrimento de conhecimento ou inteligência artificial. Em função da diversidade e do potencial de aplicação destas técnicas, não foram incluídas outras alternativas, tais como estatística espacial, estatística não-linear, sistemas de informações geográficas e regressão multi-nível.

A análise empírica foi desenvolvida com dados obtidos a partir de guias do imposto de transmissão (ITBI), registradas na Secretaria Municipal da Fazenda de Porto Alegre. A possível influência nos resultados devido às peculiaridades desta fonte de dados não foi investigada em profundidade. A extensão dos resultados a outros tipos de imóveis, tais como imóveis comerciais, ou de outros locais, pode apenas ser sugerida, com base na experiência anterior e nas evidências apresentadas na literatura consultada. Assim, as conclusões e os resultados obtidos empiricamente devem ser observados como exemplos específicos, resultantes da seqüência de análise, do conhecimento empregado, das peculiaridades dos dados e dos objetivos da análise.

Efetivamente, o trabalho desenvolvido não se trata de uma aplicação completa de

descobrimento de conhecimento, em função de limitações nos dados utilizados na análise empírica, pois não há busca e descobrimento de "padrões completamente novos" na fase de modelagem, consistindo na seleção de atributos e casos relevantes, e na busca de coeficientes e formatos mais adequados para modelos hedônicos de preços, pois há uma expectativa inicial sobre atributos e formatos, em função do conhecimento disponível sobre o mercado (Fayyad *et al.*, 1996a). Contudo, a abordagem utilizando a estrutura do descobrimento de conhecimento em bases de dados é potencialmente útil para outras aplicações no âmbito do mercado imobiliário, tais como previsão do prazo de vendas ou rentabilidade, elementos de grande interesse para os agentes do mercado. Assim, entende-se que o desenvolvimento da análise usando a mesma seqüência é útil para aplicações futuras.

Por fim, é importante ressaltar que não se busca o desenvolvimento ou aperfeiçoamento das técnicas em si, mas a aplicação das técnicas existentes a um domínio específico (o mercado imobiliário).

1.7 CONTRIBUIÇÕES

O trabalho tem a intenção de contribuir para a investigação e utilização de técnicas das áreas de descobrimento de conhecimento e de inteligência artificial no âmbito do mercado imobiliário, tendo em vista que se trata de um conjunto poderoso de técnicas, cujo desenvolvimento é bastante recente e que são relativamente desconhecidas nas Escolas de Arquitetura e Urbanismo e Engenharia Civil brasileiras[9].

Mais especificamente, as contribuições esperadas deste trabalho são as seguintes:

a) diminuir a subjetividade do processo de avaliação de imóveis através da análise sistemática e objetiva dos dados;

b) aprimorar a precisão ou confiabilidade das estimativas pelo uso de técnicas mais robustas às características do mercado imobiliário;

[9] No Brasil, os Engenheiros Civis e Arquitetos Urbanistas são os profissionais habilitados legalmente para avaliações de imóveis urbanos, enquanto que os Engenheiros Agrimensores e Agrônomos respondem pelos imóveis rurais.

c) estimular a geração de modelos gerais, que são úteis também para outras situações, como teste de hipóteses ou análise do comportamento do mercado;

d) apresentar um modelo de aplicação de descobrimento de conhecimento para o mercado imobiliário tendo em vista futuras aplicações de descobrimento de conhecimento que envolvam outros aspectos do mercado.

1.8 MÉTODO DE PESQUISA

A pesquisa foi desenvolvida em quatro etapas. Inicialmente foi desenvolvida uma ampla revisão bibliográfica sobre o mercado imobiliário, sobre o processo de descobrimento de conhecimento em bases de dados e sobre as técnicas de inteligência artificial, com o levantamento das principais técnicas utilizadas, dentro e fora do âmbito do mercado imobiliário, examinando suas potencialidades para as tarefas envolvidas na avaliação de imóveis.

A revisão bibliográfica sobre as técnicas foi acompanhada paralelamente de uma análise empírica exploratória e incremental das técnicas. Durante a revisão bibliográfica foram realizados testes através da aplicação de dados de pesquisas anteriores em *softwares* específicos. Nesta etapa foi examinada uma técnica por vez, dedicando-se atenção aos detalhes de entrada (alimentação dos dados e parâmetros requisitados) e saída (análise dos resultados) das técnicas, bem como as aplicações mais freqüentes para cada uma delas na literatura.

Na terceira etapa, foi desenvolvida uma análise geral dos resultados da etapa anterior, com cruzamento de informações, identificando-se as vantagens e desvantagens de cada técnica, em geral e considerando o contexto do domínio Mercado Imobiliário. As características do domínio e as deficiências identificadas na análise de regressão influenciaram a definição das técnicas a serem utilizadas nas etapas posteriores (geração da base de dados e dos modelos preditivos).

Com estes elementos, foi formulada a proposta de uma nova abordagem para a avaliação de imóveis. Em seguida foram coletados dados do mercado imobiliário e desenvolvidas as

aplicações, finalmente com a comparação dos resultados.

Esta forma de análise foi adotada em função do relativo desconhecimento inicial sobre algumas das técnicas por parte do autor. Assim, era necessária uma aprendizagem incremental, com redefinições periódicas em níveis progressivamente mais amplos de conhecimento sobre as técnicas exploradas antes da formatação da proposta de uma nova abordagem. Os dados e resultados da análise exploratória inicial não estão descritos nesta tese, apresentando-se apenas as aplicações finais.

1.9 ESTRUTURA DO TRABALHO

O trabalho foi estruturado em oito capítulos. O presente capítulo apresenta o tema, indicando a relevância das avaliações, os problemas com as técnicas atuais e enuncia as soluções a serem investigadas. Os demais capítulos são os descritos a seguir.

O Capítulo 2 compreende a revisão da literatura sobre os aspectos fundamentais do mercado imobiliário e dos métodos adotados atualmente para avaliação de imóveis, com destaque para a análise de regressão, explicitando-se os problemas decorrentes da ruptura de pressupostos básicos.

O Capítulo 3 trata do descobrimento de conhecimento em bases de dados, apresentado também detalhes sobre a preparação dos dados e sobre as tarefas da fase de mineração de dados.

O Capítulo 4 reúne a revisão de literatura sobre algumas técnicas alternativas, potencialmente úteis para as avaliações, incluindo técnicas das áreas de estatística e inteligência artificial.

No Capítulo 5, apresenta-se a proposição de uma nova abordagem para as avaliações, propondo os passos de análise e preparação dos dados e as formas de seleção de casos para cada tipo de avaliação. Por fim, apresenta-se o grupo de técnicas de modelagem alternativas a serem exploradas empiricamente.

O Capítulo 6 apresenta o desenvolvimento da base de dados, incluindo a coleta e a preparação dos dados, incluindo identificação de *outliers* e a seleção de atributos relevantes.

O Capítulo 7 apresenta o desenvolvimento dos modelos preditivos, iniciando pela seleção de casos para a avaliação de massa e individual. Para tanto, foi apresentado um caso-exemplo para a avaliação individual. O trabalho seguiu com a modelagem e os testes para cada técnica selecionada, finalmente com a análise dos resultados e comparação das técnicas.

Por fim, o Capítulo 8 apresenta as principais conclusões do trabalho, acrescidas de sugestões para trabalhos futuros, relacionadas à área de avaliações.

2 MERCADO IMOBILIÁRIO E TÉCNICAS DE AVALIAÇÃO

2.1 CONSIDERAÇÕES INICIAIS

Em termos gerais, o mercado é uma forma de coordenação da atividade econômica, que busca o equilíbrio através de um mecanismo de preços. Como existem distorções que impedem este equilíbrio, geralmente existem também alguns mecanismos de planejamento, tais como a regulamentação através de legislação, definidos pelas várias esferas de governo (Harvey, 1996; Lange, 1985). A expressão "mercado imobiliário" refere-se a um mercado abstrato, que agrega diversos segmentos. Podem ser identificadas parcelas que constituem sub-mercados, com funcionamento diferenciado em função das localizações, dos tipos de imóveis ou das formas usuais de transação. Contudo, os limites de cada segmento não são claros e muitas vezes há interpenetração entre estes sub-mercados. De qualquer forma, o mercado imobiliário geralmente realiza, ainda que de forma imperfeita, as funções de um mercado, embora existam autores que afirmem não existir um mercado imobiliário, no sentido estrito do termo (Balchin *et al.*, 1995; Evans, 1995; Harvey, 1996; Lavender, 1990; Maclennan, 1977; Pindyck e Rubinfeld, 2002).

Dentro do mercado imobiliário, a parcela habitacional atrai consideravelmente mais atenção, em função da importância social e do volume de recursos envolvido. A discussão a seguir concentra-se no mercado habitacional, mas se aplica aos outros segmentos em vários aspectos. Algumas características especiais diferenciam a habitação de outros bens, tais como imobilidade, heterogeneidade, durabilidade, custo elevado e influência governamental. Embora haja influência simultânea, estes aspectos podem ser divididos basicamente em três grupos, analisando-se separadamente as questões relativas à localização, às características das construções e ao governo. Neste capítulo, são examinadas algumas questões sobre o funcionamento do mercado e discute-se o conceito de valor de mercado. Por fim, discute-se as práticas correntes de avaliação de imóveis.

2.2 CARACTERÍSTICAS DO MERCADO IMOBILIÁRIO

2.2.1 Características locacionais

Uma das características fundamentais do mercado imobiliário é a de que há imobilidade da oferta e ao mesmo tempo a demanda é localizada espacialmente. A demanda é definida de acordo com condições próprias de cada local, em função dos padrões de renda, nível de emprego, preferências da população e outros fatores, sendo significativamente diferente de um local para outro. Ademais, cada local tem uma diferente oferta de serviços públicos, empregos, amenidades e externalidades negativas, gerando níveis de qualidade (ou desejabilidade) distintos. Por outro lado, o excedente de oferta em uma região não pode ser deslocado para compensar a falta em outros locais, em virtude da imobilidade (Balchin e Kieve, 1986; Harvey, 1996; Lucena, 1985).

A qualidade de localização de um imóvel pode ser explicada pelas características de acessibilidade e pelos padrões de uso na vizinhança. A importância da localização é enfatizada por diversos autores (Anselin, 1998; Balchin e Kieve, 1986; Derycke, 1971; Harvey, 1996; Lavender, 1990; Muth, 1975; Robinson, 1979). Entretanto, na prática há dificuldades de medição dos efeitos da localização, pois acessibilidade e vizinhança não contam com medidas padronizadas. Para a acessibilidade, os modelos mais comuns de análise das áreas urbanas consideram apenas um pólo de atração, chamado de distrito central de negócios (*Central Business District* - CBD). Estes modelos pressupõem que o CBD concentra as funções urbanas essenciais e a maioria dos empregos (Derycke, 1971; Muth, 1975). Esta simplificação pode ser exagerada, pois o crescimento das cidades tende a gerar estruturas mais complexas, com múltiplos centros de atração (Can, 1990; Dubin, 1992; Dubin e Sung, 1987; Wyatt, 1996a, 1996b). Para Straszheim (1987), a deficiência do modelo monocêntrico é a desconsideração dos efeitos das características de vizinhança na decisão de localização. Efetivamente, muitos estudos empíricos que usam a distância ao CBD como medida de acessibilidade encontram pouca importância estatística para esta variável, e há várias medidas alternativas (Ball, 1973; Bartik e Smith, 1987; Dubin e Sung, 1987; Smith *et al.*, 1988).

Os efeitos da vizinhança são igualmente importantes e difíceis de medir. Há uma tendência de que as densidades de usos e os padrões de construção sejam semelhantes para o mesmo local

e época. Existem estudos demonstrando os efeitos de diversos fatores, tais como padrão dos imóveis vizinhos (ambiente construído), grau de escolaridade e renda dos residentes no entorno, qualidade do ar, disponibilidade de escolas e transporte público, proximidade de depósitos de lixo e usinas nucleares (Ball, 1973; Boyle e Kiel, 2001; Din *et al.*, 2001; Jud e Watts, 1981; Kain e Quigley, 1970; Lang e Jones, 1975).

Não há um consenso na literatura sobre as medidas mais apropriadas para acessibilidade e vizinhança (Ball, 1973; Can, 1990). Por outro lado, propriedades similares e próximas tendem a apresentar um valor de mercado semelhante, ou seja, a imobilidade produz um "valor de localização" e esta semelhança tende a diminuir com o aumento da distância que os separa. Em particular, Ding *et al.* (2000) demonstraram que a renovação de um conjunto de imóveis pode valorizar os imóveis próximos, nos quais não foram realizados investimentos, apenas em função da melhoria na vizinhança. Portanto, é razoável supor que o nível dos preços de um imóvel seja influenciado pelos imóveis circundantes. Segundo Can (1998), espera-se que os preços dos imóveis variem sistematicamente ao longo da área urbana. Aparentemente, estas variações são contínuas, isto é, os valores não surgem de forma aleatória, e podem ser mapeados a partir de dados do mercado, usando as ferramentas adequadas, tais como superfícies matemáticas ou geoprocessamento, através de duas abordagens. Uma delas baseia-se na identificação e posterior modelagem dos erros em modelos sem variáveis de localização, gerando uma nova variável. Outra corrente constrói ou refina as variáveis de localização utilizando medidas objetivas aos pólos de interesse (Dubin, 1992; Gallimore *et al.*, 1996; González *et al.*, 2002a; Li e Brown, 1980; McCluskey *et al.*, 2000; Wyatt, 1996a).

2.2.2 Características do próprio imóvel

Também devem ser considerados aspectos ligados ao próprio imóvel e sua tecnologia de construção, tais como heterogeneidade, durabilidade e custo elevado. Existe uma grande variedade de produtos no mercado imobiliário. Entre outros elementos, os imóveis têm grandes diferenças em tamanho, idade e qualidade de construção, as quais são refletidas através de variações nos preços de mercado. A heterogeneidade dos imóveis e de suas localizações dificulta a comparação, pois a informação sobre os vários atributos nem sempre está disponível aos agentes (Lavender, 1990; Robinson, 1979).

Os imóveis contam com elevada vida útil, na maioria dos produtos. A durabilidade dos imóveis faz com que a maioria das transações na área urbana seja composta por unidades pré-existentes. Este fator é importante, porque diminui os efeitos das novas construções, tornando lentas as transformações urbanas e vinculando as transações do presente às decisões do passado. Como conseqüência, os preços praticados pelo mercado são definidos pelos níveis de preços dos imóveis usados, e os imóveis novos adaptam-se a estes preços (Balchin e Kieve, 1986). Ademais, o custo das unidades é muito elevado. Robinson (1979) afirma que os preços dos imóveis representam dois ou três anos de salário, exigindo financiamento para a aquisição, geralmente oferecido pelo Estado e com amplos prazos de pagamento. Em muitas economias, a residência representa o bem mais valioso adquirido pela maioria dos indivíduos, bem como uma expressiva fatia dos gastos familiares, e tem importância crucial na análise do nível de bem-estar da população (Balchin *et al.*, 1995; Lucena, 1985; Maricato, 1987, 1997; Sheppard, 1999).

2.2.3 Influência governamental e da legislação

A influência das diversas esferas de governo pode ser sentida na oferta de infra-estrutura, no sentido físico e legal, bem como na condução da economia. Os imóveis estão sujeitos às influências dos governos e das economias local, regional, nacional e global. Por sua importância e significação social, as leis geralmente propiciam tratamento diferenciado aos imóveis, com respeito às condições de uso e transferência de propriedade, com o intuito de garantir ou proteger os direitos individuais e coletivos. O poder público também tem influência decisiva nas alterações de uso e ocupação do solo, através de intervenções diretas (tais como abertura ou alargamento de vias urbanas) e pelo controle ou incentivo à atuação da iniciativa privada (através de planos diretores de desenvolvimento), alterando o comportamento do mercado imobiliário local. Além disto, existe influência macroeconômica do comportamento do Governo, na oferta de crédito e na condução da economia nacional (Balarine, 1996; Robinson, 1979; Rovatti,1990; Sheppard, 1999).

Os imóveis contam, historicamente, com uma regulamentação especial. O direito de propriedade é entendido como um "feixe de direitos" englobando os direitos de posse, alienação, uso, usufruto e reivindicação. Balchin *et al.* (1995) e Harvey (1996) ressaltam que não é o próprio imóvel que é transacionado, mas sim os direitos sobre o imóvel, e há várias

condições especiais regulando as transações para este tipo de bem.

O direito brasileiro fundamenta-se nos institutos do direito romano, mas alguns conceitos têm evoluído, como é o caso da função social da propriedade, relativizando o conteúdo absoluto do direito de propriedade (Chemeris, 2002; Ihering, 1957; Meirelles, 1996; Sciascia, 1952). A função social é uma visão contemporânea, que foi defendida na Constituição de 1988 e teve alguns aspectos recentemente regulamentados através do Estatuto da Cidade (Lei 10.257/01), com prováveis impactos sobre o mercado imobiliário. Outras influências da legislação abrangem o controle de aluguéis, que geralmente provoca alterações significativas no mercado imobiliário (Balarine, 1996; Balchin e Kieve, 1986; Marks, 1984; Olsen, 1972) e a regulamentação de uso do solo urbano, incluindo o zoneamento de uso e os planos diretores de desenvolvimento (Gondim, 1995; Pogodzinski e Sass, 1991; Rovatti, 1991; Salengue e Marques, 1993). Por fim, a tributação imobiliária é importante fonte de recursos para muitos municípios brasileiros, e também tem sido utilizada para finalidades extrafiscais, de auxílio à gestão urbana (Abramo, 2001a, 2001b; Cenecorta e Smolka, 2000; Fernandez, 2001; Furtado, 1997; George, 1972; Leal, 1990; Smolka e Furtado, 2001; Rolnik e Cymbalista, 1997) .

2.2.4 Funcionamento do mercado imobiliário

O mercado é geralmente informal, e é difícil distinguir como os agentes obtêm informações sobre as transações. Não há controle, nem ao menos registro de muitas transações, em função de questões como a evasão de tributos ou taxas de transferência, os quais podem atingir um montante de custos significativo (Balchin e Kieve, 1986; Evans, 1995; Robinson, 1979). O mercado não é "transparente", segundo Derycke (1971). Há custos de mobilidade, no sentido de custos financeiros, tempo e desgaste psicológico, dispendidos na busca pelo imóvel desejado. Em função do comportamento localizado de oferta e demanda, os agentes precisam obter informação específica sobre cada segmento de interesse, com os custos respectivos. E existem diversos outros gastos, tais como taxas de intermediação do negócio e a própria mudança (física) de um imóvel para outro. Estes custos criam barreiras à livre opção e dificultam as decisões de compra e venda (Bell, 1999; Evans, 1995; Gau, 1987; Kinnard, 1966, 1971; Robinson, 1979).

A oferta é relativamente fixa no mercado imobiliário por causa da imobilidade e do prazo de

maturação das novas construções. Entre outros efeitos, preço e qualidade não se alteram simultaneamente em resposta às mudanças de demanda, e os ajustes de equilíbrio ocorrem principalmente no nível geral de preços. As mudanças na demanda afetam os preços inicialmente e, somente após um período de tempo, os empreendedores alteram a oferta, de forma descoordenada, às vezes gerando um excesso de oferta (Lavender, 1990; Rovatti, 1990; Straszheim, 1987).

O mercado é dinâmico, embora as variações sejam relativamente lentas. Uma parte da dinâmica imobiliária está associada com o processo de estruturação intra-urbana, o qual modifica continuamente a forma da cidade, alterando os usos do solo em tipo e densidade (Abramo, 1988; Balchin e Kieve, 1986; Maclennan, 1977; Smolka, 1994b). A realização de obras, tais como escolas, parques, avenidas, *shopping centers* ou indústria, introduz modificações não só no entorno próximo, mas em uma área de maior abrangência, bem como a atuação do capital imobiliário transforma o mercado, na medida em que o mesmo se desloca espacialmente na busca do lucro, desenvolvendo ciclos econômicos espaciais no interior da área urbana (Abramo, 1988; Lucena, 1985; Maraschin, 1993). Também podem ser identificados ciclos econômicos de longo prazo no mercado imobiliário, envolvendo, entre outros fatores, as expectativas não racionais dos compradores e vendedores sobre o comportamento futuro do próprio mercado. Alguns autores citam uma média ou tendência de equilíbrio, chamada de "valor intrínseco" ou "valor fundamental" do mercado, para o qual o mercado deveria convergir, embora possa existir uma tendência permanente de crescimento dos preços, especialmente para o caso dos terrenos (Case, 2000; Case e Shiller, 1987; Clayton, 1998; Derycke, 1971; Nordvik, 1995).

É comum supor a existência de concorrência perfeita em alguns tipos de análise microeconômicas. Aceitar a perfeição de um mercado significa, em termos gerais, admitir que os bens podem ser considerados idênticos, que a entrada no mercado é livre, que as pessoas têm informação perfeita, decidem livre e prudentemente, sem pressões de qualquer ordem, e que as ações individuais não afetam os preços. Nestas condições, o valor do bem seria igual ao preço que atinge no mercado, e seria proporcional à quantidade adquirida (Balchin *et al.*, 1995; Lavender, 1990). Em um mercado com funcionamento ao menos próximo do ideal, há uma grande quantidade de agentes, de ação simultânea e não coordenada, e o número de transações anuais é significativo. Os preços são resultantes do balanço das variações de oferta e demanda

decorrentes do equilíbrio desta atividade (Balchin e Kieve, 1986; Goodall, 1972; Harvey, 1996). No mercado imobiliário, contudo, não é assim. Entre outros problemas, os bens são heterogêneos e há significativa falta de conhecimento. Assim, o mercado imobiliário funciona em regime de concorrência imperfeita (Balchin *et al.*, 1995; Evans, 1995; Harvey, 1996; Robinson, 1979).

2.2.5 O conceito de valor no mercado imobiliário

Vários autores afirmam que o conceito de valor é fundamental na execução de uma avaliação, definindo os métodos e os dados a serem utilizados. Portanto, é necessário um conceito que permita nortear os trabalhos. Entretanto, há um debate histórico sobre o valor entre os avaliadores, o qual ainda não está encerrado. Segundo Bonright (1937), todas as definições de valor, se vistas criticamente, contém ambigüidades e invocam conceitos aceitáveis apenas em casos específicos, entre outros problemas. Para este autor, definir "valor" de uma forma geral é uma tarefa difícil, mas necessária para embasar as avaliações.

Há indicações de que o conceito dominante está vinculado ao contexto ético-filosófico-econômico vigente na sociedade de cada época. Para os gregos e romanos, o valor era uma característica intrínseca do bem. Na Idade Média, sob influência da Igreja, dominava a idéia de preço justo, também vinculada ao valor intrínseco (Myrdal, 1962; Wendt, 1974). Com o mercantilismo, alguns economistas inovaram, defendendo preços de mercado e diferenciando o valor de troca da utilidade do bem, enquanto que outros se mantiveram fiéis à corrente anterior. Na época moderna, John Stuart Mill defendeu o valor como indicação do poder de troca, enquanto que Marx entendia o valor como resultado unicamente do trabalho. Já Marshall afirmou que valor e preço coincidem esporadicamente, em situações de equilíbrio de mercado, mas geralmente diferem bastante (Ring, 1965; Wendt, 1974). Na Depressão norte-americana, as fortes oscilações e a instabilidade do mercado imobiliário fizeram com que os avaliadores questionassem a existência de valores no curto prazo, apontando novamente para a existência de "valores intrínsecos" ou "reais", válidos a longo prazo (Ring, 1965; Weimer, 1953).

Os conceitos gerais tendem a se fixar em duas correntes ou formas de valor: valor de uso e valor de troca. O primeiro está vinculado a uma visão subjetiva, definindo o valor que o investimento tem para um indivíduo, enquanto que o segundo é o valor de mercado, que é

objetivo, pois é uma visão coletiva, mais geral (Kinnard, 1971; Wendt, 1974; White, 1950).

Segundo Fiker (1993), o valor é a relação entre a intensidade das necessidades econômicas do homem e a quantidade de bens disponíveis para satisfazê-los, sendo determinado pela relação entre oferta e procura do bem. Para White (1950), o valor reflete a importância relativa de um bem escasso, e não há um valor intrínseco ou inerente à coisa. O valor é derivado da comparação com muitas mercadorias, entre as quais as escolhas de alternativas são possíveis. Quando estas mercadorias são trocadas livremente no mercado, as taxas de troca definem os valores de cada uma. Em uma economia monetária, os bens são comprados com dinheiro, estabelecendo assim um nível de preços. Fernandes (1983) também entende que o valor não é uma propriedade intrínseca do bem, mas uma característica definida pelo mercado, resultante da oferta e da procura, e afirma que é único em um momento considerado.

Kinnard (1971) afirma que o valor é um preço que tende a prevalecer mesmo com variações nas condições de mercado, como resultado da interação das forças de oferta e demanda. Este preço reflete a capacidade de um bem econômico como poder de troca (valor de troca, estimado pelo valor de mercado). O valor também pode ser visto como o valor presente dos benefícios futuros antecipados ou previstos para serem recebidos (valor de uso, estimado pelo valor de investimento). Se os compradores e vendedores forem informados e racionais, não venderão por menos, nem comprarão por mais do que o valor presente dos benefícios futuros esperados, e os preços devem igualar-se ao valor de uso e ao valor de troca. Porém, como há imperfeições no mercado, preços e valores são diferentes. Neste caso, estimar o preço mais provável é conveniente, pois é menos rígido do que o valor de mercado tradicional.

Nesta visão, o valor é um conceito objetivo, não vinculado a um comprador ou vendedor particular, enquanto que os preços refletem tendências ou vantagens de uma parte sobre a outra, presentes nas negociações individuais (White, 1950). Albritton (1982) afirma que geralmente as partes concluem a transação após considerável espaço de negociação, com ofertas e contra-ofertas sucessivas. O preço final reflete as condições de cada parte, e é afetado pelas habilidades de cada uma, ou de seus intermediários. O preço é entendido como um fato histórico, ou a quantia efetivamente praticada na transação, enquanto que o valor de mercado é um elemento potencial, uma tendência, que pode ou não se realizar. É importante ressaltar que a avaliação é uma estimativa do valor, e não a predição do preço (Kinnard, 1971).

Há algum tempo, dominava a visão de que o valor de mercado deveria ser estimado como o maior preço que um imóvel poderia atingir, em uma transação normal, com os agentes dispondo de informação adequada (McMichael, 1962; Pelegrino, 1983; Wendt, 1974). Albritton (1982) argumenta que, se há suposição de que comprador e vendedor estão bem informados, o maior preço não é mais lógico do que o menor preço. Assim, a expressão passou a ser entendida como "maior preço que um comprador bem informado pagaria", e foi progressivamente substituída pela idéia de preço mais provável. Porém, ainda há discussão sobre a adequação desta relação (Kummerow, 2000; Smith, 1995).

Ventolo e Williams (1997) afirmam que o valor de mercado de um imóvel é o preço mais provável que um comprador está disposto a pagar a um vendedor pelo imóvel em uma operação normal de mercado. Bell (1999, p.49) indica uma definição mais completa. Para este autor, o valor de mercado significa o preço de venda mais provável pelo qual um imóvel seria vendido em um mercado aberto e competitivo, sob todas as condições de uma venda justa, com comprador e vendedor agindo prudentemente, com conhecimento suficiente, assumindo que o preço não é afetado por estímulos indevidos. Nesta definição está implícita a consumação da venda a uma determinada data e a passagem do título do vendedor para o comprador sob as seguintes condições: (a) comprador e vendedor estão tipicamente motivados; (b) ambas as partes estão bem informadas ou bem assessoradas, agindo de acordo com seus melhores interesses; (c) há um tempo razoável de exposição no mercado; (d) o pagamento é feito em termos monetários, à vista ou em arranjos monetários equivalentes; e (e) o preço representa a condição normal de venda, ou seja, o imóvel vendido não foi afetado por formas especiais de desconto ou de financiamento.

É conveniente ressaltar também que o valor de mercado é único, para um determinado momento e situação de mercado. Esta teoria é defendida pela "escola univalente", em oposição à outra corrente, que afirmava que existiam diversos valores e que o valor dependia da finalidade da avaliação. Evidentemente, se o valor de mercado não é característica intrínseca da coisa, pode variar, mas, considerado em dado momento, será único, pois só há um mercado para aquele bem. O preço, ao contrário, é múltiplo, variando em uma faixa, de acordo com os agentes da negociação, e também está sujeito a flutuações de curto prazo, tais como as decorrentes de situações econômicas, campanhas publicitárias, novos empreendimentos ou perspectiva de alterações na legislação.

Como exemplo da corrente plurivalente, podem ser citados Marston e Agg (1936), McMichael (1962) e Meirelles (1996). Entretanto, Meirelles, mesmo listando diversos tipos de valores, afirma que "...usualmente se busca, nas avaliações judiciais, o valor de mercado, valor de venda, valor venal, relegando-se os outros valores para situações especiais" (Meirelles, 1996, p.291).

Pode-se concluir que o valor de um imóvel pode ser identificado pelo seu valor de mercado, que é o valor médio ou preço mais provável a ser atingido em transações normais, em um dado momento. O valor de mercado não necessariamente coincide com o preço, por causa da imperfeição do mercado, que provoca dificuldades nos julgamentos dos indivíduos, formando-se uma faixa de preços aceitáveis em torno do valor de mercado. De qualquer forma, em alguns casos, o valor único é necessário, como na avaliação para tributação e nas causas discutidas na Justiça. Em algumas ocasiões é necessário estimar uma faixa de valores razoáveis, como para orientação de investidores, os quais devem tomar suas decisões posteriormente. Em outros casos, é necessário estimar o preço de venda futuro.

2.3 ANÁLISE EMPÍRICA DO MERCADO IMOBILIÁRIO

O mercado imobiliário, como um setor da economia, pode ser analisado através das visões micro e macroeconômica. A microeconomia trata do comportamento dos indivíduos (pessoas, famílias ou firmas), envolvendo o estudo sobre a produção e consumo dos bens, equilíbrio de oferta e demanda, formação de preço e, em geral, maximização do bem-estar na sociedade. A macroeconomia trata dos mesmos interesses, porém a análise é desenvolvida de forma agregada, considerando produção, consumo, renda da população e outros elementos, a partir de uma visão do conjunto (Henderson e Quandt, 1976; Pindyck e Rubinfeld, 2002; Pinho e Vasconcellos, 1997).

Há pesquisas sobre o mercado imobiliário nos dois níveis. Em macroeconomia, há mais interesse na explicação da estrutura ou de tendências do mercado (Balarine, 1996; Case e Shiller, 1987, 1990; Lucena, 1985). Na visão microeconômica do mercado imobiliário, os modelos hedônicos de preços são amplamente utilizados em pesquisas sobre o comportamento do mercado, com vários enfoques ou temas principais. Existem um razoável

arcabouço teórico sobre a construção de modelos de mercado. Estes estudos iniciaram com Court, na década de 30, e receberam também contribuições de Lancaster, nos anos 60. Mais recentemente, Griliches (1971) e Rosen (1974) aprofundaram as bases para a análise hedônica. Nos modelos hedônicos os bens são tratados como um "pacote de atributos" ou um conjunto de "serviços de habitação", reunindo as características importantes para os agentes. Assim, os preços dos imóveis podem ser compreendidos como a soma dos produtos das quantidades de cada um destes serviços pelos seus preços implícitos (Lucena, 1985).

Inicialmente não são conhecidas as importâncias relativas (participações no preço) de cada uma das características contidas no pacote, sendo conhecido apenas o preço integral do imóvel. Os preços implícitos de cada um destes atributos, também chamados de preços hedônicos ou "preços-sombra", são os preços relacionados com cada um dos atributos dos imóveis, tais como área, idade e localização. Como as parcelas referentes a cada atributo não podem ser separadas, e não há mercados específicos para cada uma, os preços são obtidos indiretamente (Rosen, 1974; Sheppard, 1999).

2.3.1 Avaliação de imóveis

As avaliações de imóveis são análises sobre o mercado imobiliário desenvolvidas no âmbito da microeconomia, utilizando métodos próprios da Engenharia de Avaliações e também técnicas econométricas tradicionais, como a análise de regressão. Mais especificamente, a avaliação de imóveis pode ser definida como a determinação técnica do valor de um imóvel ou de um direito sobre este imóvel (ABNT, 1989; Fiker, 1993; Moreira, 1997).

Segundo Marston e Agg (1936), as pessoas realizam avaliações todo o tempo, mesmo sem perceber. Quando compram, ou mesmo quando comparam preços de algum bem, estão fazendo uma avaliação simplificada. Para esse autor, a Engenharia de Avaliações é a arte da estimação dos valores justos no caso em que conhecimento e julgamento profissional são aplicados. Este autor afirma ainda que a Engenharia de Avaliações surgiu por volta de 1890, em função das novas necessidades do mundo industrializado.

Os objetos de uma avaliação podem ser terrenos para habitação, comércio ou outras atividades, glebas urbanizáveis, casas, apartamentos, salas comerciais ou prédios industriais. Para que seja feita uma boa avaliação, o profissional deve conhecer não só as técnicas

envolvidas no cálculo, mas também o funcionamento do mercado onde se situa o imóvel. Existem vários tipos de avaliações, em função do objetivo principal do trabalho, bem como diferentes requisitos, tanto do ponto de vista do trabalho em si quanto em relação aos profissionais (Abunahman, 1999; Dantas, 1998; Mendonça, 1999; Moreira, 1997).

2.3.2 Tipos de avaliações

As avaliações podem ser classificadas em individuais ou coletivas (também chamadas de avaliações de massa), em função do objeto. A avaliação individual de imóveis é uma estimativa ou opinião profissional sobre o valor de um imóvel adequadamente descrito, a uma data específica, apoiada pela apresentação e análise de dados relevantes (Dowse, 2000; Smeltzer, 1986[10], *apud* McCluskey *et al.*, 1997). Já a avaliação de massa pode ser definida como a avaliação sistemática de grupos de imóveis a uma determinada data, usando procedimentos padronizados e testes estatísticos (Eckert, 1990[11], *apud* McCluskey *et al.*, 1997). Há ainda os modelos mais gerais, com finalidade de pesquisa ou teste de hipóteses sobre o mercado, os quais geralmente são voltados para a descrição do mercado, e não para a estimação de valores. Contudo, o formato e o desenvolvimento destes modelos são similares, com diferenças em termos de forma de apresentação e detalhamento. Conta apenas com uma etapa a menos, pois a elaboração do próprio modelo é o objetivo do trabalho.

Para McCluskey *et al.* (1997), as avaliações individuais e de massa diferem apenas em termos de escala, pois o objetivo final é o mesmo: uma avaliação precisa do valor de uma ou mais propriedades. Os métodos também são essencialmente os mesmos, com algumas diferenças na análise de mercado e no controle de qualidade. Segundo estes autores, a avaliação de massa surgiu pela necessidade de uniformidade e consistência nas avaliações quando há grande quantidade de imóveis a avaliar (centenas ou milhares de casos, geralmente).

Na avaliação individual o método mais utilizado é o método comparativo de dados de mercado. Este método consiste de uma inspeção razoavelmente detalhada do imóvel, busca e

[10] SMELTZER, M. V. The application of multi-linear regression analysis and correlation to the appraisal of real estate. **Appraisal Review**, v.28, 1986.

[11] ECKERT, J. K. **Property appraisal and assessment administration**. Chicago: IAAO, 1990.

seleção de propriedades similares, ajustamento dos preços e documentação da avaliação através de um formulário, relatório ou laudo. Em alguns países, como os Estados Unidos, é comum a avaliação baseada apenas em três dados similares. Também é comum a conjugação de métodos, unindo as estimativas obtidas através dos métodos do custo de reprodução, da renda e da comparação de dados (Detweiler e Radigan, 1996; Gelbtuch *et al.*, 1997; Lentz e Wang, 1998; Waller, 1999).

Na avaliação coletiva busca-se derivar um modelo representativo do mercado, o qual deve refletir a teoria da avaliação e o comportamento do mercado, estando firmemente vinculado à teoria micro-econômica. É uma análise mais geral e normalmente há apenas um relatório geral ou um formulário gerado automaticamente pelo sistema (McCluskey *et al.*, 1997). Em vista das dificuldades na avaliação em massa os avaliadores recorreram à inferência estatística e à tecnologia da informação. Realizada desde os anos 30 nos Estados Unidos, a partir dos anos 50 a avaliação em massa passou a utilizar sistemas computadorizados, com modelos hedônicos calibrados com regressão múltipla e o contínuo desenvolvimento de computadores e programas aplicativos assegurou que a maioria dos setores de avaliação tributária tivessem acesso a sistemas de avaliação em massa relativamente baratos (Deddis *et al.*, 1998; Wachs, 1978).

Na avaliação individual, a qualidade pode ser medida pela comparação direta com vendas de imóveis comparáveis específicos. Por outro lado, na avaliação em massa, devido à quantidade de imóveis a serem avaliados, são utilizadas medidas estatísticas como o coeficiente de dispersão (COD) ou o coeficiente de variação (COV). Nos dois casos o avaliador precisa defender as avaliações, o que naturalmente é mais fácil na avaliação individual. De qualquer forma, o modelo utilizado precisa ser capaz de demonstrar como o valor foi atingido (IAAO, 1990; McCluskey *et al.*, 1997).

A avaliação baseada na comparação subjetiva é demorada e cara, requerendo diversos dias ou até semanas de trabalho para que o avaliador obtenha uma estimativa do valor (Waller, 1999). Segundo Detweiler e Radigan (1996), os avaliadores precisam reduzir os custos sem sacrificar a credibilidade. Mais ainda, se for possível simultaneamente aumentar a credibilidade e

diminuir os custos, o avaliador terá uma vantagem competitiva[12].

Para Eckert *et al.* (1993), as avaliações tradicionais ("manuais"), utilizadas para auxílio ao gerenciamento dos agentes de hipotecas nos Estados Unidos, proporcionavam bons resultados enquanto tinham um escopo local. O crescimento do mercado secundário de hipotecas tornou o sistema de âmbito nacional, exigindo o aperfeiçoamento das avaliações. Segundo Lenk *et al.* (1997), a atração por técnicas para avaliação em massa vem crescendo em função das pressões para diminuição de custos e prazos nas avaliações para análise de pedidos de financiamento. Waller (1999) relata que as duas maiores empresas do mercado secundário de hipotecas nos Estados Unidos têm recomendado o desenvolvimento e uso de sistemas automatizados.

A idéia de avaliação em massa tem recebido diversas denominações na literatura. A expressão mais comum é *Computer Assisted Mass Appraisal* (CAMA), mas os sistemas também são denominados de *Automated Valuation Models* (AVM) ou *Computer Assisted Real Estate Appraisal System* (CAREAS). A utilização mais comum para os modelos automatizados é em tributação imobiliária (McCluskey e Anand, 1999), mas há também largo uso na administração de carteiras de hipotecas ou de investimentos, com os objetivos de reduzir o tempo e o custo dos processos de tomada de decisão em hipotecas (Waller, 1999), revisar as avaliações realizadas por terceiros (Eckert *et al.*, 1993), avaliar lotes de imóveis na aquisição de carteiras de hipotecas (Bonissone *et al.*, 1998), ou na avaliação em geral, para aperfeiçoar as estimativas (Deddis *et al.*, 1998; McCluskey *et al.*, 1997).

Em resumo, podem ser identificadas três metas distintas para as avaliações, provavelmente resultando em diferentes formatos e graus de detalhamento dos modelos ou de exigências de homogeneidade dos dados empregados na análise: (a) análise de tendências de mercado; (b) avaliação em massa; e (c) avaliação individual.

O primeiro caso consiste de uma descrição ampla do mercado. Este tipo de modelo pode ser utilizado para auxiliar ou embasar o estudo de planos diretores de desenvolvimento urbano,

[12] A Caixa Econômica Federal, provavelmente a maior contratante de avaliações no Brasil, remunera as avaliações para instruir processos de financiamento por cerca de USD 40. Nos Estados Unidos, um trabalho similar é remunerado em cerca de USD 500, valor que é considerado um empecilho à utilização do formato tradicional em avaliação de massa, indicando a automação como uma necessidade (Bonissone *et al.*, 1998).

para análises de tendências ou de reconhecimento do mercado, ou ainda para auxiliar na preparação dos dados, selecionado atributos e casos relevantes. Nestas atividades, interessam mais os modelos do que as estimativas em si.

O segundo formato consiste da avaliação em massa, que é um pouco mais detalhada do que a anterior, com modelos do tipo empregado para tributação (planta genérica de valores), para desapropriações, ou para identificação dos valores em outras funções de gestão urbana (tais como o "solo criado"). Geralmente devem ser geradas estimativas para uma grande parte ou para todos os imóveis do cadastro municipal em um curto espaço de tempo, ou mesmo em uma única data de referência. Os modelos estão limitados ao uso dos atributos disponíveis na base de dados, em função de restrições de custo de coleta de informação. No caso do imposto sobre a propriedade (IPTU), os atributos são restritos aos constantes do cadastro municipal, sendo inviável a coleta de novos atributos, a não ser em momentos especiais (recastramentos plurianuais).

O terceiro tipo de avaliação consiste na avaliação individual de imóveis. Há um recorte do mercado e os modelos são "especializados" para certo tipo de imóvel e localização. Há necessidade de uma atenção especial sobre um determinado imóvel e, mesmo que eventualmente haja uma quantidade expressiva de imóveis a serem avaliados quase simultaneamente, como nos casos de avaliações para o imposto de transmissão ou para financiamento, há necessidade de modelos e estimativas individuais. A seleção de casos relevantes de forma ágil é uma tarefa importante. A avaliação pode ser desenvolvida com base nos dados pré-existentes (da base de dados), ou com auxílio de dados especialmente coletados, como dados oriundos de vistorias (inspeções *in loco*) a todos os imóveis da amostra. Esta tarefa é demorada e cara, não sendo viável a sua aplicação prévia para todo o banco de dados, devendo ser desenvolvida à medida do necessário. Assim, após a seleção dos casos relevantes, verifica-se se há necessidade deste tipo de complementação. A realização de vistorias pode ser uma exigência do cliente, mas também pode ocorrer que um volume maior de casos similares compense o menor detalhamento, gerando um modelo com o nível de precisão desejado e dispensando as informações de campo, sendo a vistoria realizada apenas no avaliando.

2.3.3 Requisitos a serem atendidos nas avaliações de imóveis

Basicamente as avaliações devem prever o valor de mercado, entendido como o preço de venda mais provável. Accetta (1999) fala em avaliações "convincentes", as quais devem enfatizar a localização dos dados (preferindo dados mais próximos aos mais recentes e mapeando os dados da amostra), apresentar fotografias dos dados da amostra, apresentar ofertas e vendas anteriores, reconhecer dados potencialmente úteis que não foram incluídos (explicando porque não foram incluídos), admitir fraquezas ou dificuldades encontradas no desenvolvimento do trabalho e usar cuidadosamente a experiência própria.

A avaliação individual deve ser objetiva, precisa, confiável e recente (Dowse, 2000; Newell e Kishore, 1998). Também deve ser consistente (a diferença entre avaliações realizadas por diferentes profissionais deve ser pequena) e com pequena variância (Parker, 1998).

Segundo McCluskey e Anand (1999), as avaliações de massa devem ter objetividade, economia de escala, eqüidade, justiça, defensibilidade, poder explicatório, transparência, facilidade de aplicação e precisão. Para Deddis *et al.* (1998) e McCluskey *et al.* (1997), os objetivos principais dos modelos CAMA são precisão, explicabilidade e estabilidade dos valores (no tempo). Geralmente estes modelos são segmentados em sub-modelos, divididos em grupos relativamente homogêneos por tipos ou categorias de imóveis. Os sub-modelos precisam ser coerentes entre si e devem ser de fácil entendimento, tendo em vista a aceitação por parte do contribuinte.

No caso de avaliações para hipotecas, Shiller e Weiss (1999) afirmam que um método de avaliação ou um sistema de avaliações deve ajudar o decisor a aprovar ou não os pedidos de hipotecas e prevenir as perdas decorrentes de faltas, sem incluir custos elevados. Para tributação, o principal requisito é a garantia de eqüidade, ou seja, que o nível de precisão seja similar para todos os tipos de imóveis (De Cesare, 1998). Estes elementos podem ser condensados em três aspectos básicos:

a) objetividade, incluindo facilidade de aplicação e economia de escala;

b) precisão, englobando eqüidade, justiça, estabilidade no tempo e confiabilidade;

c) explicabilidade, envolvendo também transparência e defensibilidade.

A objetividade pode ser atingida através do uso de métodos adequados, basicamente com a redução da subjetividade. A precisão decorre também da aplicação dos métodos adequados, e pode ser medida pela diferença entre as estimativa e os preços observados, para os imóveis da amostra.

Por outro lado, a explicabilidade é um princípio mais geral e não pode ser medida objetivamente, exigindo a formulação de alguns critérios. Gloudemans apresenta sete condições para atingir-se a explicabilidade em um sistema de avaliação em massa para tributação (Gloudemans, 1982[13], *apud* Deddis *et al.*, 1998):

a) simplicidade dos modelos em termos de forma funcional e racionalidade das variáveis empregadas[14];

b) razoabilidade do valor monetário em relação aos atributos do imóvel (os efeitos de um atributo em particular devem estar de acordo com as expectativas);

c) consistência entre os sub-modelos;

d) consistência no tempo (garantia de que os valores não mudam inexplicavelmente de um período para o outro);

e) possibilidade de decomposição entre valor da terra e das benfeitorias nos locais onde a tributação é realizada separadamente ou com alíquotas diferentes;

f) possibilidade de demonstrar em termos simplificados como foi obtido o valor e não a simples exposição de um modelo matemático;

g) explicabilidade através da apresentação de casos similares (comparáveis).

Com exceção do princípio (e), os demais podem ser estendidos para qualquer tipo e finalidade de avaliação. McCluskey *et al.* (1997) afirmam que os sistemas desenvolvidos para avaliação de massa exigem uma grande quantidade de dados, os quais devem ser apropriados,

---

[13] GLOUDEMANS, R.J. The base home approach to explainability in mass appraisal. In: **Legends of Carna**. Lincoln Institute of Land Policy/International Association of Assessing Officers. USA, 1982.

[14] Há um *trade-off* que o analista precisa resolver, entre simplicidade e precisão, pois provavelmente um modelo mais completo será mais preciso. Este elemento é conhecido como "Navalha de Occam" (*Occam's razor*).

relevantes, recentes, precisos, completos e de qualidade. Estes autores identificam a obtenção e o tratamento da informação como um dos principais problemas também na avaliação individual. Atualmente os profissionais buscam informação de forma isolada e não estruturada. Para estes autores, as barreiras à disseminação das informações aumentam a mística e a inconsistência das avaliações. Ratcliff (1956)[15], *apud* Kinnard (1966), afirma que os dados a serem escolhidos devem ser dos imóveis "competidores", não necessariamente dos "comparáveis", como normalmente são entendidos. Por exemplo, se residências em condomínios horizontais e apartamentos são opções igualmente consideradas para um determinado segmento da população, então ambos poderiam ser incluídos no mesmo modelo.

A exigência de explicabilidade é discutível para algumas aplicações, tal como o imposto sobre a propriedade. Para determinados segmentos da população, qualquer modelo convencional, mesmo uma equação de regressão linear, pode ser muito complexo. Neste caso, outras alternativas devem ser consideradas, tais como o uso de mapas, tabelas ou exemplos, como auxiliares para a explicação dos valores estimados.

Em termos formais, as avaliações individuais necessitam de um relatório, laudo ou formulário individual, enquanto que nas avaliações de massa não há relatório, ou há apenas um relatório geral. No caso dos tributos, há emissão de carnês de pagamento, e os sistemas específicos para tributação geralmente incluem esta facilidade. Accetta (1999) recomenda que o laudo seja consistente, evitando o uso de jargões e palavras dúbias, antecipando questionamentos e analisando os fatos relatados, sempre tendo em mente que o leitor provavelmente não é um técnico.

As avaliações também devem incluir uma cuidadosa análise de mercado. Kinnard (1966) aponta a necessidade de análise do mercado, aplicando inclusive simulação de cenários futuros (tal como mudança nas taxas de juros), incluindo também uma macro-análise do mercado imobiliário. Entre diversos outros autores, Fanning *et al.* (1994), Mitchell (1993) e Wincott e Mueller (1995) também afirmam que a análise de mercado é necessária e fundamental para o resultado da avaliação. Waller (1999) afirma que sistemas automatizados também podem proporcionar o conhecimento de tendências de mercado antes que estas sejam

[15] Ratcliff, R.U. **Modern real estate valuation**: Theory and applications. Chicago: AREUEA, 1956.

claramente discerníveis, aumentando o conhecimento do avaliador. Accetta (1999) afirma que o avaliador deve conhecer bem o mercado e, se não conhecer, deve recusar o trabalho.

Kinnard (1971) e Reenstirena (1996) afirmam que tradicionalmente os avaliadores indicam o valor do imóvel através de uma estimativa pontual. Em alguns casos é realmente mais adequado trabalhar com um único valor, como no caso de processos judiciais ou na tributação. Porém em algumas situações os clientes desejam saber a chance de ser obtido este valor e um intervalo de valores é mais interessante, como na orientação para vendedores. Lentz e Wang (1998) afirmam que é importante conhecer a dispersão dos preços, além da estimativa do valor de mercado, nas avaliações para hipotecas. Geralmente a questão da dispersão é abordada através do cálculo de um intervalo de confiança, procedimento comum quando a avaliação é realizada com análise de regressão.

Por outro lado, os compradores ou investidores desejam ser informados também sobre o valor de investimento (potencial), como retorno para o investimento ou garantia de um empréstimo. Assim, outro problema a ser tratado é a previsão do comportamento do mercado. Shiller e Weiss (1999) entendem que no caso das hipotecas é essencial que as avaliações indiquem as variações potenciais do valor para o futuro. Para Reenstirena (1996), nestes casos trata-se de análise de risco, a ser tratada com as técnicas específicas para este fim. Este autor reconhece que a complexidade aumenta, porém entende que esta informação pode ser um requisito fundamental para o cliente. O potencial futuro pode ser apreciado com uma análise cuidadosa do mercado e a verificação do comportamento macroeconômico do mercado, com a consideração de séries temporais e cruzamento de informações. Estas análises não se enquadram na definição convencional de avaliação de imóveis, pois tratam de previsões de médio ou longo prazos, e as avaliações tratam do mercado presente.

2.3.3.1 Requisitos a serem atendidos pelo profissional

O avaliador deve ser qualificado, independente das partes e conhecedor do mercado, segundo Accetta (1999) e Dowse (2000). Carneghi (1999) aponta requisitos para o profissional que desenvolve arbitragem (solução extrajudicial de disputas) nesta área, incluindo experiência (em geral e com o tipo de imóvel em avaliação), qualificação (através dos institutos profissionais), conhecimento do mercado local e ausência de relações anteriores com as

partes.

Gelbtuch *et al.* (1997) apresentaram um panorama mundial sobre esta área profissional. Há várias abordagens sobre a regulamentação profissional, existindo algumas formas típicas, tais como as exemplificadas a seguir. Nos Estados Unidos, há cerca de 10 anos é exigido um licenciamento ou certificação do profissional no estado em que se situa o bem a ser avaliado, obtido de acordo com normas locais. No Reino Unido não há exigência formal de licenciamento, porém o mercado prefere claramente os profissionais inscritos nos institutos profissionais, os quais passam por cursos e provas para serem aceitos. Na Argentina, o licenciamento só é exigido nas causas judiciais, e é obtido pela graduação em curso afim e registro no conselho profissional correspondente das províncias. No Brasil, o profissional avaliador deve ser Engenheiro Civil ou Arquiteto, para avaliação de imóveis urbanos, e Engenheiro Agrônomo ou Agrimensor, no caso de imóveis rurais, conforme a Lei nº 5194/66, a qual regulamenta o exercício da profissão para Engenheiros e Arquitetos, e a Resolução nº 218 do CONFEA (Conselho Federal de Engenharia, Arquitetura e Agronomia), que define as atribuições de cada categoria profissional.

Por outro lado, não há normas técnicas internacionalmente aceitas e geralmente as instituições profissionais de cada país desenvolvem as normas de uso interno (Gelbtuch *et al.*, 1997). A norma brasileira sobre a avaliação de imóveis atualmente é a NBR-5676 (ABNT, 1989), mas está em curso um processo de elaboração de uma nova norma. Uma parte desta norma já está aprovada, correspondendo à parte de definições, a qual recebeu o nome de NBR-14653-1 (ABNT, 2000).

O profissional também deve ser independente das partes ou contratantes. Worzala *et al.* (1998) pesquisaram a influência da pressão sofrida pelos avaliadores por parte dos clientes. Não encontraram evidências de que a importância do cliente para o avaliador ou o nível de ajuste requerido tenham provocado alterações nos resultados[16]. Contudo, um percentual elevado de avaliadores relatou já ter sofrido pressões para alterar resultados. Em alguns casos os honorários são proporcionais aos valores dos imóveis (o que também ocorre no Brasil,

[16] Segundo os critérios da pesquisa, um grande cliente seria aquele com 30% ou mais do trabalho mensal fornecido ao avaliador, e os níveis de ajuste foram entendidos como pequenos se não ultrapassavam 5% do valor inicial e grandes se atingiam de 15 a 20% da estimativa (Worzala *et al.*, 1998).

conforme orientações dos próprios institutos profissionais). Outro elemento que poderia contribuir para a manipulação é o pequeno porte das empresas de avaliações frente aos contratantes (nos Estados Unidos, os maiores contratantes são as corporações de investimentos imobiliários ou de hipotecas). Ademais, Worzala *et al.* (1998) relataram que na Justiça dos Estados Unidos existem diversas causas de avaliadores demitidos por não aceitarem manipulações em avaliações pelas quais eram responsáveis. A diminuição da subjetividade do processo avaliatório pode colaborar, neste aspecto.

Segundo Smith (1986), a imprecisão e as disparidades entre as avaliações e o mercado podem levar a uma gradual diminuição da confiança pública sobre os avaliadores, diminuindo também os honorários e as oportunidades de trabalho. Além disto, segundo McCluskey *et al.* (1997), o profissional deveria ser mais aberto às mudanças e mais associativo. Para estes autores, a profissão geralmente mantém uma posição conservadora e insular, e tais atitudes não são sustentáveis por muito tempo. Há muita dificuldade em formar bases de dados conjuntas, mesmo com intermediação dos institutos, por exemplo. McCluskey *et al.* (1997) entendem que os profissionais precisam adaptar-se às novas exigências dos clientes e às novas oportunidades disponibilizadas pelo desenvolvimento tecnológico. DeWeese (1998) afirma que o mercado para os avaliadores não tem aumentado porque muitas avaliações são realizadas pelas próprias corporações de investimento imobiliário (*Real Estate Investment Trusts - REIT*), as quais têm necessidades que não são atendidas pelo formato tradicional das avaliações, tais como a resposta rápida e a consideração do comportamento futuro do mercado.

2.3.3.2 Precisão (níveis e fontes de erro)

A precisão é o principal requisito das avaliações de imóveis e o único que pode ser mensurado objetivamente. Entretanto Gilbertson (2001) afirma que as avaliações não podem ser comprovadas diretamente, pois os imóveis geralmente não são vendidos logo após as avaliações. Ademais, normalmente os imóveis não são transacionados exatamente pelo valor de mercado. Assim, a precisão das avaliações é verificada indiretamente, pela comparação do valor estimado pelo modelo com os preços de venda para os imóveis da amostra (Parker, 1998).

Parker (1998) afirma que em princípio as avaliações devem ter precisão, mas a imprecisão é aceita como fato normal entre os avaliadores. Os erros seriam devidos à falta de um controle ou registro central das transações, à heterogeneidade dos imóveis, e à confidencialidade das informações. Além disto, no Reino Unido os imóveis são freqüentemente avaliados por empresas ou profissionais que participam das mesmas transações como intermediários. Parker (1998) aponta diversos autores, concluindo que limites de erros de 10% a 15% (em torno do valor avaliado) são bem aceitos entre os profissionais e mesmo nos tribunais britânicos, mas afirma que não há evidências de que os usuários tolerem bem estes níveis de erro. Da mesma forma, Crosby *et al.* (1998a, 1998b) afirmaram que os juizes na Grã-Bretanha e na Austrália aceitam erros de 10% em casos normais e até 15% ou mais em casos excepcionais. Gilbertson (2001) também indica que erros de até 15% são considerados aceitáveis pelos Tributais britânicos. Pesquisada a jurisprudência dos tribunais brasileiros, não foram encontradas causas em que fosse discutida a qualidade ou o nível de erro dos trabalhos avaliatórios.

As recomendações das instituições internacionais, como a *International Association of Assessing Officers* (IAAO), são de que os erros devem ser medidos através do coeficiente de dispersão (COD), não devendo ultrapassar 10% em avaliações individuais e 15% em avaliações em massa, com amostras de maior variabilidade, em avaliações de residências. Não há recomendações de limites similares na norma brasileira atual de avaliações (ABNT, 1989; IAAO, 1990). Revisando avaliações reais, Newell e Kishore (1998) encontraram erros médios de 9% em avaliações de propriedades comerciais em Sydney, mas somente 65% dos casos estavam na faixa de 10% em torno do preço de venda, com 9% diferindo em mais de 20%.

Por outro lado, a diferença entre preços de venda e valores estimados pode ser decorrente, além de erros de modelagem, do próprio processo de formação dos preços (de acordo com o funcionamento do mercado imobiliário), pois os preços praticados incorporam as distorções do mercado, tais como as diferenças de informação entre os agentes. Bonissone *et al.* (1998), citando Case e Shiller (1987), afirmam que haveria um erro mínimo de 5 a 7%, o qual seria intrínseco ao mercado imobiliário, em função da imperfeição do mercado. Porém, na verdade os números de Case e Shiller (1987) incluem outros fatores além do ruído inerente ao processo de vendas, tais como algumas variações nas características dos imóveis, os quais não foram completamente incorporados nos modelos desenvolvidos por estes autores. De qualquer

forma, há uma indicação de que o nível de erro mínimo é pequeno, para a área de estudo do trabalho (Atlanta, Chicago, Dallas e São Francisco). Neste sentido, Evans (1995) afirma que o erro seria de 10%, em função da ineficiência do mercado.

Por outro lado, a análise da consistência de avaliações em imóveis comerciais revela uma certa homogeneidade. Alguns autores examinaram avaliações produzidas por diversos profissionais com base nos mesmos dados e apontaram erros aleatórios de 2% a 6% nas avaliações, com os erros determinísticos em níveis um pouco maiores. Os erros determinísticos devem-se a fatores como sazonalidade ou regionalidade das variações dos preços ou a problemas metodológicos, ou seja, basicamente devidos à falta de uma análise mais cuidadosa do mercado (Diaz, 1997; Graff e Young, 1999).

2.3.3.3 Subjetividade nas avaliações

Há autores que discutem se as avaliações são arte ou ciência, geralmente concluindo que se trata de um misto entre as duas, porém muitas vezes com predominância para a arte na prática corrente (Gilbertson, 2001; Kinnard, 1966; McCluskey, 1996; Smalley, 1995). Esta discussão é baseada no método tradicional de ajustamentos manuais (correspondente à homogeneização de fatores). Para Smalley (1995), o processo de avaliação envolve ao mesmo tempo ciência (objetividade na análise) e arte (envolvendo a experiência, por exemplo no julgamento de qualidade de localização ou depreciação). A própria seleção dos dados é subjetiva e Smalley (1995) sugere incluir todos os dados disponíveis, indicando as razões adotadas para a seleção. Gilbertson (2001) afirma que, embora uma visão científica deva ser objetiva, o cliente espera uma interpretação subjetiva da análise objetiva desenvolvida através dos métodos de cálculo, bem como uma avaliação subjetiva das condições, com seus sentimentos profissionais sobre o comportamento futuro do mercado, sobre qualidade de micro-localização, depreciação e outras questões.

Já foi demonstrado que a experiência tem um papel muito importante nas avaliações e que a subjetividade é uma constante nas avaliações. Porém, a falta de objetividade dificulta até a auto-avaliação do trabalho realizado. Diaz (1997) e Diaz e Hansz (1997) pesquisaram se os avaliadores são influenciados por resultados de avaliações anteriores, desenvolvidas por outros avaliadores. No primeiro caso os avaliadores conheciam a região onde se situavam os

imóveis em análise e não foram significativamente influenciados (Diaz, 1997), enquanto que no segundo caso não conheciam bem a região e foram influenciados (Diaz e Hansz, 1997). Hansz e Diaz (2001) indicaram que os avaliadores também são afetados pela retroalimentação recebida em avaliações recentes, mesmo se as avaliações examinadas estavam corretas. Nas avaliações seguintes, os avaliadores tendem a corrigir os valores no sentido inverso ao apontado na retroalimentação. Spence e Thorson (1998) compararam as estimativas de 72 avaliadores novatos com as de 69 avaliadores experientes. Os resultados indicaram que os especialistas têm estimativas mais próximas dos valores de mercado, mas a faixa de valores apontada pelos avaliadores experientes investigados nem sempre inclui o valor obtido em uma venda posterior.

Nestas pesquisas, é preciso ter presente que os autores geralmente abordam o tradicional método de ajustamentos manuais[17], baseados em três casos similares, que é a forma usual nos Estados Unidos e em muitos países, como é o caso dos estudos de Accetta (1999) e Smith (1986). Por exemplo, Smith (1986) apontou diversas inconsistências entre a prática corrente e as teorias ou recomendações, mas as críticas deste autor estão direcionadas aos métodos "tradicionais" (custo e renda, além da comparação através de ajustes manuais).

Os níveis de erros próprios do mercado, por outro lado, tendem a ser menores nos países desenvolvidos do que os brasileiros, pois a precisão das avaliações nos Estados Unidos, baseadas em apenas três dados e análise subjetiva, pode estar repousada em um mercado estável, com baixo crescimento populacional e economia estável. Há indícios de que quando o mercado muda rapidamente surgem sérios problemas de precisão (Mitchell, 1993). Case e Shiller (1987, 1989, 1990, 1994) e Case (2000) identificaram ciclos econômicos no mercado imobiliário que podem explicar em parte os erros cometidos.

No caso brasileiro, é possível que o nível de erro inerente ao mercado seja maior, pela maior dificuldade de informação, maior desigualdade na população em termos de distribuição de renda, maior dificuldade de obter financiamentos, e altas taxas de juros praticados, entre outros fatores.

[17] Esta técnica é conhecida no exterior como *Comparable Sales Analysis - CSA* ou *Adjustment Grid Method – AGM*, e no Brasil aproxima-se da técnica de homogeneização de fatores.

2.3.4 Métodos de avaliação

Existem vários métodos para se encontrar o valor de mercado de uma propriedade. O principal é a comparação com dados de transações de imóveis semelhantes efetuadas na época da avaliação. Nem sempre isto é possível, principalmente quando se trata de imóveis singulares, tais como grandes prédios comerciais e industriais. Nestes casos, devem ser aplicados outros métodos, tais como os apresentados a seguir.

2.3.4.1 Custo de reprodução

O método do custo de reprodução compõe-se do cálculo do custo de aquisição do terreno e cálculo do custo de construção da edificação. O método fundamenta-se na premissa de que um comprador bem informado não pagará mais que o necessário para construir uma propriedade substituta, com a mesma utilidade daquela que está comprando. A base deste método é a consideração de que o valor de um imóvel é equivalente ao custo de execução da construção mais o custo do terreno (Moreira, 1997).

Este método é usado para construções que raramente mudam de dono, ou que são muito singulares, tais como hospitais e escolas, para os quais existe pouca ou nenhuma evidência em forma de preços de venda, ou para prédios comerciais e industriais. É um método também largamente utilizado para calculo de impostos prediais. Outra utilização é na avaliação de prédios inacabados ou em péssimo estado de conservação (exigindo reformas), segundo Ratcliffe *et al.* (1993) e Seeley (1976).

2.3.4.2 Residual

Se for necessário avaliar um terreno situado em área extremamente urbanizada e não há informações de vendas de terrenos livres, mas existem vendas de terrenos com construções, este método é útil. O valor do terreno é obtido a partir do valor total do imóvel, subtraindo-se deste os valores das construções existentes, que podem ter seus valores determinados por outros métodos, como o custo de reprodução, considerando depreciação e vantagem da coisa feita. Também é empregado se a finalidade é a apuração do valor das construções, em si, feita

então pela subtração do valor do terreno do valor total do imóvel, nos mesmos moldes. Outro uso comum é no cálculo das Plantas Genéricas de Valores, no formato tradicional, utilizando homogeneização de fatores e custo de reprodução (Fiker, 1993; Moreira, 1997).

2.3.4.3 Renda

Neste caso, o valor do imóvel é obtido pela capitalização de sua renda real ou prevista, a uma taxa de juros compatível com o mercado, representado pelo valor atual dos benefícios futuros que resultam dos direitos de propriedade (usufruto). É usado quando o valor depende essencialmente da capacidade de gerar lucros, como no caso de hotéis e cinemas. A abordagem usual é estimar o rendimento bruto e deduzir os custos de trabalho e juros sobre o capital (Fiker, 1993; Moreira, 1997; Seeley, 1976).

2.3.4.4 Máximo aproveitamento eficiente (involutivo)

No exterior, este método é geralmente considerado como uma forma de avaliação pela renda, enquanto que no Brasil é identificado na Norma e na literatura como um método à parte (ABNT, 1989; Moreira, 1997). Tendo em vista que a terra não é reproduzível e que as construções têm elevada vida útil, a comparação entre dois terrenos poderá ser realizada em função do seu aproveitamento. O método do máximo aproveitamento eficiente busca identificar os melhores usos, em qualidade e quantidade. Sendo uma gleba urbana, uma das alternativas é o loteamento. Todos os tipos de construções que podem ser executados devem ser investigados. O avaliador deve fazer até um anteprojeto da construção, levando em conta as utilizações permitidas pelos planos diretores e pelos usos tradicionais na região. Definida a utilização, a análise segue com a execução de orçamentos (mais ou menos detalhados, conforme o caso), verificação da viabilidade e dos frutos e despesas esperados (aluguéis, lucros na venda, taxas, custos de publicidade e corretagem, etc.) – (Moreira, 1997).

2.3.4.5 Comparação de dados de mercado

O método comparativo de dados de mercado é o mais empregado. Consiste em fazer uma

comparação direta com os preços pagos no mercado para propriedades similares, quando existem substitutos razoavelmente semelhantes e ocorrem transações com uma certa freqüência. É largamente empregado para propriedades residenciais, nas quais existe normalmente mais similaridade entre diferentes unidades. As principais dificuldades do uso do método estão associadas à impossibilidade de encontrar propriedades idênticas - há diferenças na área construída, acabamento, estado de conservação, etc. (Kinnard, 1966; Seeley, 1976).

O método é baseado em informações sobre preços de propriedades comparáveis com a que está sendo avaliada. Os avaliadores precisam conferir as condições em que são feitas as transações (motivos dos compradores e vendedores), para verificar se os preços são típicos de mercado ou se existem condições não econômicas influindo (Weimer e Hoyt, 1948).

Para se obter o valor de um imóvel por este método, é preciso que existam dados de transações com imóveis semelhantes, em número e especificação razoáveis, para permitir a obtenção de resultados com confiabilidade. A partir da amostra do mercado, atualmente três técnicas podem ser adotadas: análise intuitiva, homogeneização de valores e inferência estatística.

No primeiro caso não há um modelo explícito, com os ajustamentos sendo desenvolvidos em bases subjetivas. Na homogeneização de fatores há um modelo determinístico, com fatores de correção determinados diretamente pelo avaliador. A terceira alternativa é baseada em modelos extraídos dos dados. Nesta última, há um modelo hedônico de preços, que visa a relacionar os atributos importantes para o mercado com os preços praticados, através do estabelecimento de uma função hedônica (Dodgson e Topham, 1990; Sheppard, 1999). Como cada imóvel tem um conjunto diferente de características, apresenta um preço diferente. A precisão dos modelos de formação de preços depende das soluções para dois problemas principais, segundo Griliches (1971):

a) identificar os atributos relevantes (neste caso, atributos dos imóveis e do mercado);

b) encontrar a forma real do relacionamento entre os preços e estes atributos;

Estas questões geralmente são resolvidas empiricamente. Os modelos hedônicos vêm sendo

usados há décadas, especialmente em estudos que examinam os efeitos de vários atributos sobre os preços dos imóveis, mas a literatura corrente demonstra que os pesquisadores na área de economia urbana têm utilizado múltiplos caminhos para este tipo de análise, sem que seja encontrado um resultado definitivo (por exemplo, ver Ball, 1973; Bartik e Smith, 1987; Bible e Hsieh, 1996; Boyle e Kiel, 2001; Ding *et al.*, 2000; Isakson, 1998; Smith *et al.*, 1988).

2.3.4.5.1 Homogeneização de valores

A idéia geral de utilizar fatores determinísticos para ajustar as diferenças nos imóveis tem diversas variantes, mas basicamente trata-se da utilização de fatores derivados da experiência, gráficos gerais ou estatística descritiva. No Brasil, é utilizada a homogeneização de fatores. Esta técnica consiste de uma espécie de ponderação arbitrária das diferenças entre os imóveis. Por este processo, os elementos da amostra são alterados por fatores ou coeficientes corretivos, de modo a torná-los mais semelhantes. Tais coeficientes dependem do julgamento do avaliador, sendo baseados em métodos e bibliografia correntes. A deficiência do método consiste justamente na obtenção dos fatores. Na bibliografia há diversas tabelas e fórmulas para calcular os fatores de homogeneização. Porém, não se pode afirmar que o comportamento do mercado de uma região se repita em outras regiões, de características distintas. Ao contrário, tudo leva a crer que existem modificações substanciais de um local para outro, inclusive pela diversidade sócio-econômica-cultural das diferentes populações. As transformações constantes das cidades e da economia também invalidam estas tabelas. Portanto, a interferência do avaliador é fundamental para obter-se o valor, tendo seus critérios e considerações muita influência no resultado final. A homogeneização está bastante detalhada e documentada nas obras de diversos autores, tais como Berrini (1957), Caires (1978), Caires e Caires (1984), Chandias (1954), Fiker (1993), IBAPE (1974, 1983), Moreira (1997) e Vegni-Neri (1968, 1979).

2.3.4.5.2 Inferência estatística

Esta linha de avaliação trata o processo de ajustamento das diferenças de forma mais objetiva. O embasamento teórico vem dos modelos hedônicos de preços, enquanto que a estimação é geralmente realizada através de análise de regressão, em uma visão econométrica. O analista deve estipular modelos com as hipóteses de relacionamento entre as variáveis, que devem ser testados segundo critérios estatísticos, verificando-se a validade destas hipóteses, ou seja, se os modelos são capazes de representar o segmento de mercado em questão. Para tanto, devem ser coletados dados de transações (evidências do mercado), analisando-se o ajuste dos modelos considerados a estes dados, dentro de um determinado grau de precisão. Os testes estatísticos permitem avaliar o próprio modelo e a importância individual das variáveis incluídas, indicando a qualidade geral do modelo formulado. O modelo convencional assume o formato apresentado na Equação 1 (Neter *et al.*,1990):

$$Y = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 + ... + \alpha_k X_k + \varepsilon_\alpha = Y^h + \varepsilon_\alpha \qquad \text{(Equação 1)}$$

Este formato é chamado de "modelo linear clássico", no qual Y é a variável dependente ou explicada (geralmente o preço), $X_1,...,X_k$ são as variáveis independentes ou explicativas (as características dos imóveis e da região), α_0 é o intercepto da equação, $\alpha_1,...,\alpha_k$ são os coeficientes parciais da regressão (preços hedônicos implícitos), ε_α é o termo de erro (desvio da estimativa) e Y^h é a estimativa para a variável dependente, calculada em função das variáveis explicativas incluídas (Judge *et al.*, 1985; Neter *et al.*, 1990, Ramanathan, 1998).

Os coeficientes α_i são estimados geralmente pelo Método dos Mínimos Quadrados, que busca um conjunto de coeficientes que minimize o quadrado dos erros do modelo. O processo de análise de regressão exige que sejam atendidos alguns requisitos essenciais, chamados de pressupostos básicos, e ainda outras condições relacionadas, os quais precisam ser respeitados para que a análise seja válida, e as inferências (estimativas) possam ser realizadas com a equação determinada. Os princípios a serem atendidos para garantia de validade dos modelos, e as conseqüências da violação destes pressupostos, são os seguintes, conforme Cuthbertson *et al.* (1992), Maddala (1988) e Neter *et al.*(1990):

a) o modelo é linear nos parâmetros: este pressuposto decorre da própria forma do modelo clássico (Equação 1) e a fuga, por não-linearidade da função, provoca

tendências nos resíduos.

b) as variáveis independentes são constantes: é necessário para garantir a estabilidade dos coeficientes na repetição de amostras da mesma população. Além disto, se as variáveis explicativas são aleatórias ocorre a diminuição do poder dos testes de hipóteses.

c) os resíduos seguem a distribuição Normal: a suposição de normalidade dos resíduos simplifica a teoria de análise de regressão e é necessária para garantir a validade dos testes de hipóteses e a estimação de intervalos de confiança. Pequenas fugas não são importantes, pois os testes de hipóteses são baseados na distribuição *t*, que não é muito sensível a estes desvios.

d) os resíduos têm média nula: geralmente este pressuposto é garantido pela fixação conveniente do termo constante (α_0), mas deve ser verificado para evitar tendências nos resíduos.

e) há homocedasticidade dos resíduos (a variância é constante): as conseqüências da heterocedasticidade são de que as estimativas dos parâmetros da regressão são ineficientes (ou seja, a variância não é mínima), as estimativas das variâncias são tendenciosas e os testes de hipóteses (*t*, *F*) tendem a fornecer resultados incorretos.

f) os resíduos são independentes entre si: ou seja, não há autocorrelação. Existindo correlação entre os resíduos, os estimadores de mínimos quadrados não são mais os melhores estimadores lineares não tendenciosos e os testes *t* e *F* indicam conclusões incorretas.

g) não há colinearidade entre quaisquer variáveis independentes: a perfeita correlação entre duas ou mais variáveis, ou seja, uma é combinação linear de outra(s), implica a existência de diversos modelos com o mesmo grau de ajustamento, não sendo possível selecionar um dos modelos, o que impede a interpretação sobre os coeficientes. Na prática, é mais comum a existência de colinearidade em grau menor ($|r|<1$), que provoca alteração nos coeficientes das variáveis afetadas, inclusive invertendo sinais.

O modelo deve ainda atender aos seguintes requisitos, decorrentes dos pressupostos e da forma de cálculo dos coeficientes (Neter *et al.*,1990):

h) não existem observações espúrias: a existência de elementos claramente não adaptados ao modelo (chamados de *outliers*), provoca distorções nos coeficientes, quando estes são calculados pelo MMQ, pois um erro relativamente grande tem influência sensível nos coeficientes, mascarando os resultados.

i) as variáveis importantes foram incluídas: o modelo especificado deve ser similar ao real, e a falta de variáveis importantes provoca tendências nos resíduos, por falta de explicação do fenômeno (variação da variável dependente).

j) a amostra de dados é suficientemente grande: ou seja, o número de observações é maior que o de coeficientes a serem estimados. Este requisito é necessário para que possam ser realizados os cálculos dos coeficientes (se o número de casos e de coeficientes for igual, trata-se de um sistema de equações, nos domínios da Matemática).

No mercado imobiliário, diante de suas características peculiares e da dificuldade de obtenção de dados, há risco de ruptura de várias destas condições. É comum a ocorrência de multicolinearidade, *outliers* e não-linearidade dos dados (De Cesare, 1998; Dubin, 1988; Worzala *et al.*, 1995; Wyatt, 1996a). Porém, a ruptura de algum dos pressupostos não impede a utilização desta técnica, apenas dificulta ou impede a generalização dos resultados. Se não estão presentes as condições básicas, a análise passa a ter um caráter determinístico, podendo ser vista como um ajustamento de curvas, e os resultados são restritos ao conjunto específico de dados considerado. Se todas as características do avaliando estiverem contidas dentro dos intervalos respectivos, não há prejuízos maiores, e também não impedem, em tese, o prosseguimento da análise. Porém, para avaliação de massa trata-se de um empecilho sério, em função da diversidade dos "avaliandos". Ademais, há diversas críticas à aplicação desta técnica, na prática. Kummerow (2000) e Smith (1995) contestam a hipótese de normalidade dos preços, basicamente afirmando que não há repetição de eventos, nem quanto à venda (preço praticado), nem quanto à avaliação (valor estimado) de um imóvel específico, portanto não poderia haver distribuição de probabilidade, invalidando o restante do arcabouço teórico

das avaliações através de inferência estatística. Porém há uma deficiência nesta argumentação, pois a aleatoriedade que deve ser confirmada é sobre o comportamento dos erros, e não dos preços ou dos valores. Cada transação que ocorre no mercado apresenta uma diferença entre o preço observado e o valor de mercado (ou preço mais provável) correspondente. Este erro, que decorre dos efeitos da concorrência imperfeita e da ineficiência do mercado, é aleatório, e geralmente pode ser demonstrado empiricamente que é, ao menos aproximadamente, Normal.

Quase todas as condições podem ser razoavelmente garantidas, ou mesmo bem controladas, existindo técnicas estatísticas para corrigir eventuais problemas. Por exemplo, para o caso de surgimento de *outliers*, há tratamentos bem documentados e a solução é razoavelmente fácil, bastando tomar os cuidados necessários para a seleção dos casos relevantes (Belsley *et al.*, 1980; Daniel e Wood, 1980). A multicolinearidade, causada por inter-relações das variáveis explicativas, pode ser eliminada pelo emprego de técnicas que transformam os dados, como a análise fatorial ou outras combinações entre as variáveis (González, 1993; Harmann, 1976; Maddala, 1988; Morton, 1997). Entretanto, alguns autores afirmam que a multicolinearidade não é um problema em modelos preditivos quando o relacionamento encontrado reflete um padrão do mercado, ou seja, quando não é apenas uma característica dos dados coletados (Gujarati, 2000; Judge *et al.*, 1985). Também a heterocedasticidade pode ser contornada, com o emprego de mínimos quadrados generalizados, desde que sejam obtidas boas estimativas para a matriz de ponderação, ou segmentação dos dados em grupos mais homogêneos (Judge *et al.*, 1985; Neter *et al.*, 1990; Newsome e Zietz, 1992) e assim por diante. Entretanto, para a autocorrelação e para a definição da forma funcional as soluções paliativas sugeridas não são suficientes para eliminar as dificuldades geradas, como descrito a seguir.

2.3.4.5.2.1 Autocorrelação espacial

A autocorrelação é o relacionamento entre elementos consecutivos. A forma mais comum é a correlação serial. Devido às características espaciais do mercado imobiliário, os modelos de regressão podem sofrer com um tipo de autocorrelação especial, denominado geralmente de "autocorrelação espacial", principalmente em análises baseadas em recortes espaciais ou no tempo, ou seja, amostragens do tipo *cross section* (Cliff e Ord, 1973; Cuthbertson *et al.*, 1992; Judge *et al.*, 1985; Wyatt, 1996a).

Nos modelos econômicos gerais, que lidam principalmente com a análise de séries de tempo,

é muito comum o aparecimento de relações seriais entre as medidas dos desvios das estimativas para os valores reais (erros), provocando dificuldades de análise. Como enunciado acima, se ocorre autocorrelação, os estimadores obtidos por Mínimos Quadrados são não-viesados, mas não tem a variância mínima e podem ser ineficientes, os testes *t* e *F* são imprecisos e os modelos não são plenamente válidos, havendo restrições para serem empregados na inferência de valores (Cuthbertson *et al.*, 1992; Neter *et al.*, 1990). A correlação serial (ou autocorrelação) implica a existência de uma relação indesejada, do tipo apresentado na Equação 2 (Neter *et al.*, 1990):

$$\varepsilon_t = \phi_0 + \phi_1 \varepsilon_{t-1} + \phi_2 \varepsilon_{t-2} + \phi_3 \varepsilon_{t-3} + ... + \phi_p \varepsilon_{t-p} + \upsilon_t \qquad \text{(Equação 2)}$$

Ou seja, o erro ε_t pode ser determinado em função dos erros anteriores, quando deveria ser aleatório, para atender aos pressupostos básicos da ARM (como foi visto, espera-se que seja independente, seguindo uma distribuição próxima da Normal, com média nula e desvio-padrão constante). Na Equação 2 está indicado o caso geral, com correlação de ordem "p". É mais comum ocorrer correlação de primeira ordem (p=1) ou de segunda ordem (p=2). Se for detectada correlação serial, uma solução possível seria corrigir o modelo teórico e recomeçar a estimação. Neste caso, a autocorrelação não poderia ser encarada como um problema dos dados, a ser removido por transformação das séries ou dos resíduos, mas por especificação correta do modelo. Pode ocorrer autocorrelação por omissão de uma variável importante ou por inclusão de variáveis em formato inadequado, por exemplo, geralmente com respeito às variáveis que medem a passagem do tempo, neste caso (Dubin, 1988; Maddala, 1988).

É importante verificar que, nas séries temporais, as variáveis consistem em medidas repetidas sobre o mesmo objeto, ocorrendo variação em apenas uma direção e sentido de análise (variação cronológica). Por exemplo, uma análise sobre o comportamento dos custos de construção pode utilizar uma série de medidas do Custo Unitário Básico (CUB), o qual é calculado mensalmente, há algumas décadas, com a mesma metodologia. A análise de *time series* é fundamental no âmbito da Economia e, por isso, há intensa pesquisa sobre este tipo de situação. No caso de correlação serial, os testes e métodos de estimação estão bastante desenvolvidos e a análise pode ser realizada por diversos processos, com boa probabilidade de sucesso. Por exemplo, existindo cointegração (relação de longo prazo) entre as variáveis, podem ser aplicados métodos complexos, mas eficientes, como o Procedimento de Johansen

ou o Mecanismo de Correção de Erros. Outras alternativas para o caso de autocorrelação incluem a estimação por processos baseados em máxima verossimilhança, mínimos quadrados generalizados, modelos ARIMA ou o emprego de técnicas iterativas, como o Filtro de Kahlman (Cuthbertson *et al.*, 1992; Engle e Granger, 1987; Hendry e Mizon, 1978; Pereira, 1991).

Entretanto, no mercado imobiliário os objetos de análise são distintos entre si, não existindo propriamente uma "série" (os imóveis são heterogêneos). A questão principal é a distribuição espacial. Conforme Camargo *et al.* (1999), "alguns processos espaciais, principalmente aqueles observados em aplicações ambientais, apresentam indexação no espaço e trazem como característica comum a continuidade, observando-se que seus valores variam de forma gradual numa determinada vizinhança". É o caso também do mercado imobiliário. Se a variável tem distribuição espacial, as variações nos dados ocorrem em todas as direções, o que dificulta a análise. Enquanto a correlação serial lida com a relação dos termos de erros entre si, a autocorrelação espacial é mais geral e está ligada a características peculiares dos dados, espacialmente distribuídos, que têm maior relação com os imóveis da vizinhança próxima do que com os mais distantes. De forma semelhante com o que ocorre com a correlação serial, a correlação espacial surge por falta de explicação correta das variações de preços no espaço, através das variáveis incluídas no modelo hedônico (Cuthbertson *et al.*, 1992; Judge *et al.*, 1985).

Anselin (1988[18], *apud* Macedo, 1998) afirma que, com dados distribuídos no espaço, podem ocorrer "efeitos espaciais", basicamente de dois tipos: "dependência espacial" e "heterogeneidade espacial". Assim, deve-se adaptar os tradicionais modelos econométricos, considerando estes efeitos e empregando modelos econométricos "espaciais", que levam em conta explicitamente estes efeitos.

Kelejian e Robinson (1992) relatam que muitos testes tem sido desenvolvidos para investigar a correlação espacial. São testes sofisticados, mas todos com restrições teóricas. Na verdade, os testes de avaliação da dependência espacial dos resíduos não são poderosos (no sentido da confiabilidade da conclusão) e não estão implementados em muitos dos *softwares* disponíveis.

[18] ANSELIN, L. **Spatial econometrics**: Methods and models. Dordrecht: Kluwer Academic, 1988.

Segundo Bennett e Hordijk (1986), em econometria espacial as técnicas de mínimos quadrados ordinários geralmente não funcionam bem, porque há violações nas condições básicas necessárias para validade da técnica, surgindo problemas como multicolinearidade, heterocedasticidade e autocorrelação espacial, simultaneamente, em função da consideração deficiente da distribuição espacial dos dados.

As análises que empregam modelos hedônicos apresentam dificuldades na consideração das características espaciais. Por serem de difícil determinação empírica, a qualidade de vizinhança e a localização são elementos quase intangíveis, na prática. Geralmente, busca-se identificar regiões "homogêneas", atribuindo-se valores para as características de interesse através da participação de especialistas (por exemplo, ver Lapolli *et al.*, 1994). Além dos problemas de julgamentos viesados, e do custo e tempo dispendidos, ainda se deve considerar que é muito difícil definir regiões realmente homogêneas, por causa de heterogeneidades internas (o espaço é multivariado, heterogêneo) e da definição arbitrária das próprias fronteiras (Dubin, 1992; Gallimore *et al.*, 1996; Moscovitsch, 1997).

Para Wyatt (1996a), a localização pode ser incorporada no modelo de regressão através do fracionamento da área em estudo em áreas homogêneas menores, nas quais os efeitos de vizinhança e acessibilidade sejam considerados similares, cada uma gerando uma equação de regressão. Porém, este tipo de zoneamento deve ser realizado por um profissional experiente, e se as zonas homogêneas tornam-se pequenas, o número de modelos cresce muito, diminuindo a eficiência do sistema.

Anselin (1998) observa que, até recentemente, o tratamento espacial explícito não era comum, por questões metodológicas e operacionais. A questão metodológica envolve a dificuldade de reconhecer a natureza bi-dimensional da interação espacial (correlação espacial) e suas implicações na análise estatística. Já foi demonstrado que ignorar este aspecto pode levar a estimativas tendenciosas ou ineficientes dos coeficientes ou inferências incorretas. Por outro lado, do ponto de vista operacional, a deficiência era devida à falta de *softwares* adequados, o que já não é mais problema, segundo o autor, devido à oferta de vários pacotes estatísticos com ferramentas espaciais e, principalmente, ao desenvolvimento dos Sistemas de Informação Geográfica.

Desta forma, pode-se concluir que a correlação espacial é um dos principais problemas estatísticos nas análises econométricas realizadas sobre o mercado imobiliário e a busca de

soluções a serem aplicadas em modelos hedônicos ou de outros métodos que permitam diminuir a dificuldade de estimação é importante para o aperfeiçoamento da avaliação de imóveis.

2.3.4.5.2.2 Forma funcional

Outro problema é o da formatação do modelo, que afeta os pressupostos de linearidade da equação e de que as variáveis importantes tenham sido incluídas. Quais variáveis incluir e em que formato é um problema estatístico não trivial e as revisões da literatura indicam estas dificuldades através da falta de uniformidade nos textos sobre economia urbana (Ball, 1973; Macedo, 1998; Smith *et al.*, 1988).

Os coeficientes da equação de regressão geralmente são estimados através do Método dos Mínimos Quadrados e uma das condições é que a forma escolhida para a equação seja adequada. Se a forma da equação não é conhecida, os modelos ajustados não podem ser utilizados para inferência. A especificação inadequada do modelo pode provocar autocorrelação ou heterocedasticidade, como indicado acima (Daniel e Wood, 1980; Neter *et al.*, 1990; Weisberg, 1985).

Um dos problemas mais comuns é a possível não-linearidade dos dados. Os relacionamentos entre as variáveis são complexos, nem sempre os dados ajustam-se linearmente ao modelo e a escolha da transformação matemática a ser aplicada não é tarefa fácil, exigindo forte intervenção do especialista no processo de modelagem (De Cesare, 1998; Worzala *et al.*, 1995). Uma tentativa de tornar objetivas as transformações a serem realizadas é a aplicação do procedimento de Box-Cox, o qual utiliza transformações do tipo $V_T=(V^{\lambda L}-1)/\lambda_L$, onde V_T é a variável transformada, V é a variável original e λ_L é o parâmetro de transformação (Neter *et al.*, 1990). Entretanto, não havendo indicações teóricas da forma de relacionamento das variáveis, a análise de uma função linear de regressão para verificar se ela é apropriada para os dados ou não geralmente é realizada por tentativas, através de gráficos de resíduos contra as variáveis dependentes ou independentes do modelo. Se houver uma forma definida nos erros, com tendências de crescimento ou curvaturas, pode ser que o modelo testado não seja o mais adequado, e deve-se tentar o ajuste de funções não lineares ou linearizar a função, por transformações nas variáveis, tais como logaritmos, inversas ou potências. Muitas vezes,

contudo, não há indicações claras do caminho a seguir, e as transformações a serem aplicadas, se mal escolhidas, podem até prejudicar o modelo (Kmenta, 1978; Maddala, 1988).

2.3.5 Técnicas alternativas

Há muitos exemplos de proposições de técnicas auxiliares ou substitutas à regressão. Quanto ao tratamento ou preparação dos dados, por exemplo, alguns autores utilizaram clusterização para identificar segmentos de mercado, dividindo os dados em sub-mercados, com imóveis relativamente homogêneos (Bourassa e Hoesli, 1999; Bourassa *et al.*, 1999; Kauko, 1997, 2000; Lewis *et al.*, 2001). Outros autores apresentaram a análise de componentes principais para compensar ou corrigir os efeitos de atributos multicolineares (González, 1993; Kain e Quigley, 1970; Morton, 1977; Wilkinson e Archer, 1973). Para a identificação do formato dos modelos, a principal sugestão da literatura é a exploração de transformações dos atributos através de Box-Cox (Anglin e Gençay, 1996; Barbosa e Bidurin, 1991; Dantas e Cordeiro, 2001; Kang e Reichert, 1987; Milon *et al.*, 1984). Entretanto, a busca por *outliers* é desenvolvida apenas com os dados da amostra coletada, e não há medidas cautelares, tal como a utilização de validação cruzada para testar os modelos, exceto em Barbosa e Bidurin (1991).

A identificação de atributos para representar a localização geralmente é desenvolvida através de medidas expeditas ou subjetivas (Ball, 1973; Bartik e Smith, 1987; Boyle e Kiel, 2001; Smith *et al.*, 1998), mas algumas tentativas de extrair medidas dos próprios dados vem sendo realizadas, usando sistemas de informação geográfica e superfícies de resposta. Esta última técnica envolve duas abordagens: gerar uma variável de localização posteriormente incluída no modelo de regressão ou embutir a "superfície" no modelo de regressão (Eichenbaum, 1989; Gallimore *et al.*, 1996; González, 1995a, 1995b; González e Erba, 1997; González *et al.*, 2002a; McCluskey *et al.*, 2000; Siu e Yu, 2001; Ward *et al.*, 1999). Lang e Jones (1975), por outro lado, demonstraram que medidas objetivas, tais como renda e educação dos residentes em determinado local, podem alcançar desempenho similar ao de medidas subjetivas, com vantagens em termos de tempo e custo.

Com respeito às técnicas de estimação propriamente ditas, os modelos podem ser classificados em dois grupos. De um lado estão os modelos baseados em conhecimento, tais como sistemas especialistas, raciocínio baseado em casos e sistemas de regras difusas e, de

outro, estão os modelos baseados diretamente nas técnicas de predição, tais como a regressão, as redes neurais e a vizinhança próxima.

Czernkowski (1989), Hsia e Byrne (1989) e Yao (1994) apresentaram sistemas especialistas para avaliação de imóveis, usando regras e atributos fixos no ajustamento das estimativas, geralmente com base em três casos similares selecionados de uma base de dados. Em uma abordagem mais sofisticada, Jensen (1990) apresentou um sistema que utiliza algumas formas de preparação dos dados, tais como seleção de casos e atributos, e que pode automaticamente gerar, verificar e selecionar diversos modelos de regressão linear e não-linear. Porém, os SE enfrentam as dificuldades de elicitação de conhecimento e manutenção dos sistemas. Para vencer estes problemas, sistemas utilizando raciocínio baseado em casos têm sido propostos para avaliação de imóveis, tais como os de González e Laureano-Ortiz (1992), O'Roarty *et al.* (1997a, 1997b), Pacharavanich *et al.* (2000), Pacharavanich e Rossini (2001) e Ribeiro (1999), embora apenas os dois primeiros tenham sido implementados. A dificuldade ainda enfrentada por estes sistemas é o ajustamento das diferenças (adaptação dos casos), que é desenvolvida com base em regras definidas *a priori* por especialistas.

A construção de modelos vinculados à própria ferramenta de modelagem, como análise de regressão e redes neurais, ainda é a abordagem mais comum para pesquisa ou avaliação de imóveis. A regressão linear é amplamente utilizada em pesquisa (Ball, 1973; Bartik e Smith, 1987; Boyle e Kiel, 2001; Smith *et al.*, 1988) e em avaliações (De Cesare, 1998; Dodgson e Topham, 1990; Isakson, 1998; Mark e Goldberg, 1988). Alguns autores têm sugerido formas alternativas para substituir a regressão linear, tais como modelos baseados em estatística espacial (Dubin, 1992, 1988; Pace *et al.*, 1998), regressão não-linear paramétrica (Dantas e Cordeiro, 1988, 2001) e não-paramétrica (Anglin e Gençay, 1996; Mason e Quigley, 1996; Pace, 1998).

Nos últimos anos, as redes neurais têm recebido muita atenção, com vários trabalhos publicados, tais como os de Bonissone *et al.* (1998), Borst (1991), Cechin *et al.* (1999, 2000), Connellan e James (1998), Evans *et al.* (1995), González *et al.* (2002b), Lenk *et al.*, (1997), Kathman (1993), Kauko (1997), Lewis *et al.* (1997), McCluskey (1996), McCluskey e Borst (1997), McGreal *et al.*, (1998), Nguyen e Cripps (2001), Rossini (1997), Tay e Ho (1994) e Worzala *et al.* (1995). A maioria destes trabalhos desenvolve comparações entre as redes neurais e a regressão linear. Nos estudos mais antigos havia divergência quanto aos

resultados. Mais recentemente, a comunidade de pesquisa parece ter aceitado a existência de um equilíbrio entre regressão e redes neurais, com eventual vantagem para um ou outro conforme peculiaridades das aplicações.

Contudo, existem outras críticas a alguns destes trabalhos, tais como conflitos entre os resultados, diferentes medidas de desempenho, amostras pequenas (com suspeita de *overfitting*[19]) e falta de um processo de validação dos resultados, os quais dificultam a avaliações dos resultados gerais. Em outros trabalhos falta ainda uma fundamentação teórica sobre o mercado imobiliário, consistindo apenas de uma aplicação para teste da ferramenta (Lewis *et al.*, 1997, 2001; Panayiotou *et al.*, 2000). Embora existam aplicações com redes neurais há mais de uma década no exterior, no Brasil o setor de avaliações praticamente ignorou estes avanços tecnológicos, sem apresentar pesquisas ou aplicações, com poucas exceções (Guedes, 1995; Cechin *et al.*, 1999, 2000). Entretanto, no formato apresentado nestes trabalhos as redes não podem ser consideradas como substitutas da regressão, pois não proporcionam a explicação dos resultados obtidos, o que é importante para algumas aplicações, como no caso de tributação. O único trabalho que desenvolveu este assunto foi o de González *et al.* (2002b), que utiliza regras extraídas com o uso de lógica difusa para explicar as redes neurais.

Neste sentido, alguns autores têm adotado abordagens híbridas, visando compensar as deficiências individuais das ferramentas. Bonissone *et al.* (1998) apresentou um estudo com dois sistemas híbridos. Um dos sistemas utiliza raciocínio baseado em casos, com as medidas de similaridade obtidas através de lógica difusa, visando a aprimorar a seleção dos casos. O outro é um sistema de regras difusas sintonizado através de uma rede neural. O modelo neuro-difuso apresentado em González *et al.* (2002b) é outro exemplo de modelo híbrido. McCluskey e Anand (1999) apresentaram uma aplicação de mineração de dados, com o objetivo de construir um sistema de avaliação em massa. Os valores estimados são calculados através do algoritmo de vizinhança próxima (k-NN), com o conjunto de pesos determinado por uma abordagem neurogenética[20].

---

[19] Excesso de ajustamento aos dados. Neste caso, o "modelo" inclui também o ruído dos dados (Haykin, 2001).

[20] O algoritmo de vizinhança próxima (*K-Nearest Neighbors*) é uma técnica para classificação ou estimação de valores (Han e Kamber, 2001; Pyle, 1999). É muito empregada para tratamento de casos com valores omitidos e seleção de casos nos sistemas de Raciocínio Baseado em Casos (ver capítulo 4), e também é utilizada para

2.4 CONSIDERAÇÕES FINAIS

A análise qualitativa e quantitativa do mercado imobiliário é importante para o reconhecimento do funcionamento no momento da avaliação, permitindo melhores estimativas. Da mesma forma, o conceito de valor adotado como premissa do trabalho é fundamental nas avaliações. Mais especificamente, a análise geral do mercado permite guiar o desenvolvimento da avaliação, fornecendo um contexto que pode indicar mais claramente a relevância de casos e atributos, e também para ampliar a segurança das estimativas.

Pode-se observar que não foram encontrados trabalhos com a proposição de análises sistemáticas para a preparação dos dados do mercado imobiliário, sendo uma questão abordada superficialmente. Dados com problemas são considerados *outliers* e sumariamente removidos da amostra. A seleção de casos e de atributos também não foi resolvida de forma consistente, sendo desenvolvida subjetivamente. Assim, ainda não foram propostas alternativas que pudessem contribuir para a redução da subjetividade contida nestas tarefas de seleção. Algumas técnicas sugeridas, tais como a estatística espacial e a regressão não linear, por serem técnicas paramétricas como a regressão, exigem o respeito a algumas condições sobre os dados ou dependem de definições ou decisões do analista sobre os parâmetros. Por fim, as redes neurais, embora apresentem bom desempenho em termos de precisão, foram exploradas isoladamente, sem um modelo explícito, o qual é necessário para a maioria das aplicações na área de avaliação de imóveis.

---

avaliação de imóveis (Isakson, 1986; Kolodner, 1993; McCluskey e Anand, 1999; Pyle, 1999; Watson, 1997). O algoritmo fundamenta-se na identificação de casos similares através de uma medida de distância, tal como a distância Euclidiana ou a distância de Mahalanobis. Na abordagem mais simples, os casos são ordenados pela semelhança (proximidade) com o caso-alvo e o mais similar é escolhido (k=1). Se for escolhido um conjunto (k>1), pode-se adotar a média simples, ou ponderar os valores pela importância relativa dos atributos utilizados que descrevem o caso. Estes pesos podem ser determinados pelo analista ou obtidos a partir de alguma técnica de estimação, como análise fatorial (Isakson, 1986) ou o sistema híbrido sugerido por McCluskey e Anand (1999).

3 DESCOBRIMENTO DE CONHECIMENTO EM BASES DE DADOS

3.1 CONSIDERAÇÕES INICIAIS

Inicialmente foram identificadas as alternativas disponíveis na área. Verificou-se que existem várias formas ou técnicas de análise que podem ser utilizadas para desenvolver as tarefas necessárias para avaliação de imóveis, além da estatística, tais como a inteligência artificial, a aprendizagem de máquina, a inteligência ,computacional, a mineração de dados e o descobrimento de conhecimento em base de dados. A delimitação destes campos de conhecimento não é clara, e normalmente há utilização de técnicas ou pressupostos comuns. Portanto, é interessante examiná-las antes de abordar mais detidamente o descobrimento de conhecimento, examinando as etapas que compõem o processo de análise[21].

O paradigma tradicional de análise de dados é a estatística, que conta com um largo e variado espectro de ferramentas. A análise de regressão é uma das técnicas mais conhecidas, pertencendo a um ramo denominado de estatística paramétrica. Neste segmento são exigidas diversas condições para os dados, tais como o conhecimento prévio do modelo a ser estimado e da distribuição de probabilidades dos erros produzidos pelo modelo. Outras técnicas muito utilizadas são a análise de agrupamento (*clustering*) e a análise fatorial. Geralmente as técnicas estatísticas utilizam um conjunto pequeno de dados, do qual se exige boa qualidade, e existem diversos testes para assegurar a confiabilidade dos resultados obtidos (Gujarati, 2000;

[21] O processo de descobrimento de conhecimento em bases de dados é desenvolvido com diversas técnicas de

Hair *et al.*, 1998; Hayter, 1996; Moore e McCabe, 1998). Recentemente surgiu uma nova corrente na econometria, conhecida como "abordagem inglesa", que sugere a exploração dos dados disponíveis, sem adotar inicialmente modelos restritivos (Barossi Filho e Braga, 2000; Hendry, 1988; Spanos, 1989)[22].

3.2 IDENTIFICAÇÃO DOS PARADIGMAS ALTERNATIVOS

A Inteligência Artificial (IA) é uma área de pesquisa ampla, voltada para o desenvolvimento de técnicas que permitam construir sistemas com comportamento inteligente. Envolve conhecimentos de várias áreas, tais como matemática, estatística, psicologia e ciência da computação. Há pesquisas nesta área desde a criação dos computadores, quando se pensava em reproduzir o funcionamento do cérebro humano. Porém o entusiasmo inicial arrefeceu com os maus resultados nos primeiros estudos (Adriaans e Zantinge, 1996; Dean *et al.*, 1995; Haykin, 2001; Nikolopoulos, 1997; Nilsson, 1998). Em parte, estes resultados devem-se à dificuldade da tarefa de emular o comportamento do cérebro. Por exemplo, Bell e Gray (1997) prevêem que a capacidade do cérebro humano, estimada em 10^{15} operações por segundo e memória de processamento de 10 terabytes, somente será atingida pelos computadores em 2047, se a taxa de crescimento atual for mantida. Por isto, a pesquisa em IA foi relativamente limitada por muitos anos. Esta fase prolongou-se até os anos 80, quando novos pesquisadores adotaram outro ponto de vista: ao invés de buscar sistemas que tivessem inteligência próxima da inteligência dos humanos, buscaram desenvolver programas e algoritmos que pudessem realizar tarefas práticas, resultando nos sistemas especialistas e nos algoritmos para *Machine Learning*. Também foi importante a disponibilidade de novas máquinas, mais poderosas (Adriaans e Zantinge, 1996; Haykin, 2001; Mitchell, 1999).

Podem ser identificados alguns ramos na inteligência artificial, tais como sistemas especialistas (*Expert Systems*), aprendizagem de máquina (*Machine Learning*) e inteligência

---

análise, as quais, em função do volume de informações, serão abordadas no próximo capítulo.

[22] É interessante verificar que esta nova visão da econometria aproxima essa área do descobrimento de conhecimento em bases de dados, pois admite uma modelagem mais flexível, sem a exigência de um modelo inicial embasado nas teorias existentes. Ao contrário, os modelos são guiados pelos dados.

computacional (*Soft Computing*), os quais também não têm uma delimitação clara, e podem ser reunidos em um grande grupo, todos identificados como aplicações de IA, mas com características próprias (Bonissone, 1997, Cordón *et al.*, 2001; Haykin, 2001; Mitchell, 1999; Nikolopoulos, 1997; Reich, 1997).

Os sistemas especialistas são programas de computador que reúnem conhecimento especializado, geralmente na forma de regras contendo o conhecimento elicitado dos especialistas, e desenvolvem tarefas específicas, buscando emular ou substituir um especialista humano. Os sistemas especialistas representaram as primeiras aplicações de sucesso comercial na área de inteligência artificial. Porém enfrentam o problema de extração do conhecimento dos especialistas, conhecido como "*knowledge bottleneck*". Outro problema sério é a difícil atualização dos sistemas (Nikolopoulos, 1997). Watson (1997) lembra que o desenvolvimento e a manutenção destes sistemas é um processo dispendioso, sendo mais conveniente adotar uma abordagem que permita a atualização automática, através da experiência, tal como o raciocínio baseado em casos[23]. Neste sentido, Adriaans e Zantinge (1996) afirmam que um sistema inteligente deve ter um mecanismo de aprendizagem, por isto há necessidade das técnicas de aprendizagem de máquina (ou aprendizagem artificial). Nos últimos anos, os sistemas especialistas têm sido desenvolvidos com outros formatos, mais flexíveis, prevendo mecanismos de aprendizagem e adaptação automáticas do conhecimento do domínio, incorporando também regras difusas (Cordón *et al*., 2001).

A aprendizagem de máquina (*Machine Learning*) é um ramo da inteligência artificial que se preocupa em investigar como as máquinas podem aprender e busca o desenvolvimento de algoritmos para a realização prática das tarefas de aprendizagem. Estes algoritmos envolvem a busca exaustiva de soluções em um largo espaço de soluções possíveis para determinar aquela que melhor se ajusta aos dados observados e ao conhecimento anterior. São utilizados diferentes algoritmos, tais como funções lineares, árvores de decisão, redes neurais, aprendizagem baseada em exemplos e conjuntos de regras. Há inúmeras aplicações utilizando estas ferramentas, em diversas áreas de conhecimento (Mitchell, 1999; Reich, 1997; Watson, 1997).

Uma área muito próxima à aprendizagem de máquina e aos sistemas especialistas é a

[23] *Case-Based Reasoning* (CBR).

inteligência computacional (ou *Soft Computing* – SC). Conforme Bonissone (1997), *Soft Computing* consiste na exploração ao máximo do potencial de algumas ferramentas, buscando criar máquinas inteligentes e gerando modelos através da associação de métodos computacionais. Os elementos principais dos sistemas neste campo são a lógica difusa[24], as redes neurais, os algoritmos genéticos e o raciocínio probabilístico. Geralmente os sistemas são híbridos, explorando as vantagens de cada um dos componentes. Há sinergia entre eles, segundo Bonissone (1997). As soluções baseadas na inteligência computacional têm em comum o tratamento de problemas dificilmente tratados pelos métodos computacionais tradicionais, que em geral necessitam de informações precisas e conhecimento suficiente sobre o *background*. Em outras palavras, adaptam-se a sistemas reais, os quais geralmente são mal definidos, difíceis de modelar, e com espaços de solução muito amplos (Bonissone *et al.*, 1999; Mitra *et al.*, 2002; Nguyen e Walker, 2000). Para Zadeh (1994, *apud* Bonissone, 1997)[25], a inteligência computacional é tolerante a ambientes com imprecisão, incerteza e verdade parcial. Outros autores também sugerem a combinação de técnicas em sistemas híbridos como alternativa para ampliar a confiança e a precisão das estimativas, através da compensação de falhas em uma ou outra técnica, embora nem sempre identificando estes sistemas como aplicações em *Soft Computing* (Braga *et al.*, 2000; Cordón *et al.*, 2001).

As aplicações no âmbito da inteligência artificial são focadas na solução de problemas específicos, aplicando as técnicas adequadas aos dados do domínio. As técnicas desta área em geral são não-paramétricas, ou seja, normalmente não há condições especiais a serem respeitadas pelos dados, porém exigem expressiva quantidade de exemplos e muito esforço computacional para processá-los. Geralmente o desempenho é verificado apenas pelo erro cometido pelo sistema, sem outras condições.

Uma outra vertente de soluções de problemas é um novo processo de análise de dados, conhecido como descobrimento de conhecimento em bases de dados (DCBD)[26]. Este processo busca extrair o conhecimento que se encontra oculto nas bases de dados utilizando uma ou mais técnicas, envolvendo visualização, tecnologia de bancos de dados, estatística,

[24] Tambémn conhecida como lógica nebulosa.

[25] ZADEH, L.A. Fuzzy logic, neural networks, and soft computing. **Communications ACM**, v.37, p.77-84, 1994.

[26] *Knowledge Discovery in Databases* (KDD).

inteligência artificial e outras áreas. O DCBD geralmente é definido como "o processo não-trivial de identificar padrões válidos, novos, potencialmente úteis e finalmente compreensíveis nos dados" (Frawley *et al.*, 1991[27], *apud* Fayyad *et al.*, 1996a, p.6). Há uma certa confusão com a mineração de dados (*Data Mining*), mas esta última pode ser considerada uma das etapas do processo de descobrimento de conhecimento (Adriaans e Zantinge, 1996; Fayyad *et al.*, 1996b).

Estas duas grandes áreas têm visões diferentes, mas não são dissociadas. Enquanto o foco da IA é construir sistemas inteligentes, o DCBD busca extrair conhecimento. Em certo sentido, são visões complementares. Pode-se identificar **conhecimento** com termos como informação, experiência e idéias, enquanto que a **inteligência** está associada à capacidade de aprender, à destreza de raciocínio, enfim, à capacidade de resolver problemas mediante a (re)estruturação do conhecimento disponível. Assim, DCBD e IA podem complementar-se, pois há necessidade de reunir o conhecimento e o raciocínio para solucionar problemas. Em termos mais específicos, pode-se identificar as dificuldades com os dados do mercado imobiliário como problemas de falta de conhecimento, enquanto que a obtenção de modelos preditivos é um problema de inteligência, ou raciocínio, sobre o conhecimento disponível.

3.3 O PROCESSO DE ANÁLISE NO DESCOBRIMENTO DE CONHECIMENTO

Segundo Fayyad *et al.* (1996a, p.9), "descobrimento de conhecimento em bases de dados é o processo de utilização de métodos ou algoritmos de mineração de dados para extrair ou identificar o que parece conhecimento, de acordo com especificações de medidas e limites, usando a base de dados, com qualquer pré-processamento, sub-amostragem ou transformação necessários".

Uma visão geral do processo de análise é apresentada por Brachman e Anand (1996): dada uma pesquisa ou meta, um analista consulta a base de dados, extraindo dados aparentemente relevantes. Analisa estes dados, usando diversas técnicas. Esta análise leva a algum tipo de

[27] FRAWLEY, W.J.; PIATETSKY-SHAPIRO, G.; MATHEUS, C.J. Knowledge discovery in databases: An overview. **Knowledge discovery in databases**, p.1-27. Menlo Park (CA)/Cambridge (MA):AAAI/MIT, 1991.

insight sobre os dados. O analista usa, então, algumas ferramentas de apresentação ou relatório para disseminar estes *insights* (por exemplo, para quem fez a solicitação). A apreciação destes usuários sobre os resultados pode reiniciar ou redirecionar a análise.

Para Cios *et al.* (1998), o processo de descobrimento de conhecimento é dinâmico, altamente interativo, iterativo e totalmente visualizável. Suas metas principais são extrair relatórios úteis, ressaltar tendências e eventos interessantes, apoiar o processo de decisão e explorar os dados para atingir metas científicas, comerciais ou operacionais. Na visão desses autores, o descobrimento de conhecimento em bases de dados é uma abordagem multidisciplinar, usando diversos algoritmos e métodos, sendo indicada para problemas com bases de grandes dimensões, requerendo adicionalmente uma forte revisão nos métodos anteriormente existentes para que sejam significativos neste novo contexto.

3.3.1 Motivações para o desenvolvimento de um novo processo de análise

A informatização de muitas atividades, devido à disseminação dos computadores em todos os setores da sociedade, especialmente no setor comercial, tem provocado um rápido aumento na quantidade de informações armazenadas em formato digital, fazendo com que as bases de dados atingissem dimensões que impedem a análise pelos métodos empregados anteriormente (Amaral, 2001; Berry e Linoff, 2000; Westphal e Blaxton, 1998).

Alguns fatores contribuíram para esta nova situação, tais como a automatização da entrada de dados em procedimentos corriqueiros, através de códigos de barras ou cartões magnéticos, a geração automática de imagens através de sensoriamento remoto e a circulação eletrônica de documentos e de transações comerciais na Internet. Por exemplo, Cios *et al.* (1998) afirmam que a rede de lojas Wal-Mart gera 20 milhões de transações por dia, e que o novo sistema da NASA produz 50 GB de dados de imagem por hora. Para Witten *et al.* (1999), a Internet e o CD-ROM são a maior revolução na área de informação, superando largamente o impacto do surgimento dos computadores pessoais. Há muitas atividades executadas permanentemente *on-line*. Os mesmos autores afirmam que havia quatro Terabytes de dados na Internet em 1999, estimando que esta quantidade dobra a cada seis meses, com vida média de 75 dias para os documentos, e que a quantidade total de informação existente no mundo dobra a cada 20 meses.

As grandes bases têm estimulado a adoção de técnicas automatizadas de análise. Neste sentido, Cios *et al.* (1998) afirmam que atualmente há bases de dados enormes à disposição, mas a busca de padrões através de métodos manuais ou de estatística convencional é muito difícil, em função do nível de detalhamento, agilidade e custo que são exigidos das aplicações comerciais em ambientes de forte e globalizada concorrência. À medida que as bases de dados crescem, torna-se cada vez mais importante o uso de métodos automatizados de descobrimento de conhecimento. Diante deste contexto, fica clara a necessidade de buscar técnicas e ferramentas que possam analisar grandes bases de dados, de forma razoavelmente automatizada e inteligente, visando a encontrar elementos de conhecimento útil (Carvalho, 2001; Fayyad *et al.*, 1996b; Han e Kamber, 2001; Mitchell, 1999).

3.3.2 Terminologia

Existe alguma confusão entre os termos descobrimento de conhecimento em bases de dados (DCBD, ou *knowledge discovery in databases, KDD*) e mineração de dados (*data mining*). Conforme Fayyad *et al.* (1996a), a idéia de encontrar padrões úteis nos dados brutos, armazenados em bases de dados, recebeu historicamente diversos nomes, tais como descobrimento de conhecimento em bases de dados, mineração de dados, extração de conhecimento, descobrimento de informação, colheita de informação, arqueologia de dados ou processamento de padrões de dados. Segundo estes autores, o termo DCBD surgiu em 1989, referindo-se ao processo amplo de busca de conhecimento nos dados, através da aplicação de uma ou mais técnicas de mineração de dados, e está vinculado ao processo geral de descoberta de conhecimento útil, enquanto que a mineração de dados refere-se à aplicação específica de algoritmos para a extração de padrões nos dados, contudo sem envolver os passos adicionais do descobrimento de conhecimento em bases de dados, tais como considerar conhecimentos anteriores apropriados ou interpretar os resultados. Esses passos adicionais são essenciais para assegurar que a informação útil (conhecimento) seja extraída dos dados e chegue ao usuário final. Estes autores enfatizam que a mineração de dados "é um passo do processo de descobrimento de conhecimento em bases de dados, consistindo de algoritmos particulares que, sob algumas limitações computacionais aceitáveis, produzem uma enumeração particular de padrões a partir dos dados". Vários outros autores seguem esta visão, tais como Aamodt *et al.* (1998) , Adriaans e Zantinge (1996), Amaral (2001); Cios *et*

al. (1998), Han e Kamber (2001); Klemettinen *et al.* (1997), Kodratoff (1995), Mannila (1997) e Soibelman e Kim (2002).

Por outro lado, existem alguns autores que preferem a expressão “mineração de dados”, afirmando que o descobrimento de conhecimento é uma das metas da mineração de dados, tais como Berry e Linoff (2000), Carvalho (2001); Hand *et al.* (2001); Weiss e Indurkhya (1998), Westphal e Blaxton (1998) e Witten e Frank (2000). Segundo Fayyad *et al.* (1996a), o termo mineração de dados tem sido utilizado por estatísticos, analistas de dados e gerentes de sistemas de informação, enquanto que descobrimento de conhecimento em bases de dados é mais usado por pesquisadores.

3.3.3 Aplicações de DCBD

O descobrimento de conhecimento em bases de dados pode ser empregado em situações muito diferentes. Um dos exemplos mais comuns de aplicação nesta área é a busca de padrões de transações em grandes bases de dados em bancos ou supermercados. Estes padrões podem ser utilizados para incrementar o relacionamento com o consumidor ou para decidir sobre o lançamento de novo produtos (Klemettinen *et al.*, 1994).

Para Cios *et al.* (1998), há interesse em vários campos em função das possíveis aplicações. Na área comercial, os dados podem conter informações sobre mercados, consumidores e competidores. No setor industrial, busca-se novas tecnologias, maior quantidade dos produtos e melhor performance de todo o processo, e na ciência, busca-se entender um fenômeno, estabelecendo novas direções ou sugerindo a exploração de novos aspectos. Alguns exemplos de aplicação em casos reais são os seguintes.

Hätönen *et al.* (1996) indicam a análise de grandes bases de dados com casos de alarme (situação anormais), com tipicamente milhares de alarmes por dia, de milhares de tipos diferentes. A identificação e correção de falhas é uma tarefa crítica de gerenciamento nesta área. Eles usaram a técnica de regras episódicas (uma modificação da técnica de regras de associação), obtendo regras do tipo "em 23% dos casos em que ocorre 'alarme' e 'falha de ligação' em um intervalo de 20 segundos, também ocorre 'alta taxa de falha', em 40 segundos".

Klemettinen *et al.* (1997) apresentaram um exemplo de aplicação de regras de associação, visando analisar os padrões de matrículas em disciplinas dos cursos de computação na Universidade de Helsinque. Geraram regras do tipo "84% dos alunos que cursam 'Introdução ao Unix', cursam também 'Programação em C', e 34% cursam as duas simultaneamente".

Li e Biswas (1995) desenvolveram um estudo sobre localização de plataformas de exploração de petróleo, buscando aumentar as chances de acerto, que geralmente são de 10%. Estes autores utilizaram análise de clusterização para segmentar uma base de dados em grupos relativamente homogêneos, para os quais foram estimadas equações individuais.

Soibelman e Kim (2002) apresentaram um protótipo para uso de DCBD na análise de bancos de dados do *U.S. Army Corps of Engineers*, investigando causas para o sistemático atraso de obras nas quais as escavações eram elemento importante do cronograma. Foi possível identificar o desconhecimento do subsolo como principal problema. Mesmo com a realização de sondagens prévias, existiam matacões e velhas tubulações, que exigiam freqüentemente a intervenção de máquinas especiais e alterações de projeto.

3.4 ETAPAS DO DESENVOLVIMENTO DA ANÁLISE NO DESCOBRIMENTO DE CONHECIMENTO EM BASES DE DADOS

Um dos elementos importantes é que o DCBD é um processo centrado no analista, segundo Brachman e Anand (1996). Para estes autores, é importante definir quem é o usuário do processo, pois ele está envolvido em quase todas as etapas. Eles relatam que a maior parte das aplicações consistem de três partes, basicamente: (a) descobrimento inicial de conhecimento, que é trabalhoso e é realizado por alguém que entende o domínio e também as ferramentas de análise; (b) organização do conhecimento descoberto em uma arquitetura específica de solução de problemas; e (c) aplicação do conhecimento descoberto no contexto de uma aplicação real, por uma classe bem definida de usuários finais, geralmente gerentes.

Existem vários entendimentos de quais sejam as etapas deste processo, conforme a fonte consultada. Apresentam-se a seguir algumas delas, iniciando por uma definição ampla dos passos básicos do processo de DCBD, proposta por Fayyad *et al.* (1996a):

a) entendimento do domínio da aplicação, do conhecimento inicial relevante e das metas para o usuário final;

b) criação do conjunto de dados-alvo: amostragem ou seleção dos dados nos quais o descobrimento será realizado;

c) limpeza e pré-processamento dos dados: envolve operações básicas como remover ou tratar ruídos ou *outliers*, decisão sobre estratégias para lidar com dados omissos e seqüências temporais ou variações conhecidas;

d) projeção e redução de dados: encontrar características úteis para representar os dados e, dependendo da meta a ser atingida, transformar os dados ou reduzir as dimensões da base, para reduzir o número efetivo de variáveis sob consideração;

e) definição da tarefa de mineração de dados: decidir se a meta do processo de DCBD é classificação, regressão, clusterização, etc;

f) escolha dos algoritmos de mineração: selecionar os métodos mais adequados para a tarefa a ser realizada;

g) mineração dos dados, propriamente dita: buscar os padrões interessantes;

h) interpretação dos padrões encontrados, com possível retorno a qualquer dos passos anteriores;

i) consolidação do conhecimento descoberto: incorporar este conhecimento ao sistema ou documentar e relatar às partes interessadas. Também inclui conferir e resolver conflitos com conhecimentos anteriores.

Outros autores dividem as etapas do processo de descobrimento de conhecimento de forma mais agregada. Aamodt *et al.* (1998) indicaram apenas três passos, incluindo preparação e seleção dos dados, mineração e pós-processamento para apresentação da informação obtida ao usuário. Ester *et al.* (2000) e Han e Kamber (2001) resumiram o processo de DCBD em quatro passos: (a) seleção de um subconjunto de variáveis; (b) redução dos dados, através de transformações ou redução de dimensões, diminuindo o número de atributos a ser

considerado; (c) *data mining* (aplicação dos algoritmos apropriados); e (d) avaliação e interpretação dos padrões encontrados, com respeito à sua utilidade.

Para Mannila (1997), o processo de DCBD é desenvolvido através dos seguintes passos: entendimento do domínio, preparação do conjunto de dados, descobrimento de padrões (mineração dos dados), pós-processamento do conhecimento descoberto e colocação dos resultados em uso. Cios *et al.* (1998) indicaram os mesmos cinco passos. Para Klemettinen *et al.* (1997), o descobrimento de conhecimento em bases de dados é composto de preprocessamento, transformação dos dados, descobrimento de padrões, apresentação e utilização dos resultados. Já Cabena *et al.* (1997) e Soibelman e Kim (2002) indicam cinco passos: identificação do problema, preparação dos dados, mineração dos dados, análise dos resultados e refinamento do processo. Analisando as indicações destes autores, as etapas a serem realizadas podem ser adaptadas ou resumidas como exposto na Figura 1, e detalhado a seguir.

O início da análise compreende o projeto da aplicação, e geralmente é construído com base nas necessidades do usuário final. Em seguida são desenvolvidas as atividades de preparação e mineração dos dados. A preparação é necessária para garantir a qualidade dos dados, facilitando a modelagem. A mineração dos dados consiste na extração de padrões ou determinação de modelos sobre o comportamento dos dados (Berry e Linoff, 2000; Han e Kamber, 2001). Estas duas etapas estão desenvolvidas em maior detalhe adiante, em função da importância para a aplicação desenvolvida.

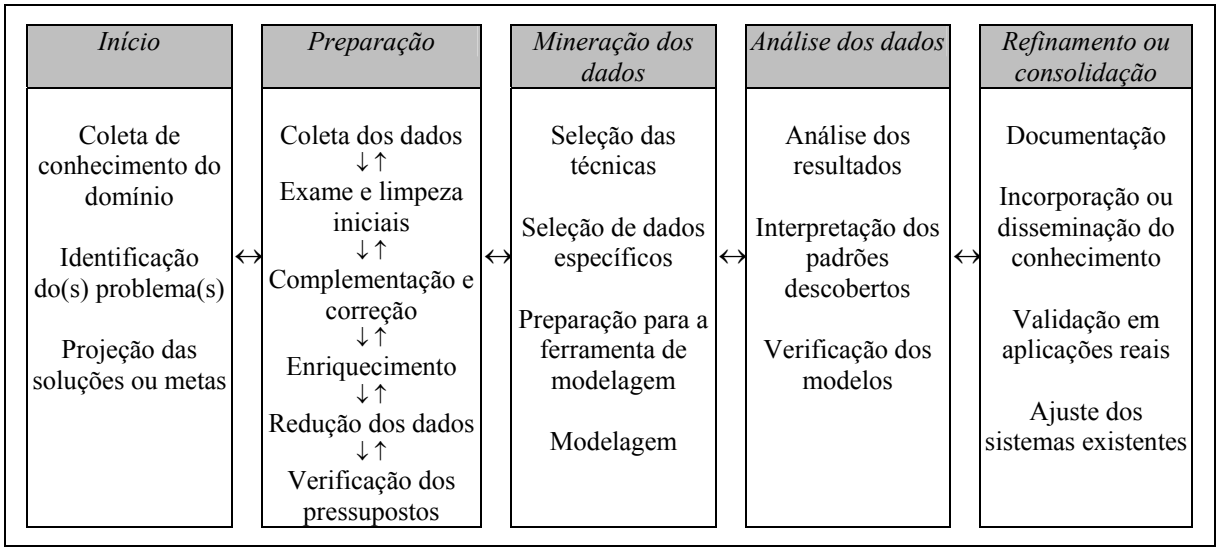


Figura 1: O processo de análise no descobrimento de conhecimento

em bases de dados (adaptado de Cabena *et al.*, 1997 e Soibelman e Kim, 2002)

Após a modelagem, os resultados devem ser examinados, verificando-se a adequação dos padrões ou modelos gerados. Uma das formas de verificar os resultados é através do teste do modelo usando uma parte dos dados, reservada especialmente. O conhecimento obtido pode estar em um formato de difícil utilização, tal como em amplos conjuntos de regras ou nos pesos de uma rede neural. Assim, pode ser necessário desenvolver a simplificação, poda ou refinamento do conhecimento gerado. Por fim, conclui-se o descobrimento de conhecimento com a comunicação ou disponibilização das informações aos usuários finais, através de relatórios, gráficos ou outros instrumentos. Estes usuários serão responsáveis pela interpretação e verificação da real utilidade dos conhecimentos. Como se trata de um processo cíclico e dinâmico, o usuário poderá sugerir a necessidade de atualização ou de aperfeiçoamentos nos modelos (Berry e Linoff, 2000; Bruha, 2001; Cios *et al.*, 1998).

3.4.1 Preparação dos dados

Algumas aplicações de DCBD lidam com bases de dados especialmente projetadas, com as entradas já em formato eletrônico enviadas diretamente às bases de dados. Existe um baixo nível de ruído no processamento dos dados. Em outros casos, porém, os dados não são gerados automaticamente e os erros, omissões e inconsistências nos dados são muito comuns, em função da participação humana na geração dos dados (Han e Kamber, 2001). Os dados imobiliários são deste segundo tipo e contam com um componente "social", pois o comportamento dos agentes não é completamente racional, sendo que as negociações geralmente conduzem a resultados bastante diferentes das ofertas e contra-ofertas iniciais. A dispersão dos preços praticados em relação às médias ou valores de equilíbrio adiciona ruído ao processo (o componente aleatório dos preços). Ademais, existem diversos outros problemas, tais como erros de transcrição, erro ou omissão de informações, diferenças devidas a peculiaridades das fontes dos dados, etc.

A preparação dos dados consiste de uma seqüência de operações destinadas a converter dados brutos ou outras características dos casos em um formato adequado para as tarefas de

processamento (Hair *et al.*, 1998; Pyle, 1999).

Assim, a preparação dos dados é muito importante para aplicações no mercado imobiliário. Entretanto, nos estudos empíricos sobre o mercado imobiliário não foram encontradas aplicações com uma metodologia clara de tratamento dos dados, que geralmente é desenvolvida de forma expedita, apenas com a remoção de *outliers* e a escolha subjetiva dos casos e dos atributos a serem utilizados nos modelos. Na área de DCBD existem descrições para a preparação dos dados, mas sem especial menção aos dados de mercado imobiliário, sugerindo a necessidade da análise da adequação dos processos para o mercado imobiliário (Berry e Linoff, 2000; Cabena *et al.*, 1997; Cios *et al.*, 1998; Fayyad *et al.*, 1996a; Han e Kamber, 2001; Weiss e Indurkhya, 1998).

A exploração dos dados inicia antes mesmo de se ter os dados em si, segundo Pyle (1999). Para este autor, é preciso preparar o terreno, reunindo conhecimento sobre o domínio antes de realizar a aquisição dos dados. O analista precisa conhecer os parâmetros importantes da análise para que possa tomar decisões sobre quantidade de dados, necessidade ou importância de cada variável, considerando também os custos de obtenção, entre outros elementos. Ademais, segundo o mesmo autor, os usuários finais são consumidores de informação, e precisam saber qual a informação disponível, sua faixa de aplicação, prazo e limites de uso, retorno provável, custo de obtenção e outros detalhes. Com este conhecimento, poderão ser escolhidas as informações mais adequadas. Para Weiss e Indurkhya (1998), a preparação de dados tem duas metas básicas:

a) organizar os dados em um formato padrão, adequado ao processamento pelos programas de modelagem;

b) preparar e selecionar características que levem a um melhor desempenho na modelagem.

Esta etapa, embora freqüentemente desprezada, geralmente é responsável pela maior parte do tempo e esforço empregados na análise (Weiss e Indurkhya, 1998). Efetivamente, segundo Cabena *et al.* (1997) e Pyle (1999), a preparação dos dados é responsável por cerca de 60% do tempo e do esforço no processamento, enquanto que a fase de mineração dos dados é responsável por cerca de 5 a 10% do tempo gasto.

3.4.1.1 Tipos de dados

É interessante diferenciar dados, informação e conhecimento. Conforme Pyle (1999), um conjunto de dados pode apresentar a informação de uma forma que os analistas não podem compreender diretamente, principalmente no caso de grandes bases de dados. Assim, a informação pode não estar disponível, embora os dados estejam, exigindo o uso das técnicas adequadas para ressaltar esta informação. Já o conhecimento consiste na identificação da parcela útil das informações, ou seja, é a apropriação ou percepção dos significados contidos nas informações.

No âmbito do DCBD, as informações geralmente são armazenadas na forma de casos, cada um deles contendo um conjunto de atributos, codificados em alguns formatos convencionais. Os casos, também chamados de observações, padrões, exemplos ou objetos, são os elementos de análise, os quais devem refletir a realidade que está sendo modelada. Um caso pode ser entendido como a representação ou caracterização de uma entidade, de um conceito abstrato, de um objeto físico, ou uma experiência. A descrição de um caso pode conter diversos atributos ou características, descrita genericamente como um vetor, tal como $X = \{x_1,..., x_n\}$, onde X é o caso, e x_i (i=1,...,n) são os atributos (Cios *et al.*, 1998; Pyle, 1999).

Um atributo (também chamado de característica ou variável) é um elemento da caracterização de um caso. Há dois tipos principais de atributos, conforme Hair *et al.* (1998) e Pyle (1999):

a) quantitativos (numéricos) – divididos em atributos contínuos (assumem valores reais) e discretos (números inteiros ou códigos binários);

b) qualitativos (não-numéricos) – simbólicos.

A maior parte das ferramentas de análise não aceita dados simbólicos, exigindo a conversão para formatos numéricos. Segundo Hair *et al.* (1998), as variáveis não-numéricas estão inicialmente em escalas nominais (com categorias não ordenadas) ou ordinais (categorias ordenadas, mas sem afastamento definido entre as categorias), e podem ser consideradas na análise através de uma ou mais variáveis binárias (também chamadas de *dummy* ou dicotômicas) ou através de variáveis discretas, com uma codificação definida pelo analista.

Para um conjunto de k categorias ou situações, utiliza-se k-1 variáveis binárias e a categoria excluída é chamada de "grupo de comparação". A inclusão de k variáveis geralmente provoca erros nas técnicas de modelagem, pois há uma combinação linear entre as variáveis do conjunto.

3.4.1.2 Etapas da preparação de dados

A seqüência de análise depende do tipo de dados e da meta de processamento, e é bastante distinta para processamento de imagens, séries temporais ou outros dados. Basicamente a preparação inicia com a coleta e armazenamento dos dados, que podem conter informações obtidas em diferentes estágios do processamento, inclusive com dados armazenados em formatos específicos e, ao final, deverá estar disponível um arquivo com dados preparados para a análise (Cios *et al.*, 1998; Pyle, 1999).

Cios *et al.* (1998) denominam estes dados de intermediários, enquanto que Pyle (1999) fala em um ambiente de informação preparada, que inclui os dados e também as ferramentas de análise. Já Weiss e Indurkhya (1998) utilizam a expressão "formato padrão". Seguindo as indicações destes autores, a seqüência de preparação pode envolver as seguintes etapas:

a) coleta dos dados:

- reconhecimento dos dados brutos: identificação das fontes, significado dos dados, limitações, escopo (abrangência espacial e temporal);
- dimensionamento e extração da amostra;
- formatação ou conversão de formatos dos arquivos;

b) exploração e limpeza iniciais:

- identificação dos atributos (significado dos campos);
- análise geral dos dados (exploração);
- limpeza inicial (identificação e remoção de erros grosseiros);

c) complementação e correção: identificação e soluções para dados omitidos ou com falhas;

d) enriquecimento: acréscimo de atributos de outras fontes ou extraídos dos dados brutos (guiado por conhecimento anterior ou extraído diretamente dos dados,

sem exame da relevância para a modelagem);

e) redução da base de dados (seleção de informação relevante):

- redução horizontal: identificação e seleção dos atributos mais relevantes;
- redução vertical (casos): identificação e soluções para *outliers*, identificação e seleção dos casos mais relevantes (sub-amostras);
- redução interna dos atributos: diminuição da variação numérica (suavização ou outras transformações);

f) verificação dos pressupostos ou condições da ferramenta de análise da mineração de dados.

3.4.1.2.1 Coleta dos dados

Esta etapa consiste da identificação das possíveis fontes de dados, exame dos dados existentes (identificação das limitações), obtenção de autorização para consulta, dimensionamento e extração de uma amostra, conversão dos arquivos para o sistema a ser utilizado na análise e armazenamento inicial (Hair *et al.*, 1998; Pyle, 1999).

3.4.1.2.2 Exame inicial dos dados

O exame dos dados é uma tarefa demorada, mas necessária. A análise cuidadosa pode levar a melhores predições e uma avaliação mais precisa da dimensionalidade dos dados. As técnicas gráficas, por exemplo, propiciam ao pesquisador meios simples e compreensivos para examinar as variáveis e os relacionamentos entre elas. Na análise inicial, são importantes os passos de exame da natureza dos dados, identificação de *outliers*, avaliação de dados incompletos ou omitidos, e o teste das condições (pressupostos) das técnicas a serem utilizadas (especialmente no caso de técnicas estatísticas) (Hair *et al.*, 1998).

Para Moore e McCabe (1998), o exame inicial dos dados consiste no processo conhecido na área estatística como "análise exploratória de dados", na qual as duas estratégias básicas são:

a) examinar as variáveis individualmente, de início, e só depois analisar em

conjunto;

b) iniciar a análise com gráficos e em seguida utilizar sumários numéricos dos dados.

Hair *et al.* (1998) indicam o esforço demandado para o exame dos dados como uma garantia dos resultados da análise posterior. Mesmo que a técnica escolhida apresentasse bons resultados, os problemas que não foram detectados nos dados poderiam abalar os resultados.

Segundo Pyle (1999), a preparação dos dados ajuda o analista a obter resultados melhores e mais rápidos na modelagem. Este autor lembra que algumas das técnicas utilizadas no descobrimento de conhecimento na verdade têm sido usadas há anos e outras evoluíram ou foram desenvolvidas nas duas últimas décadas. O que não mudou, e que é quase uma lei da natureza, segundo o mesmo autor, é o aforismo "*garbage in, garbage out*", indicando a necessidade de bons dados como pré-requisito para obtenção de bons modelos. Embora alguns problemas dos dados possam ser corrigidos ou remediados, a solução nem sempre é fácil.

O objetivo do exame inicial é o entendimento dos dados coletados, verificando as variáveis e identificando suas limitações e potencialidades, de forma a guiar o restante da análise. Esta etapa é composta de uma exploração inicial, incluindo a identificação dos atributos existentes (significados dos campos), exploração dos dados (usando técnicas de visualização e estatística descritiva), e uma limpeza inicial (remoção dados com erros grosseiros). Durante a análise inicial é importante examinar a natureza dos dados, verificando a presença de inconsistências entre fontes e no tempo, de *outliers* ou de dados omitidos ou incompletos, e verificando as condições da ferramenta de modelagem, identificando problemas a serem resolvidos nas etapas seguintes. O objetivo é preparar os dados e o analista simultaneamente. Ampliando seu conhecimento sobre os dados, provavelmente as decisões do analista serão melhores (Hair *et al.*, 1998; Han e Kamber, 2001).

3.4.1.2.3 Complementação e correção dos dados

É comum a ocorrência de dados omitidos. A primeira ação é buscar os motivos (ampliando o conhecimento) e, secundariamente, preencher os valores, se for possível. Para Hair *et al.*

(1998), existem vários tipos ou categorias de dados omitidos: dados ignoráveis (a situação não se aplica àquele caso), censurados (existe algum impedimento legal ou de conveniência da fonte dos dados), procedimentais (erros de entrada, tais como códigos inválidos ou dificuldade de entendimento de formulários) e negativa de fornecimento (em caso de questões controversas ou declaração de renda, por exemplo).

Pyle (1999) diferencia dois tipos de dados omitidos, denominando de **ausentes** os que não tiveram o valor real registrado pelo sistema, e **vazios** aqueles para os quais não há realmente uma medida. Este autor lembra que é necessário resolver o problema antes da modelagem, mas que a própria ausência de valores pode ser uma informação interessante, tal como em um formulário sistematicamente preenchido de forma incompleta. Também recomenda que a informação sobre o que está faltando seja retida, pois é importante entender os padrões de omissão.

Os valores de substituição não devem introduzir tendências, embora seja comum que isto ocorra. As melhores formas são aquelas que são bem compreendidas pelo analista e estão sob seu controle (Pyle, 1999). Se existe aleatoriedade nas omissões, é mais fácil obter soluções, que podem ser embasadas no exame dos padrões de distribuição dos dados. Algumas das soluções são as seguintes (Hair *et al.*, 1998; Pyle, 1999):

a) utilizar apenas os dados completos: adequado quando existem poucos casos com problemas. Se as omissões não são aleatórias, mas concentradas em determinadas parcelas dos dados, os resultados da modelagem podem ser afetados;

b) imputação de padrões: podem ser utilizados desvio-padrão, média ou correlações dos dados completos, como representantes para toda a amostra. Não são estimados valores individuais;

c) preenchimento utilizando o conhecimento disponível: os casos com problemas podem ser substituídos por casos completos, pode-se preencher os valores omitidos com uma constante ou com a média dos casos válidos, ou utilizar uma técnica de estimação, tais como redes neurais, vizinhança próxima ou regressão para estimar os valores correspondentes. Estas soluções podem gerar ou

reforçar distorções, alterando o comportamento da amostra;

d) modelagem interativa: os casos com dados omitidos são utilizados na geração de modelos que estimam estes valores. Os modelos são re-estimados até que as variações sejam mínimas. Pode-se utilizar o conjunto total de dados ou parte da amostra.

Também pode ser utilizada uma combinação destas técnicas, alternando as soluções ou usando a média das estimativas obtidas. Porém, para qualquer técnica, a estimação de dados numéricos é mais fácil do que para dados não-numéricos (Hair *et al.*, 1998).

3.4.1.2.4 Enriquecimento

Algumas informações auxiliares podem ser acrescentadas à base de dados, incluindo na análise elementos de fontes diversas, tais como Censos, pesquisas de organismos profissionais e conhecimento geral do mercado. Podem ser realizadas combinações de variáveis usando produtos e razões, ou convertendo escalas de medida. Esta etapa envolve o acréscimo sem o teste empírico de relevância, de forma univariada, apenas acrescentando atributos que aparentemente são úteis (Hair *et al.*, 1998; Pyle, 1999).

3.4.1.2.5 Redução nos dados

O excesso de tamanho da base de dados é considerado um problema, porque aumenta o tempo de análise (tempo de máquina e do próprio analista) e a complexidade de modelagem. Por outro lado, informações redundantes ou desnecessárias (irrelevantes para a aplicação em questão) podem ser especialmente problemáticas porque a performance da maioria das ferramentas é afetada pelo tamanho e qualidade da base de dados. Assim, a redução de dimensionalidade é interessante, para reduzir o espaço de soluções possíveis, auxiliando os algoritmos a operarem mais rápida e efetivamente. Em alguns casos, a precisão pode aumentar e, em outros, a representação do conhecimento será mais compacta ou mais fácil de interpretar (Hall, 2000; Pyle, 1999).

Para Blum e Langley (1997), com o incremento constante de tamanho das tarefas

apresentadas às ferramentas de *Machine Learning*, o problema de focar a solução na informação mais relevante torna-se progressivamente mais importante.

Muitos fatores afetam o sucesso da aprendizagem para uma dada tarefa e a qualidade dos dados é um dos elementos mais importantes, Se a informação é irrelevante ou redundante, ou se os dados são inconfiáveis ou têm muito ruído, o descobrimento de conhecimento durante o treinamento é mais difícil (Hall, 2000; Han e Kamber, 2001).

Uma parte desta tarefa é realizada no exame inicial dos dados, com a identificação de erros de digitação ou variáveis praticamente constantes, na fase de limpeza dos erros grosseiros. A redução da base pode ser realizada nos atributos, nos casos ou nas variações internas dos valores assumidos pelos atributos. O número de valores distintos que um atributo pode assumir pode ser suavizado ou mesmo convertido em variáveis discretas (um caso extremo é a conversão de variáveis discretas com várias categorias em um conjunto de variáveis binárias), conforme Pyle (1999) e Weiss e Indurkhya (1998).

A dimensionalidade pode ser reduzida sem afetar a informação presente, através da exclusão ou combinação de variáveis altamente correlacionadas. A análise de componentes principais é outra alternativa, convertendo um grupo de atributos que apresenta multi-colinearidade em um novo conjunto de atributos, os quais são não-colineares. Assim, além de se reduzir o conjunto de dados, os riscos de multicolinearidade são diminuídos. Pode-se estipular quantos fatores devem ser gerados ou buscar a combinação de melhor desempenho no modelo preditivo. A dificuldade com esta técnica é que as novas variáveis não têm relação aparente com as originais, e a interpretação e comunicação aos usuários torna-se difícil (Han e Kamber, 2001; Pyle, 1999).

A seleção de casos por amostragem ou seguindo algum critério de similaridade é utilizada na produção de modelos específicos (por exemplo, na avaliação individual). Assim, as duas etapas importantes na redução dos dados são a seleção de atributos e a investigação de *outliers*. São operações quase simultâneas e interdependentes. A explicação de casos suspeitos (potenciais *outliers*) pode ser proporcionada por um modelo mais ajustado aos dados, construído com um conjunto de atributos refinado, mas a eliminação de alguns casos problemáticos pode melhorar o modelo. É um processo iterativo, pois cada decisão tomada sobre manutenção ou eliminação de casos e atributos depende do conhecimento disponível

sobre os resultados prévios, mas também afeta as etapas seguintes (Han e Kamber, 2001; Pyle, 1999).

3.3.1.2.5.1 Seleção de um subconjunto de atributos

A seleção de atributos pode ser definida como a busca de um subconjunto ótimo de atributos, ou como o processo de identificar e remover tanto quanto possível à informação redundante (Hall, 2000). Para Kohavi e John (1997), o acréscimo de variáveis desnecessárias diminui a precisão de diversos algoritmos de aprendizagem, ou seja, o erro pode ser reduzido com a redução do número de atributos. Já Hallinan e Jackway (1999) afirmam que freqüentemente desconsidera-se que o subconjunto ótimo de atributos depende das características do conjunto de dados de treinamento e da ferramenta de aprendizagem utilizada. Por exemplo, um atributo pode ter desempenho fraco em sistemas lineares, tal como análise de discriminante linear, mas ser adequado para redes neurais.

Os atributos podem ser classificados como relevantes, fracamente relevantes ou irrelevantes. Existem várias formas de selecionar atributos, as quais podem ser agrupadas em basicamente três categorias: abordagens de filtro, envoltório (*wrapper approach*) e embutidas, segundo Blum e Langley (1997), John *et al.* (1994), e Kohavi e John (1997).

Quando a seleção de atributos é realizada pelo próprio algoritmo de aprendizagem, como parte do processo de aprendizagem, Blum e Langley (1997) denominam o mecanismo de seleção de embutido.

A abordagem de filtro consiste em identificar ou classificar os atributos de acordo com algum critério de medida, e repassar apenas atributos relevantes (aqueles que ultrapassam o nível mínimo especificado) para o algoritmo de aprendizagem (John *et al.*, 1994). Neste caso, a seleção ocorre antes do algoritmo, como uma etapa independente (Figura 2). O exemplo mais comum de filtro é a matriz de correlações, identificando os relacionamentos mais fortes entre variáveis dependentes e independentes e também identificando variáveis independentes colineares (variáveis que apresentam aproximadamente a mesma informação). No caso da correlação, costuma-se utilizar um limite definido *a priori*, tal como 0.8, para indicar relacionamentos importantes (Diaz, 2000). Um tipo especial de algoritmos de filtro são os que substituem os atributos originais por outros, tais como os componentes principais (Blum e

Langley, 1997; John *et al.*, 1994; Kohavi e John, 1997).

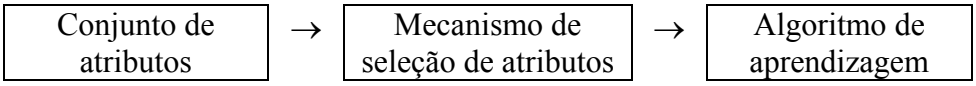


Figura 2: Seleção de atributos por filtro (baseada em Kohavi e John, 1997).

Para grandes bases de dados, os filtros são mais rápidos e práticos. A seleção pode ser realizada pela busca exaustiva (teste de todos os possíveis subconjuntos) ou heurística. Na presença de muitos atributos a busca exaustiva é inviável, na prática, e podem ser empregados métodos heurísticos, tais como Relief, Preset, LVF e CFS (Hall, 2000; Liu e Setiono, 1996; Robnik-Sikonja e Kononenko, 1997).

Na abordagem por envoltório, o algoritmo de seleção de subconjuntos de atributos utiliza a própria ferramenta de análise como sub-rotina para o exame das variáveis, como parte da função de avaliação (Figura 3). A avaliação utiliza apenas um subconjunto de atributos por vez. Como nas outras abordagens, a seleção também ocorre antes da aplicação da fase de aprendizagem, mas utiliza a ferramenta de aprendizagem como uma sub-rotina do processo de seleção, empregando os indicadores de precisão desta ferramenta como métrica de seleção. A idéia por trás desta abordagem é que os métodos de indução que irão usar o subconjunto de atributos proporcionam uma estimativa melhor de precisão do que uma medida independente, que poderá ter um viés de indução diferente. Porém, o próprio algoritmo de aprendizagem poderá ter tendências diferentes, e a utilização de diferentes algoritmos poderá levar a diferentes conjuntos de atributos selecionados (Blum e Langley, 1997; John *et al.*, 1994; Kohavi e John, 1997; Liu e Setiono, 1996).

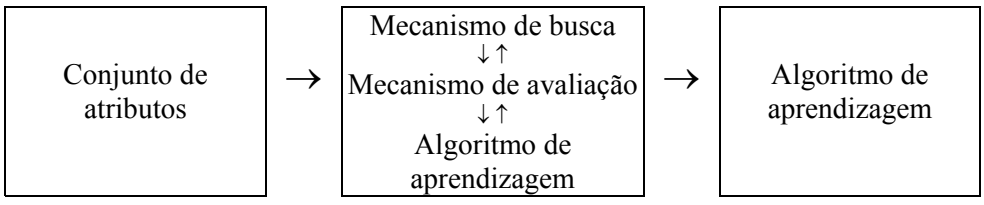


Figura 3: Seleção de atributos por envoltório (baseada em Kohavi e

John, 1997 e Wang *et al.*, 1997)

Quando o objetivo da análise é classificação, são utilizadas árvores de decisão, vizinhança próxima e redes neurais. Na tarefa de predição de valores de atributos contínuos, as ferramentas normalmente utilizadas na abordagem de envoltório são a análise de regressão e as redes neurais (Blum e Langley, 1997; Hall, 2000; Kohavi e John, 1997; Nath *et al.*, 1997). A seleção de atributos pode ser realizada pela estatística *t* calculada para cada atributo na regressão ou através de medidas baseadas nos pesos das redes neurais, como, por exemplo, o Índice Geral (GI), apresentado na Equação 3, o qual foi proposto por Howes e Crook (1999):

$$GI(x_{i*}) = \sum_{j=1}^{J} \left| \left(w_{i*j} \Big/ \sum_{i=0}^{I} \left| w_{ij} \right| \right) . w_{jk} \right| \Big/ \sum_{j=0}^{J} \left| w_{jk} \right| \qquad \text{(Equação 3)}$$

Onde $GI(x_{i*})$ é a influência potencial de cada neurônio de entrada, x_{i*} é a entrada em análise, w_{ij} e w_{jk} são os pesos entre as camadas de entrada e oculta, e entre as camadas oculta e de saída, respectivamente, I é o número de entradas (i=1,...,I), J é o número de neurônios ocultos (j=1,...,J), e existe apenas um neurônio na camada de saída (k=1). Nas duas camadas, o termo de *bias* é identificado pelo subscrito 0 (i=0, j=0), conforme Howes e Crook (1999).

A busca do subconjunto ideal de atributos é guiada por um mecanismo de busca, tais como o *stepwise* ou algoritmos genéticos. O algoritmo *stepwise* é bastante conhecido na estatística, sendo implementado em muitos pacotes de regressão para seleção automática de atributos (Neter *et al.*, 1990; Nie *et al.*, 1975).

3.3.1.2.5.2 Identificação de *outliers*

Outra questão importante, especialmente com dados do mercado imobiliário, é a investigação de *outliers*, que é uma segunda etapa de limpeza dos dados. Os *outliers* são observações diferenciadas, com uma combinação particular de características. Não podem ser categoricamente definidos como bons ou ruins. Ao contrário, devem ser analisados no contexto, verificando-se o potencial de informação que podem conter. Por exemplo, podem ser indicativos de novas categorias ou situações, desconhecidas ou indicativas de alterações na realidade conhecida anteriormente (Hair *et al.*, 1998).

Para Moore e McCabe (1998), um *outlier* é uma informação que excede os limites, considerando o padrão das outras informações (é grande no sentido do eixo Y), enquanto que uma observação influente é aquela que afeta significativamente uma análise, alterando o resultado final (geralmente é grande no eixo X).

Podem ser problemas sérios na análise, caso não representem situações reais. Por outro lado, podem indicar modelos incorretos ou incompletos, que não conseguem explicar alguns casos. Neste caso, remover os dados não irá melhorar o modelo, ao contrário, o afastará da realidade que se deseja modelar.

De qualquer forma, geralmente as técnicas de modelagem são sensíveis a dados diferenciados, e é necessário analisar os dados, verificando se existem tais tipos de dados. Hair *et al.* (1998) classificam os *outliers* em quatro categorias:

a) procedimentais: erros de digitação, leitura ou codificação. São identificados na primeira fase de análise (limpeza dos dados) e podem ser eliminados ou tratados como dados omitidos;

b) eventos extraordinários com explicação: existindo conhecimento suficiente, o analista deve decidir se o caso é representativo (mantendo) ou não (removendo o caso);

c) eventos extraordinários sem explicação: neste caso, geralmente os dados são excluídos, mas ainda podem indicar segmentos válidos da população, havendo interesse no conhecimento potencial que representam;

d) casos aparentemente válidos: os valores estão dentro das faixas de variação das variáveis, mas a análise conjunta (multivariada) indica combinações não válidas. Neste caso, as observações devem ser mantidas, até que haja evidências de que não são casos válidos.

Embora algumas técnicas sejam mais robustas à presença de dados espúrios (*outliers*), é conveniente remover os dados potencialmente problemáticos. Porém, esta é uma tarefa delicada, pois se a análise está se iniciando, é possível que o conhecimento disponível não seja suficiente para identificar as possíveis explicações para elementos distintos.

O volume de dados geralmente impede uma análise detalhada de cada caso, de forma a examinar erros, como é usualmente desenvolvido em amostras pequenas. O processo deve ser ágil e robusto, garantindo que os dados não sejam removidos apenas por serem diferentes dos outros. A identificação de *outliers* também pode utilizar as abordagens de filtro ou de envoltório. Porém a utilização do mecanismo de filtro é difícil, pela falta de limites iniciais para a seleção. A abordagem envoltório é preferida, utilizando as mesmas técnicas escolhidas para a etapa de modelagem.

Como o ajustamento dos modelos pode ser afetado pela presença de dados com erros, a investigação deve ser desenvolvida em diversos estágios, excluindo progressivamente casos com resíduos que atinjam um limite especificado (aumentando o ajustamento do modelo, o desvio-padrão é reduzido, atingindo casos não identificados previamente como *outliers*) (Hair *et al.*, 1998).

Os dados podem ser examinados de forma univariada, bivariada ou multivariada. No primeiro caso, o analista examina as distribuições, geralmente padronizando as variáveis. A análise bivariada usa basicamente gráficos de dispersão. Casos marcadamente fora do grupo podem ser *outliers*. Por fim, a análise multivariada utiliza técnicas de modelagem, tais como redes neurais e análise de regressão, em uma abordagem envoltório, por exemplo (Hair *et al.*, 1998).

A análise dos erros cometidos pelo modelo é essencial para identificar *outliers*. Na análise univariada, os resíduos são examinados através de algumas medidas convencionais, tais como as seguintes. Os resíduos padronizados são o resultado da divisão dos erros pelo seu desvio-padrão ($\varepsilon_p=\varepsilon/\sigma$). Com 50 casos ou mais, aproximam-se bastante da curva *t* de Student. Utiliza-se o limite de $\pm 2\sigma$, que corresponde aproximadamente ao valor de *t* para $\alpha=0,05$. Em amostras maiores, com maior variabilidade, podem ser adotados limites de $\pm 3\sigma$, ou $\pm 4\sigma$. Esta é a medida mais utilizada com análise de regressão (Neter *et al.*, 1990).

Os resíduos deletados são calculados por modelos que envolvem todos os casos da amostra, menos um. O erro de cada caso é calculado com um modelo estimado sem a sua participação. Já os resíduos *studentizados* são calculados dividindo o erro por um desvio-padrão especial, calculado sem aquele caso. A justificativa é que, se a observação é extremamente influente, poderá ampliar o desvio-padrão geral, mascarando seu próprio efeito. Para grandes bases de dados, estes dois tipos de resíduos não são indicados porque o tempo de processamento cresce

muito e porque a influência de um caso isolado sobre o modelo é potencialmente menor (Hayter, 1996; Moore e McCabe, 1998).

A distância de Mahalanobis (D2) indica a distância de cada observação do centro médio das observações (guarda uma certa semelhança com a clusterização). A razão entre a medida D2 e os graus de liberdade é distribuída aproximadamente como uma curva *t*, e pode-se adotar o limite de 0,001 como teste de significância. Há uma dificuldade em definir limites para a medida D2, embora seja claro que casos como D2 muito maior do que os demais são suspeitos. Por fim, outra medida é a DDFIT, que mede o grau pelo qual o valor ajustado (estimado) muda quando seu caso é excluído da modelagem (modelo com e sem ele). Geralmente adota-se a medida padronizada (SDDFIT), com o limite de $2*[(k+1)/(n-k-1)]^{0,5}$, onde k é o número de variáveis independentes e n é o número de dados (Hair *et al.*, 1998).

A decisão de reter ou eliminar o caso deve se dar após a análise detalhada dos dados. Somente devem ser removidos casos verdadeiramente aberrantes, segundo Hair *et al.* (1998), pois a remoção de *outliers* pode diminuir a generalidade dos modelos. A pesquisa por *outliers* pode indicar a falta de um atributo, gerando a demanda por enriquecimento da base. Neste caso, após a coleta das informações correspondentes e inclusão na base, é necessário realizar novamente o processo de exame dos *outliers*, para investigar o efeito desta nova variável na explicação dos casos (recuperando casos excluídos anteriormente). Em função desta forma iterativa de análise, não é conveniente excluir os casos suspeitos definitivamente, apenas identificando e removendo os mesmos da amostra para o modelo corrente, mas não da base de dados.

3.4.1.2.6 Análise dos pressupostos ou requisitos da ferramenta

Por fim, antes da aplicação das técnicas de modelagem, devem ser verificados as condições exigidas, tais como os pressupostos da análise de regressão, ou preparações específicas, como a normalização de dados para as redes neurais ou a discretização para árvores de decisão. Nesta etapa também podem ser separados sub-conjuntos para treinamento e teste (Haykin, 2001; Pyle, 1999).

3.4.2 Mineração dos dados

A mineração de dados (MD) é a fase de desenvolvimento dos modelos ou identificação dos padrões, e envolve a seleção das técnicas a serem aplicadas, seleção dos dados específicos a serem utilizados na construção dos modelos ou busca de padrões (amostras ou segmentos de maior interesse) e preparação destes dados para a ferramenta escolhida (Berry e Linoff, 1997; Weiss e Indurkhya, 1998).

Para Fayyad *et al.* (1996a), as duas metas principais da mineração de dados são predição e descrição, na prática. Segundo estes autores, no contexto do DCBD, a descrição tende a ser mais importante. Berry e Linoff (1997) indicam como metas o teste de hipóteses (verificação de conceitos prévios) e o descobrimento de conhecimento (os dados "falam por si mesmos", segundo eles). Já Weiss e Indurkhya (1998) dizem que há duas categorias de problemas: predição e descobrimento de conhecimento, sendo a primeira a meta mais diretamente ligada com a MD. Por fim, Kohavi (2000) indica como metas a busca de *insights* (identificação de padrões e tendências compreensíveis) e a predição (proposição de um modelo para). Pode-se sintetizar as metas para a MD como sendo descobrimento de conhecimento de uma forma mais ampla (buscar informação, aumentando o entendimento através do descobrimento de novos padrões ou relacionamentos nos dados) ou estimação (geração de modelos para predição), as quais estão vinculadas, basicamente, às categorias de aprendizagem não supervisionada e aprendizagem supervisionada.

Estas metas são atingidas através do desenvolvimento de uma ou mais ações ou tarefas de mineração. Segundo Berry e Linoff (2000), as tarefas são classificação, estimação, predição, agrupamento por afinidade, clusterização e descrição. Já Weiss e Indurkhya (1998) afirmam que os problemas de predição podem ser divididos em questões de classificação, regressão e séries de tempo (como caso especial de regressão ou de classificação), enquanto que o descobrimento de conhecimento lida com detecção de desvios, segmentação de bancos de dados, *clustering*, associação de regras, análise de ligações, sumarização, visualização e mineração de textos. Segundo Fayyad *et al.* (1996a), as principais tarefas são as seguintes: classificação, agrupamento, regressão, sumarização, modelagem de dependências e detecção de variações ou desvios. Já para Deogun *et al.* (1997), as tarefas são classificação, agrupamento, caracterização e modelagem de dependências nos dados. Diante da variedade de tarefas apontadas, uma forma de classificar as tarefas (e as técnicas) em MD é pela forma de

aprendizagem. Agrupando estas definições, pode-se resumir as metas e tarefas em dois grupos, basicamente:

a) descobrimento de conhecimento (aprendizagem não supervisionada) – trata de uma análise mais exploratória. Há conhecimento geral, mas a meta principal é o descobrimento de elementos específicos. As tarefas são:

- agrupamento (clusterização): busca-se identificar um conjunto finito de grupos homogêneos (*clusters*), ou seja, os dados devem ser divididos em categorias por um algoritmo, sendo que estas categorias podem ser mutuamente exclusivas, ou consistirem de uma representação mais rica, com categorias hierárquicas ou sobrepostas;
- sumarização e visualização: concentram-se na generalização dos dados, envolvendo métodos para encontrar uma descrição compacta para um subconjunto de dados, tais como médias e desvios-padrão, ou técnicas de visualização multivariadas; são freqüentemente empregadas na exploração interativa ou geração automática de relatórios;
- modelagem de dependências: busca-se um modelo que descreva os vínculos significativos entre os dados, incluindo regras de associação, árvores de decisão, etc.;

b) estimação (geração de modelos para predição, aprendizagem supervisionada) – há conhecimento inicial, por exemplo de casos anteriores, mas não há um modelo explícito que possa ser empregado no teste de hipóteses ou na previsão de valores). Esta forma pode ser resumida como "exemplos do passado, com respostas conhecidas, são generalizados para analisar casos futuros" (Weiss e Indurkhya, 1998, p.8), ou seja, busca-se determinar modelos a partir dos dados. Pode ser dividida em duas tarefas:

- classificação: é o aprendizado de uma função que relaciona cada item dos dados a uma entre diversas classes pré-definidas (variável-alvo discreta);
- regressão: é o aprendizado de uma função que relaciona cada um dos dados a uma variável de predição, geralmente contínua (intervalo real).

Algumas observações podem ser feitas sobre estas tarefas de mineração. Embora similares, a diferença entre agrupamento e classificação é que no primeiro caso as categorias não são conhecidas (e são o resultado da clusterização) e no segundo existe um grupo de casos corretamente classificado, como base para a aprendizagem (supervisão). Assim, a

segmentação de um banco de dados pode ser um caso de classificação ou de agrupamento, conforme o critério de divisão seja previamente conhecido ou não, respectivamente.

A análise de seqüências e séries temporais ou de dados distribuídos espacialmente tem particularidades quanto ao ordenamento dos dados (no tempo ou no espaço), mas pode ser entendida como um caso de classificação ou de regressão, conforme o tipo de variável-alvo (discreta ou contínua).

A mineração de textos, que é a busca de expressões significativas (*keywords*) em textos livres, também pode ser enquadrada nos dois grupos, conforme haja a definição prévia de um dicionário de palavras-chave (é um caso de classificação) ou ele seja gerado durante o processo de análise (pode ser *clustering* ou sumarização de dados).

3.5 CONSIDERAÇÕES FINAIS

A preparação dos dados deve aprimorar os modelos gerados, em função da maior qualidade dos dados (redução do ruído e seleção de dados relevantes). Neste sentido, a aplicação de DCBD pode aprimorar os resultados das técnicas atuais. Entretanto, um dos empecilhos que podem ser apontados é a falta de bases de dados, e a dificuldade de obter informações sobre alguns aspectos das transações, tal como a motivação dos agentes. Para compensar esta dificuldade, é necessário que haja uma coleta sistemática de dados, como ocorre em outros países.

4 TÉCNICAS EMPREGADAS EM DESCOBRIMENTO DE CONHECIMENTO E MODELAGEM

4.1 CONSIDERAÇÕES INICIAIS

A revisão das características das técnicas disponíveis foi desenvolvida tendo como base as indicações gerais da literatura. Foram identificadas as características básicas das técnicas utilizadas nas áreas de descobrimento de conhecimento e inteligência artificial, tendo em vista o potencial sugerido por diversas aplicações relatadas, especialmente nas áreas de avaliação de imóveis e Engenharia Civil.

Foram selecionadas algumas técnicas na área de inteligência artificial, tais como algoritmos genéticos, árvores de decisão regras clássicas e difusas, árvores de decisão, redes neurais e raciocínio baseado em casos, e outras relacionadas com a estatística, tais como análise de agrupamento e análise de componentes principais, as quais aparentemente podem contribuir para a formulação de um sistema destinado à avaliação de imóveis. Algumas destas técnicas tiveram desenvolvimento mais recente, e apresentam um grau de complexidade maior, então foram discutidas em maior detalhe, neste capítulo, enquanto que outras tiveram apenas as características básicas apresentadas.

4.2 ANÁLISE DE AGRUPAMENTO (*CLUSTERING*)

A tarefa principal da análise de agrupamento é organizar os dados em um pequeno número de grupos (*clusters*), de forma que os elementos similares estejam alocados no mesmo grupo e os padrões muito distintos estejam em grupos diferentes (Berry e Linoff, 2000; Kaski, 1997). Embora existam diversos algoritmos para realizar esta tarefa, há dois aspectos fundamentais, da natureza dos métodos de agrupamento (Cios *et al.*, 1998):

a) não existe efeito explícito de supervisão: os padrões são organizados de acordo com um critério de agrupamento, mas os grupos não conhecidos no início;

b) a noção de similaridade (ou distância) entre dois padrões é essencial: a função de distância é a quantificação da similaridade; quanto menor a distância, maior a similaridade; se a distância é zero, os padrões são iguais.

Na análise de agrupamento, não existe pré-classificação dos dados, nem distinção entre variáveis dependentes ou independentes. Trata-se de um tipo de aprendizagem não supervisionada, no qual o método lida diretamente com os dados, visando determinar a estrutura dos padrões (Berry e Linoff, 2000; Cios *et al.*, 1998).

Em alguns casos de *data mining*, uma das dificuldades a ser enfrentada é o excesso de informações. Torna-se necessário utilizar técnicas que permitam particionar ou reduzir os dados, facilitando o entendimento dos padrões. Freqüentemente não há informação inicial para apoiar as tarefas de segmentação ou classificação dos dados (Berry e Linoff, 2000).

Para Kaski (1997), um dos principais problemas com esta técnica é a interpretação dos *clusters* obtidos, que pode ser difícil. Assim, se a meta não é apenas compactar o arquivo de dados, mas também realizar inferências sobre a estrutura de *clusters* resultante, nem sempre podem ser obtidos bons resultados.

Por este motivo, a análise de agrupamento raramente é utilizada isoladamente. Detectados os *clusters*, outros métodos são aplicados para descobrir o que significam. Esta técnica tem aplicação em uma larga variedade de campos, incluindo descobrimento de conhecimento, análise estatística, compressão de dados, quantização vetorial, reconhecimento e classificação de padrões. Uma das aplicações comuns é a construção automatizada de categorias ou taxonomias (Berry e Linoff, 1997; Kanungo *et al.*, 1999; Kaski, 1997).

4.2.1 Classes de algoritmos

Existem duas classes básicas de algoritmos: agrupamento hierárquico ou particional. O agrupamento hierárquico busca reunir sucessivamente grupos menores, formando grupos maiores, ou dividir grupos grandes em outros de maior similaridade interna. Os métodos diferem pela regra adotada para decidir quais grupos devem ser reunidos ou divididos. O resultado do algoritmo é um gráfico tipo árvore, chamado de "dendograma", que mostra como os grupos são inter-relacionados (Cios *et al.*, 1998; Kaski, 1997).

Já o agrupamento particional busca dividir o conjunto de dados em um conjunto de *clusters* distintos entre si, maximizando as dissimilaridades dos diferentes *clusters* (Kaski, 1997; Ng e Han, 1994; Bradley *et al.* 1998). As técnicas que seguem o agrupamento particional, incluindo K-means e diversas outras, geralmente são baseadas na otimização de uma função de custo, que envolve a minimização do erro quadrático, por exemplo, e são de natureza combinatorial. Em função do tamanho dos arquivos, o tempo de processamento é muito grande. Por isto, técnicas de otimização, tais como *simulated annealing* e algoritmos genéticos, são empregadas para acelerar o processamento (Berry e Linoff, 2000; Cios *et al.*, 1998; Kaski, 1997).

4.2.2 Algoritmos para agrupamento

Em função de ser o método de implementação mais comum em *softwares* dentre todos os métodos, K-means é uma referência, contra a qual os outros algoritmos são comparados. Muitos algoritmos para *clustering* tem sido propostos, especialmente para *clustering* espacial, alguns dos quais são apresentados a seguir. Por ser uma das técnicas mais importantes em mineração de dados, há muito esforço de aperfeiçoamento (Ester *et al.*, 1998; Xu *et al.*, 1997; Wang *et al.*, 1997).

4.2.2.1 Algoritmo K-means

K-means é um algoritmo de clusterização particional, proposto por MacQueen em 1967 (Kaski, 1997). É o algoritmo mais conhecido e aplicado para clusterização, embora existam

muitas variantes. Este algoritmo requer que os dados sejam compostos de variáveis numéricas, pois uma parte do processo é baseada no cálculo das médias (Bradley *et al.*, 1998).

Em termos mais precisos, a clusterização por K-means pode ser descrita como: "dado um conjunto de n pontos no espaço real d-dimensional R^d e um número inteiro k, definir os k conjuntos de pontos em R^d que minimizem a distância média quadrada de cada ponto ao centróide do conjunto mais próximo" (Kanungo *et al.*, 1999, p.1). O processo de cálculo consiste basicamente das seguintes etapas:

a) selecionar k pontos;

b) determinar as coordenadas destes pontos como sendo os centróides dos *clusters*;

c) calcular a distância do próximo ponto aos k centróides (geralmente empregando a distância euclidiana);

d) incorporar o ponto ao *cluster* mais próximo;

e) recalcular o centróide deste *cluster*;

f) passar ao próximo ponto – se terminarem os pontos, recomeçar do primeiro ponto, revendo seu posicionamento;

g) encerrar o processo se não houver possibilidade dos dados mudarem de *cluster* ou retornar à etapa (c).

Embora simples e razoavelmente eficiente, o algoritmo K-means tem algumas desvantagens. Um dos problemas apontados é o tempo de processamento, em função da aplicação em espaços de muitas dimensões (muitas variáveis) e da necessidade de muitas iterações até a condição de final (Kanungo *et al.*, 1999). Outro problema é a escolha das condições iniciais. O número de *clusters* (k) e a inicialização dos centróides (escolha dos primeiros k pontos) pode influir decisivamente nos resultados. A própria limitação a um determinado número de *clusters* pode não ser adequada, quando a análise é exploratória. Por fim, o algoritmo freqüentemente encerra o processo com ótimos locais, desprezando valores mais globais. Não obstante, ainda é o algoritmo mais utilizado, na prática, em função de ser a variante implementada nos pacotes estatísticos comerciais (Kaski, 1997).

4.2.2.2 Algoritmos para agrupamento espacial

Há diferenças significativas na análise de dados espaciais, provenientes de imagens ou geo-referenciados. Por isto, há necessidade da proposição de novos métodos. Vários pesquisadores têm proposto algoritmos para substituir K-Means, com desempenho superior, em termos de tempo de processamento e precisão de agrupamento:

a) K-medoids - É a base de vários algoritmos. A principal diferença é que em K-medoids os centróides são dados da base, enquanto que em K-means são médias dos pontos, nem sempre coincidentes com pontos reais, existindo, por isto, a sensibilidade a *outliers*. Outro ponto importante é que os resultados deste algoritmo não dependem da ordem em que os dados são apreciados (Ng e Han, 1994).

b) K-medianas - O objetivo do algoritmo é minimizar a soma das distâncias dos pontos aos centróides dos k *clusters* (Kanungo *et al.*, 1999).

c) CLARANS - Ng e Han (1994) desenvolveram um algoritmo para clusterização espacial denominado CLARANS. Usaram *clusters* determinados através de K-medoids e seleção aleatória de amostras para determinação dos *clusters*, ao invés de trabalhar com toda a base de dados, visando a aumentar a eficiência para grandes bancos de dados.

d) DBSCAN – A idéia básica é que, para cada ponto de um *cluster*, a vizinhança a um certo raio deve conter ao menos um número mínimo de pontos, ou seja, a densidade da vizinhança precisa atingir um limite mínimo especificado (Xu *et al.*, 1997a).

e) COD-CLARANS - Tung *et al.* (2001) propuseram uma formulação para a clusterização espacial na presença de obstáculos, tais como rios ou avenidas que criam barreiras físicas. Geralmente, não tem sentido unir em um mesmo *cluster* pontos dos dois lados do obstáculo, embora possam ser próximos. Estes autores apresentaram o algoritmo COD-CLARANS, que é uma evolução de CLARANS. Segundo Tung *et al.* (2001), os resultados empíricos indicam que é mais adequado para este tipo de situação.

4.3 REGRAS DE ASSOCIAÇÃO

As regras de associação são outra uma importante ferramenta de análise para o descobrimento de conhecimento. Uma regra é um tipo de regularidade em um banco de dados, que pode indicar um padrão, embora nem sempre existam relações de causalidade. As regras são úteis para a análise exploratória (Berry e Linoff, 2000; Westphal e Blaxton, 1998).

Agrawal *et al.* (1993) abordaram originalmente esta técnica, no contexto de descobrimento de conhecimento. As regras assumem um formato do tipo $X \Rightarrow Y$, onde X e Y são conjuntos de atributos. O significado intuitivo de uma regra deste tipo é de que nos casos da base em que os atributos do conjunto X têm valor "verdade" (ocorrem), os atributos do conjunto Y também têm valor "verdade" (Klemettinen *et al.*, 1994; Toivonen, 1996).

Na presença de grandes bases de dados, com muitos casos (milhares) e muitos atributos (dezenas ou centenas), o maior problema é o possível excesso de regras de associação. Klemettinen *et al.* (1994) afirmam que, paradoxalmente, a mineração de dados pode gerar tantas regras que surge um novo problema, de gerenciamento do novo conhecimento e é necessário examinar a relevância das regras. Eles chamam isto de mineração de segunda ordem, e sugerem técnicas de poda ou limitações nos relacionamentos. Toivonen (1996), por exemplo, relatam a existência de mais de 2 mil regras em uma base de dados de matrículas de estudantes universitários, e de mais de 30 mil regras em uma base de casos de alarme em redes de telecomunicações.

Porém, é preciso evitar a poda de regras importantes. Por outro lado, mesmo com elevada confiança, algumas regras podem não interessar, principalmente porque correspondem a conhecimento anterior, previamente dominado, porque se referem a atributos, ou combinações de atributos, não interessantes, ou porque são redundantes (Klemettinen *et al.*, 1994). Outro problema deste tipo de regra é que são baseadas na lógica clássica, com baixa tolerância à imprecisão e ao raciocínio imperfeito. Se estes elementos estiverem presentes, é mais conveniente utilizar regras difusas (Cordón *et al.*, 2001).

Existem dois tipos especiais de regras, conhecidos como análise de ligações (*link analysis*) e análise da cesta de consumo (*market basket analysis*), também muito utilizadas na busca por conhecimento, especialmente na área comercial.

4.3.1 Análise de Ligações (*Link Analysis*)

Para Berry e Linoff (2000), a análise de ligações (AL) persegue relacionamentos entre casos, desenvolvendo modelos baseados nos padrões dos relacionamentos. Esta é uma aplicação da teoria dos grafos à mineração de dados. É utilizada principalmente para investigar relacionamentos de consumidores. Por exemplo, na área de telecomunicações, cada chamada pode significar informação com potencialidades de consumo. Outra área de uso intenso é a investigação criminal. Para estes autores, o mundo dos negócios é um mundo de relacionamentos, conectando pessoas, lugares e coisas através de caminhões e aviões, telecomunicações, contatos pessoais, etc. A análise de ligações não é aplicável a quaisquer tipos de dados ou problemas. Algumas áreas em que têm sido obtidos bons resultados são:

a) análise de padrões de chamadas telefônicas: cada ligação representa um relacionamento entre dois pontos;

b) entendimento de padrões médicos;

c) combinação de fontes de informação na investigação criminal (FBI).

Existem poucas ferramentas que explicitamente desenvolvem AL e, na maioria, são especializados na área criminal. Geralmente são focadas na visualização dos *links*, auxiliando os analistas para que eles mesmos encontrem padrões nos dados. Todavia, mesmo consultas SQL e bancos de dados relacionais podem ser utilizados para AL, embora sofram com a performance, pois a análise de ligações geralmente exige muito processamento. Para pequenas quantidades de dados, a tecnologia orientada a objetos pode proporcionar uma alternativa eficiente (Berry e Linoff, 1997).

Outro problema a ser enfrentado é o reconhecimento de quando os *links* existem realmente entre os itens. No caso das chamadas telefônicas e nos transportes, os links são óbvios, enquanto que nas aplicações criminais os relacionamentos são encontrados com o auxílio de outras técnicas (Berry e Linoff, 1997).

Algumas das vantagens da AL são a explicitação dos relacionamentos, a facilidade para visualização dos dados e a criação de atributos derivados, tal como "área de influência". As desvantagens são a dificuldade de aplicação em muitos tipos de dados, a existência de poucas

ferramentas (em geral muito caras) e as dificuldades na implantação em bancos de dados relacionais (Berry e Linoff, 1997).

4.3.2 Análise da Cesta de Consumo (*Market Basket Analysis*)

Segundo Ganti *et al.* (1999), uma cesta de consumo é uma coleção de itens comprados por um consumidor em uma transação individual. Uma das análises comuns sobre cestas de consumo é buscar conjuntos de itens ("*itemsets*", na linguagem de Agrawal *et al.*, 1993) que aparecem juntos em muitas transações.

A análise da cesta de consumo indica a probabilidade de certos produtos serem adquiridos juntos. É uma forma de análise similar à clusterização, usada para encontrar grupos de itens que tendem a ocorrer juntos em uma transação (cesta de consumo). Os modelos construídos apresentam a probabilidade dos diferentes produtos serem comparados em conjunto. É uma técnica muito empregada no comércio varejista, no qual a informação da cesta de consumo é o principal dado disponível para minerar os padrões de consumo (Berry e Linoff, 1997).

4.4 ÁRVORES DE DECISÃO

Uma árvore de decisão é um método de classificação que recursivamente particiona o conjunto de dados de treinamento até que cada partição consista inteiramente ou ao menos prioritariamente de exemplos de uma classe. A árvore é composta por folhas e nós de decisão. As folhas estão nos extremos dos ramos, indicando as classes. Cada nó contém uma divisão, a qual consiste em um teste aplicado a um ou mais atributos, gerando um ramo (e uma sub-árvore) para cada resultado possível do teste (Quinlan, 1993; Shafer *et al.*, 1996).

As árvores dividem os casos do conjunto de treinamento em subconjuntos distintos, cada um descrito por uma regra simples sobre um ou dois atributos. Uma das vantagens principais das árvores de decisão é que o modelo é muito explicável, o que facilita a avaliação dos resultados por parte dos usuários, identificando as características-chave do processo. É útil também na presença de novos dados de qualidade incerta: resultados espúrios ficam óbvios com regras explícitas. As regras podem estar como proposições lógicas, em uma linguagem tal como

SQL, podendo então ser aplicadas diretamente aos novos casos (Berry e Linoff, 1997).

Para Quinlan (1993), alguns modelos de classificação são obtidos através do conhecimento de especialistas, como nos sistemas especialistas. Outro caminho é a construção indutiva dos modelos, generalizando exemplos específicos conhecidos. Não são todas as tarefas de classificação que se adaptam a esta abordagem indutiva, existindo um conjunto de requisitos que devem ser atendidos (Berry e Linoff, 2000; Quinlan, 1993):

a) existe uma descrição atributo-valor: os dados a serem analisados devem ser expressas em termos de uma coleção fixa de propriedades ou atributos: o valor destes atributos pode variar, mas a estrutura (o tipo de variável) é fixa;

b) as classes são pré-definidas: trata-se de aprendizagem supervisionada;

c) as classes são discretas: um caso pertence ou não pertence a uma classe;

d) existem dados suficientes: em função dos testes estatísticos, é preciso dispor de dados em quantidade suficiente (devem existir muito mais casos do que classes);

e) a variável dependente tem valores discretos, embora alguns algoritmos aceitem variáveis contínuas, as quais são particionadas recursivamente.

4.4.1 Algoritmos para árvores de decisão

Segundo Quinlan (1993), as idéias iniciais desta técnica foram enunciadas na década de 50, por Hoveland e Hunt, com os sistemas de aprendizagem de conceitos. Para Quinlan, o esquema básico para construir uma árvore de decisão a partir de um conjunto de casos de treinamento T é elegantemente simples. Dadas as classes $\{c_1, c_2, ..., c_k\}$, existem três possibilidades:

a) T contém um ou mais casos, todos de uma única classe c_j: a árvore de decisão para T é uma folha identificando a classe c_j;

b) T não contém casos: a árvore de decisão é novamente uma folha, mas as classes

a serem associadas com ela devem ser determinadas por informações externas a T (conhecimento do domínio); por exemplo, C4.5 usa a classe mais freqüente nos ramos-pais deste nó;

c) T contém casos que incluem uma mistura de classes: nesta situação, deve-se refinar T em subconjuntos de casos que são, ou parecem ser, coleções de casos de uma única classe. Um teste é escolhido, baseado em um único atributo, que tenha um ou mais resultados mutuamente exclusivos $\{O_1, O_2, ..., O_n\}$. T é particionado em subconjuntos $\{T_1, T_2, ..., T_n\}$, onde T_i contém todos os casos em T que tem o resultado O_i do teste escolhido. A árvore para T consiste de um nó de decisão identificando o teste, e um ramo para cada O_i.

O algoritmo é aplicado recursivamente ao subconjunto de treinamento. Qualquer teste que divida T de uma forma não-trivial, desde que ao menos dois dos subconjuntos $\{T_i\}$ sejam não vazios, deve resultar em uma partição de subconjuntos de uma única classe, mesmo se todos ou a maioria tenha apenas um caso. Todavia, o processo de construção da árvore não busca meramente encontrar qualquer partição, mas construir uma árvore que revele a estrutura do domínio e tenha poder preditivo. Para isto, é necessário um número significativo de casos a cada folha ou que a partição tenha tão poucos blocos quanto possível. Não é viável examinar todas as árvores possíveis, em função do grande número de possibilidades (Quinlan, 1993).

O algoritmo constrói a árvore identificando repetidamente qual atributo deve ser utilizado em cada subdivisão da árvore. Para esta definição, usa uma medida da capacidade de classificação de cada atributo disponível. O atributo com maior capacidade é colocado na raiz da árvore (Quinlan, 1993).

Os algoritmos para árvores de decisão consistem basicamente de duas etapas: construção da árvore e poda dos excessos. Na construção da árvore, é bastante importante o critério de divisão, o qual tem importância na qualidade da árvore. A maioria dos algoritmos é do tipo "sem retorno" e exige um grande número de casos para o treinamento. Alguns dos algoritmos mais conhecidos são ID3, C4.5, CART e CHAID. Os três primeiros tendem a gerar árvores com muitos ramos, exigindo a etapa de poda. O algoritmo CHAID geralmente cria árvores mais compactas, porém é adequado apenas para variáveis categóricas (Berry e Linoff, 2000; Quinlan, 1983). No caso de bases muito grandes, que não cabem na memória, é necessário adotar uma estratégia de abordagem dos dados. Entre vários critérios disponíveis, Sprint e

RainForest estão entre os mais freqüentemente citados (Ganti *et al.*, 1999; Gehrke *et al.*, 1998; Shafer *et al.*, 1996).

4.4.2 Utilização das árvores de decisão

Westphal e Blaxton (1998) afirmam que as árvores de decisão são utilizadas para descobrir regras e relacionamentos através da divisão e subdivisão sistemática das informações contidas na base de dados. As árvores de decisão são usadas particularmente na classificação de dados, e têm algumas vantagens (Berry e Linoff, 2000; Ganti *et al.*, 1999):

a) têm uma representação intuitiva, que faz o modelo de classificação resultante fácil de entender;

b) a construção da árvore de decisão não requer parâmetros a serem definidos pelo analista;

c) a precisão preditiva das árvores de decisão é igual ou superior a de outros modelos de predição;

d) existem algoritmos razoavelmente rápidos e escaláveis, facilitando o trabalho com grandes bases de dados;

e) as árvores de decisão fornecem uma indicação clara de quais são os atributos mais importantes.

Além disto, as árvores podem ser facilmente convertidas para uma estrutura de regras, com vantagens em relação ao entendimento de árvores de grandes dimensões (Weiss e Indurkhya, 1998). São utilizadas em diversos tipos de tarefas, tais como diagnósticos médicos e análise de crédito. Não foram encontradas aplicações na área do mercado imobiliário e, segundo Berry e Linoff (2000) e Westphal e Blaxton (1998), as árvores de decisão não são apropriadas para tarefas de predição de uma variável contínua.

4.5 ANÁLISE FATORIAL

A análise fatorial tem como principal objetivo obter uma interpretação mais simples de um conjunto de dados. É também uma forma de obter a redução do número de variáveis, e torna-se especialmente útil quando há suspeita de multicolinearidade. As variáveis são convertidas em fatores, os quais são combinações lineares destas variáveis, ou seja, os fatores são médias ponderadas das variáveis originais, com o formato apresentado na Equação 4 (Hair *et al.*, 1998; Harmann, 1976; Nie *et al.*, 1975):

$$F_j = a_1X_i + a_2X_2 + ... + a_kX_k \qquad \text{(Equação 4)}$$

Onde F_j é um fator, a_i são os pesos (também chamados de cargas) e X_i são as variáveis. Existem diversos métodos para obter os fatores, e a análise de componentes principais é um dos mais empregados (Frank, 1971; Harmann, 1976; Nie *et al.*, 1975).

As variáveis que apresentam maior correlação tendem a apresentar pesos elevados no mesmo fator. Assim, através do exame da matriz de pesos, pode-se identificar quais as variáveis que têm potencialmente a mesma informação, excluindo-se então uma ou mais variáveis da análise posterior. Outra alternativa é o emprego dos próprios fatores como variáveis nos modelos desenvolvidos, com as vantagens de redução no número de variáveis ou de redução da colinearidade. No caso da substituição, um dos efeitos perceptíveis na análise de regressão é a diminuição do coeficiente de determinação, provocada pela redução da colinearidade. Se as relações entre as variáveis não são fortes, ou se expressam relações que espelham a realidade, pode ser interessante manter as variáveis originais, em função da interpretação dos modelos (Gujarati, 2000; Hair *et al.*, 1998; Judge *et al.*, 1985).

Há alguns aspectos a serem considerados na aplicação deste método, tais como a quantidade de fatores a serem gerados e a aplicação ou não de rotação nos fatores (e também a escolha do método de rotação). O número de fatores não é conhecido inicialmente, e deve ser ajustado por tentativas. A rotação geralmente facilita a interpretação dos fatores, concentra os pesos de cada variável em poucos fatores, diminuindo nos outros. A aplicação de um método de rotação ortogonal, tal como Varimax, permite manter a independência (colinearidade zero) entre os fatores gerados (Hair *et al.*, 1998; Harmann, 1976; Nie *et al.*, 1975).

González (1993), Kain e Quigley (1970), Morton (1977) e Wilkinson e Archer (1973)

empregaram esta técnica na área do mercado imobiliário, através do método de componentes principais e, com exceção de Kain e Quigley (1970), utilizaram rotação Varimax. Segundo Kain e Quigley (1970), há motivos para acreditar que o mercado avalia a qualidade de habitação através de índices agregados, portanto os modelos também devem utilizar agregações das variáveis. Estes autores utilizaram análise fatorial para condensar 39 variáveis qualitativas em apenas 5 fatores. Os fatores foram então incluídos nas equações de regressão em lugar das variáveis. Wilkinson e Archer (1973) também desenvolveram modelos substituindo as variáveis originais pelos fatores. Em outra abordagem, González (1993) e Morton (1977) utilizaram a análise fatorial para identificar grupos de variáveis colineares. Examinando cada fator, era escolhida apenas uma das variáveis entre aquelas com os maiores pesos. Em seguida, as variáveis selecionadas foram incluídas em modelos de regressão.

4.6 RACIOCÍNIO BASEADO EM CASOS

O Raciocínio Baseado em Casos é uma técnica de solução de problemas baseada em conhecimento, fundamentada na reutilização de experiências prévias (sintetizadas em "casos"). Ao contrário de outras técnicas baseadas em conhecimento, que resolvem problemas a partir de uma base de conhecimentos gerais, o RBC preocupa-se com experiências específicas, representadas pelos casos contidos na base de dados. As premissas do RBC são de que (a) problemas similares têm soluções similares, e (b) a reutilização de soluções de casos anteriores é mais adequada do que uma solução baseada na generalização de regras (Althoff e Bartsch-Spörl, 1996; Kolodner, 1993; Leake, 1996).

Segundo Watson (1997), as vantagens do RBC incluem o fato de que o método não depende de um modelo explícito para a solução do problema, a flexibilidade de trabalhar com grandes quantidades de dados e a possibilidade de aprender com novos casos, sendo fácil manter atualizado o sistema.

A técnica de Raciocínio Baseado em Casos é ligada a uma área relativamente recente da Inteligência Artificial. Podem ser encontrados alguns elementos relacionados com RBC desde as pesquisas filosóficas de Wittgenstein, em 1953, mas grandes avanços foram obtidos em 1977, com estudos de Schank e Abelson, e em 1982, com trabalhos de Schank sobre memória

dinâmica e outros modelos de memória. O primeiro sistema, denominado Cyrus, foi desenvolvido por Kolodner, logo após, como uma implementação do modelo de Roger Schank (Aamodt e Plaza, 1994; Kolodner, 1993; Marir e Watson, 1994; Watson, 1997).

Segundo Aamodt e Plaza (1994), diversas pesquisas em psicologia cognitiva demonstraram evidências da utilização de regras de situações anteriores na resolução de problemas, mesmo no caso de especialistas. Para Althoff e Bartsch-Spörl (1996), o Raciocínio Baseado em Casos pode ser considerado uma abordagem cognitiva de modelagem da forma humana de solução de problemas, em domínios nos quais a experiência é importante. Eles apontam os seguintes elementos-chave para o sucesso de um sistema de RBC:

a) deve ser obtida uma quantidade representativa de casos na área de aplicação;

b) os casos devem estar representados em um formato acessível pelo *software* utilizado;

c) os índices de busca devem ser eficientes;

d) a rotina que avalia a similaridade dos casos tem de ser precisa e robusta;

e) o sistema RBC tem de ser apto para realizar as adaptações necessárias;

f) o sistema deve ser bem aceito pelos usuários.

A principal parte do conhecimento nos sistemas RBC está no adequado tratamento dos casos conhecidos. A representação dos casos é uma tarefa complexa e importante para o sucesso do sistema. Um caso pode ser entendido como a abstração de uma experiência descrita em termos de seu conteúdo e contexto, podendo assumir diferentes formas de representação. Não existe consenso, na literatura especializada, sobre as informações que devem ser consideradas na descrição de um caso, mas podem ser enunciadas duas regras básicas, quais sejam: a funcionalidade do sistema e a facilidade de aquisição das informações (Watson, 1997). O conhecimento, neste nível de abrangência, pode ser interpretado como sendo um conjunto de métodos que modelam um conhecimento especializado para disponibilizá-lo em um sistema inteligente (Aamodt e Plaza, 1994).

Os casos têm três partes a serem descritas: o problema (contextualizado), a solução aplicada e o resultado final. Podem ser representados de várias formas, tais como simples vetores de

características ou objetos estruturados, com descrições simbólicas ou com informação multimídia (Althoff e Bartsch-Spörl, 1996). Neste sentido, Oliveira *et al.* (1997) e Watson e Oliveira (1998), apresentam aplicações de realidade virtual na representação dos casos. Segundo esses autores, a interface com o usuário é importante, e a realidade virtual permite apresentar melhor o caso para os usuários, pois é um meio "ativo", isto é, o usuário pode interagir com o caso.

O Raciocínio Baseado em Casos é um processo cíclico, composto basicamente por quatro fases[28]: (a) recuperação de casos similares, (b) reutilização das informações como proposta de solução, (c) revisão da solução e (d) retenção da experiência para utilização futura (Aamodt e Plaza, 1994). Este ciclo está representado na Figura 4:

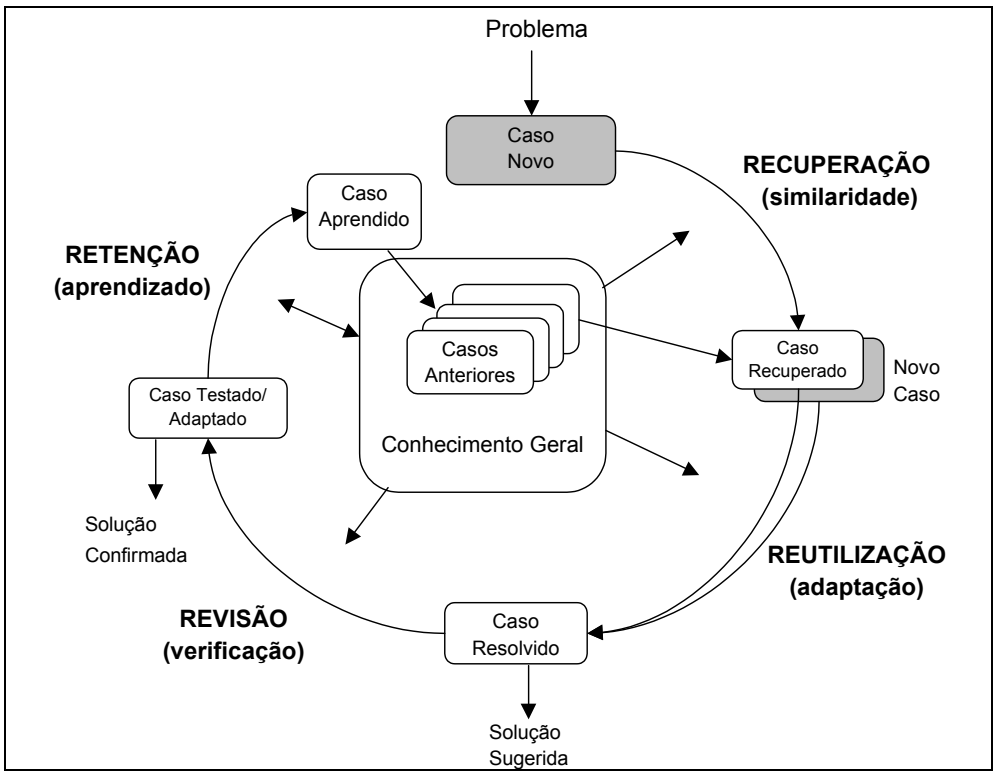


Figura 4: Ciclo do raciocínio baseado em casos (Adaptado de Aamodt e Plaza, 1994, p.45)

A tarefa de recuperação inicia com uma descrição parcial do problema, e termina quando um caso similar é encontrado. As sub-tarefas são "identificar características", "realizar busca

inicial", "procurar" e "selecionar", nesta ordem. A identificação trata de ajustar um conjunto de descritores relevantes, para que a busca possa encontrar casos suficientemente similares, dado um limiar de similaridade (Aamodt e Plaza, 1994).

A indexação facilita a recuperação dos casos de interesse na solução do problema. Os índices devem ser preditivos, indicando os propósitos do caso, sendo abstratos o suficiente para abranger toda a base e, por outro lado, concretos o bastante para serem reconhecidos no futuro (Watson, 1997).

A busca também pode ser realizada de várias formas. A mais simples é a seqüencial, que não é recomendada para sistemas que possuam um grande número de casos, por causa do tempo consumido. Sistemas mais complexos requerem um ou mais índices. Como em geral os casos não coincidem exatamente, deve ser utilizado um mecanismo ou algoritmo flexível de busca que permita esta recuperação. Para cada caso "candidato" encontrado, deve ser determinada uma medida de similaridade em relação ao caso que se pretende resolver, ordenando-se os casos recuperados em ordem decrescente de similaridade (Althoff e Bartsch-Spörl, 1996).

Os algoritmos de busca mais conhecidos são: vizinhança próxima, indução, indução guiada pelo conhecimento e recuperação segundo um modelo, empregados isoladamente ou em conjunto (Kolodner, 1993; Watson, 1997). Em alguns casos, podem ser empregadas redes neurais ou lógica difusa como auxiliares ao sistema RBC, na tarefa de busca de casos similares ou desenvolvimento de índices (Reategui *et al.*, 1995; Sovat e Carvalho, 1999).

No mecanismo de vizinhança próxima, a similaridade é indicada por uma soma ponderada das características. Naturalmente, o problema é determinar o peso das características. O tempo de busca cresce linearmente com o tamanho da base (número total de casos), sendo mais indicado para bases pequenas (Kolodner, 1993; Watson, 1997).

Na indução simples, o algoritmo determina quais características trazem a melhor solução na escolha de casos e gera uma árvore de decisão organizando os casos na memória. É mais adequado para os casos em que se requer um único caso como solução, e quando as características do caso dependem de outras características. A indução guiada pelo conhecimento aplica informações fornecidas pelo usuário para a recuperação, através da

[28] Estas fases são denominadas de "4 RE's" (*retrieve*, *reuse*, *revise, retain*) por Aamodt e Plaza (1994).

seleção prévia das características que são reconhecidas como principais (Watson, 1997).

Por outro lado, a recuperação utilizando um padrão retorna apenas os casos que se ajustam a certos parâmetros (limites). Pode ser utilizada antes dos outros métodos, para selecionar um grupo relevante de casos (Watson e Marir, 1994).

A reutilização de casos está ligada com a adaptação dos casos passados. A adaptação é a fase da escolha entre a solução recuperada diretamente ou o método que construiu esta solução. No primeiro caso, a solução é adaptada através de operadores de transformação, aplicados à solução. A segunda forma é conhecida como utilização derivada, buscando refazer o caminho da solução, sob o novo contexto. A atividade de adaptação pode ser tão simples quanto substituir apenas um componente do caso recuperado ou tão complexa quanto modificar a estrutura geral da solução (Aamodt e Plaza, 1994; Kolodner, 1993; Watson e Marir, 1994).

A dificuldade de adaptação com dados numéricos é um problema corrente na área, mas que pode ser contornado com o uso de outras técnicas, tal como a análise de regressão. Outra alternativa é a lógica difusa, que pode ser empregada para gerar regras de adaptação, conforme sugerido por Sovat e Carvalho (1999). Dubois *et al.* (1997) e Weber-Lee *et al.* (1995) também apresentaram trabalhos com a participação da lógica difusa em sistemas de RBC. No mesmo sentido, Yager (1997) afirma que a lógica difusa deve ser utilizada em ambientes em que soluções numéricas sejam requeridas.

A fase de revisão consiste em avaliar a solução gerada. Se a solução for bem sucedida, o caso pode ser retido como aprendizado. Se não for, a solução deve ser revista, usando conhecimentos específicos ou a intervenção do usuário ou de um especialista. Geralmente é uma etapa externa ao sistema informatizado, pois envolve a aplicação da solução ao mundo real. Encerrando o ciclo, a etapa de retenção tem a finalidade de preparar os casos para uso futuro, atualizando a base de casos e proporcionando a aprendizagem do sistema (Aamodt e Plaza, 1994).

Em síntese, a partir de um problema dado, o sistema RBC deve recuperar um ou mais casos similares relevantes e avaliar como estes casos encaixam-se na nova situação. Se forem detectadas diferenças, o sistema adapta os casos para a nova situação e retorna ao início da rotina de avaliação. O sistema aprende relembrando sucessos e falhas como novos casos (Watson, 1997).

4.6.1 Aplicações de raciocínio baseado em casos em avaliações

O RBC adapta-se a um grande número de aplicações. Althoff e Bartsch-Spörl (1996) afirmam que a maior parte dos problemas comerciais refere-se a tarefas analíticas, compreendendo análise da situação, classificação e obtenção de um diagnóstico ou sugestão de uma solução. São aplicações deste tipo a análise de alternativas de investimentos, de compra de imóveis ou de compra de novos equipamentos para uma empresa. Um outro tipo de problema é o que pertence à categoria denominada por eles de tarefas sintéticas, nas quais é necessário gerar construções ou planos complexos, em aplicações como logística de transportes e planejamento da produção.

Uma das aplicações mais comuns do RBC atualmente é nos serviços de atendimento a clientes. Watson (1997) elencou mais de 130 sistemas de RBC, dentre os quais cerca de 80 eram de *help-desk*. Outro campo de utilização freqüente é a área jurídica, com diversas aplicações, inclusive no Brasil (Bueno *et al.*, 1999; Sycara, 1988; Watson, 1997; Weber-Lee, 1998; Weber-Lee *et al.*, 1997). São comuns as aplicações em projeto, planejamento e diagnósticos (Watson e Abdullah, 1994; Watson, 1997).

Não são comuns aplicações com resultados predominantemente numéricos, principalmente se existe grande volume de dados (Watson, 1997). Entretanto, foram encontrados alguns trabalhos relacionados com o mercado imobiliário (Bonissone *et al.*, 1998; González e Laureano-Ortiz, 1992; McSherry, 1998; O'Roarty *et al.*, 1997a, 1997b; Pacharavanich *et al.*, 2000; Ribeiro, 1999).

Ribeiro (1999) descreveu, em linhas gerais, a relação entre RBC e sistemas multi-agentes, propondo um sistema voltado para a avaliação de imóveis. Segundo este autor, na avaliação do valor de uma propriedade, o avaliador utiliza mecanismos de analogia para encontrar as relações entre o avaliando e as propriedades similares utilizadas como amostra do comportamento do mercado. Entretanto, apresentou um formato genérico, sem o efetivo desenvolvimento do sistema.

González e Laureano-Ortiz (1992) afirmam que as técnicas usuais para avaliação de imóveis por comparação de dados não capturam o conhecimento heurístico, o que poderia ser realizado por RBC. Esses autores entendem que o RBC parece ser mais adequado para processos que decorrem do comportamento psicológico (humano), como o mercado

imobiliário. O sistema proposto por González e Laureano-Ortiz (1992) exige a intervenção de especialistas para a definição dos fatores de ajuste a serem aplicados na fase de revisão. A aplicação desenvolvida recupera os dez casos mais similares ao avaliando e o grau de similaridade depende da importância relativa das características consideradas. A adaptação é realizada ajustando o valor de venda de cada caso selecionado, através de um método chamado "aplicação crítica", que utiliza parâmetros de ajuste obtidos de um especialista. Por fim, a solução decorre da aplicação dos ajustes determinados, que são adicionados ou subtraídos ao valor de venda de cada propriedade recuperada do arquivo. Após este ajuste, o sistema seleciona três imóveis e calcula a média dos preços destes imóveis.

O'Roarty *et al.* (1997a, 1997b) apresentaram um sistema para a avaliação de imóveis comerciais, utilizado especialmente para identificar casos similares. Os pesos foram derivados de especialistas, através da literatura ou de entrevistas. A base de casos foi estruturada com 82 casos e 76 atributos, utilizando um índice composto por 32 destes atributos. Bonissone *et al.* (1998) apresentaram um sistema híbrido, no qual as medidas de similaridade são aperfeiçoadas através de lógica difusa. A base de casos deste trabalho foi de mais de 35 mil casos, com um grande conjunto atributos. O sistema identifica os casos mais similares e ajusta estes casos ao avaliando através de regras previamente definidas. Após os ajustamentos, novamente identifica e seleciona os casos mais semelhantes ao avaliando e finalmente calcula a média destes casos. Pacharavanich *et al.* (2000) desenvolveram um sistema para avaliações residenciais em Bangkok, também utilizando um conjunto de pesos pré-determinado para identificar a similaridade, selecionando então três casos. A base de casos continha 236 imóveis e os pesos foram calculados através de regressão múltipla.

Outro trabalho, de McSherry (1998), descreve uma forma alternativa de adaptação, que foi implementada em um sistema RBC, no qual o conceito de "caso dominante" é fundamental, na adaptação e na manutenção da consistência da base de casos. Um caso é dominante a outro se todos os seus atributos são iguais ou superiores. Assume-se que os valores dos atributos dos casos podem ser organizados em ordem crescente. Assim, a heurística de adaptação segue regras simples: dado um caso para o qual se deseja determinar o valor, denominado C_1, o sistema busca um caso que seja diferente deste apenas em um atributo (C_2). Em seguida, busca outros dois casos, C_3 e C_4, que também são distintos apenas neste atributo, de tal forma que C_3 tem valor igual a C_1 para o atributo em questão, bem como C_2 e C_4 conferem. Então,

calcula-se o valor de C_1 por: $est(C_1)=est(C_2)+est(C_3)–est(C_4)$. Entretanto, esta forma é adequada para sistemas em que o valor seja determinado por uma função linear e aditiva, o que pode não ocorrer em muitas situações. Ademais, encontrar vários casos em que apenas um atributo seja distinto também não é tarefa fácil, diante da heterogeneidade dos imóveis.

4.7 ALGORITMOS EVOLUCIONÁRIOS

A computação evolucionária exibe um comportamento adaptativo, que ajuda a lidar com problemas não lineares e de grandes dimensões, sem requerer diferenciabilidade de funções ou conhecimento explícito da estrutura do problema. Como resultado, estes algoritmos são robustos ao comportamento variável no tempo, mesmo que possam ter ocasionalmente baixa velocidade de convergência, também com o risco de parar em ótimos locais. A computação evolucionária reúne uma família de algoritmos estocásticos, incluindo três correntes básicas (Bonissone *et al.*, 1999):

a) algoritmos genéticos;

b) estratégias de evolução;

c) programação evolutiva.

Foi apontado por Fogel (1995[29], *apud* Bonissone *et al.*, 1999) que os três têm várias características comuns, pois todos mantém uma população de soluções tentadas, impõe mudanças aleatórias a estas soluções e incorporam seleção para determinar quais soluções manter em futuras gerações. Fogel também afirma que os algoritmos genéticos enfatizam modelos de operadores genéticos similares aos encontrados na natureza. Bonissone *et al.* (1999) afirmam que os vários componentes evolutivos têm ampliado suas semelhanças, com o passar do tempo. Bäck e Kursawe (1995) afirmam que ainda existe discussão teórica sobre as vantagens e desvantagens de aplicação dos algoritmos evolucionários.

[29] FOGEL, D. B. **Evolucionary computation**. New York: IEEE Press, 1995.

4.7.1 Algoritmos Genéticos

Os algoritmos genéticos são os mais conhecidos e empregados dentre os algoritmos evolucionários. São mecanismos de busca ou otimização, inspirados nos mecanismos de seleção natural e na genética natural. Eles combinam sobrevivência da estrutura mais ajustada com troca de informação aleatorizada, formando um algoritmo de busca com elementos que lembram o comportamento inovativo humano. Em cada geração, um novo conjunto de criaturas artificiais é gerado, usando partes dos elementos mais ajustados da geração anterior. Algumas novas partes podem ser incluídas ou alteradas aleatoriamente (Goldberg, 1989).

Os dados são convertidos em um código, representando um gene ou cromossomo (indivíduo), e o processamento ocorre através dos operadores genéticos de seleção natural (indivíduos mais aptos), cruzamento e mutação. Estes operadores alteram a população inicial composta de soluções candidatas, gerando indivíduos novos, potencialmente melhores. A seleção é realizada através de um mecanismo probabilístico (geralmente através de um mecanismo conhecido como "roleta") garantindo a tendência de que os indivíduos com maior aptidão (mais próximos da solução) permanecem para a próxima geração. O cruzamento ocorre com a fusão de dois indivíduos, divididos em uma posição aleatória e recombinados. Já a mutação consiste na alteração aleatória de parte do código de alguns indivíduos. Muitas vezes os cromossomos são binários (Goldberg, 1989; Man *et al.*, 1999), mas também podem ser números reais (Goldberg, 1991; Blekas e Stafylopatis, 1996).

Por outro lado, segundo Mitchell (1999), não existe uma definição aceita de algoritmos genéticos que permita diferenciar totalmente os algoritmos genéticos dos outros métodos evolucionários. Porém, pode-se dizer que os métodos chamados genericamente de algoritmos genéticos têm ao menos os seguintes elementos em comum: populações de cromossomos (representando potenciais soluções), seleção de acordo com uma função-objetivo (*fitness*), cruzamento para produzir uma nova descendência e mutação aleatória destes novos descendentes. A inversão, que consiste no quarto elemento dos algoritmos genéticos e foi introduzido por Holland, raramente é utilizado, pois suas vantagens ainda não são bem conhecidas (Man *et al.*, 1999; Mitchell, 1999).

Para Goldberg (1989, p.7), os algoritmos genéticos são diferentes dos procedimentos comuns de otimização em quatro aspectos fundamentais:

a) trabalham com uma codificação do conjunto de parâmetros e não com os próprios parâmetros;

b) buscam a solução a partir de uma população de pontos, e não de um ponto simples;

c) usam diretamente funções-objetivo e não derivadas ou outro conhecimento auxiliar;

d) usam regras de transição probabilísticas e não determinísticas.

Estes quatro elementos também diferenciam os algoritmos genéticos dos outros mecanismos evolucionários, como é discutido adiante. Mesmo contando com componentes aleatorizados, os algoritmos genéticos não são simples passeios aleatórios. Eles exploram eficientemente a informação histórica para especular novos pontos de busca. O ponto central de pesquisa em algoritmos genéticos tem sido sua robustez, ou seja, o balanço entre eficiência e eficácia (Goldberg, 1989). O fluxo de funcionamento dos algoritmos genéticos pode ser resumido como na Figura 5 (adaptada de Goldberg, 1989).

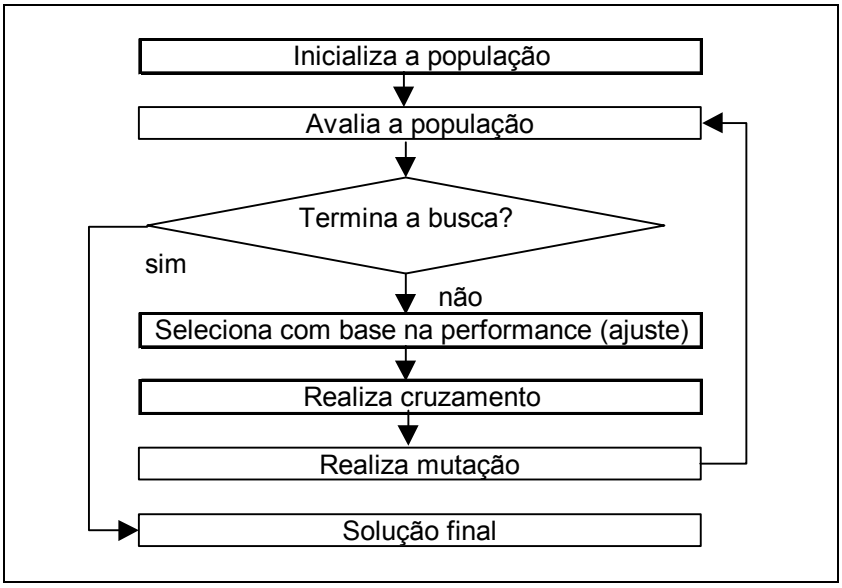


Figura 5: Mecanismo de funcionamento dos algoritmos genéticos (adaptada de Goldberg, 1989).

Embora a versão básica dos algoritmos genéticos seja simples, existem aplicações importantes, tais como otimização, programação automática, aprendizagem de máquina, economia, ecologia (modelagem de fenômenos como co-evolução de hospedeiro/parasita, simbiose e proliferação de armas biológicas), genética populacional, evolução e aprendizagem, sistemas sociais (tais como modelagem de comportamento social de colônias de insetos ou esquemas de cooperação em sistemas multi-agentes), otimização de recursos (redes de distribuição de água, cargas e dimensões em estruturas), e diversos outros casos (Mitchell, 1999). Também são muito empregados em sistemas híbridos, no ajuste de topologias neurais ou na geração de bases de regras difusas (Bonissone *et al.*, 1999; Man *et al.*, 1999).

As razões para o crescente número de aplicações são claras, segundo Goldberg (1989, p.2): os algoritmos são poderosos mas computacionalmente simples e não são limitados por condições restritivas sobre o espaço de busca (tais como continuidade ou existência de derivadas nas funções), vantagens que paradoxalmente podem ser limitações, como apontam Illich e Simovic (1998), pois tornam a busca excessivamente ampla, em alguns casos.

A programação genética (*genetic programming*) , proposta por Koza, é um sub-campo dos algoritmos genéticos. É uma forma especial de algoritmos genéticos, na qual os cromossomos têm uma estrutura hierárquica e não linear, e cujo tamanho não é predefinido. Os indivíduos são programas estruturados em árvores, e os operadores genéticos são aplicados a seus ramos (sub-árvores) ou a nós isolados. Este tipo de cromossomo pode ser representado por uma expressão executável, tal como uma expressão S em Lisp, uma expressão aritmética, etc. Este formato tem sido usado em predições de séries temporais, controle, otimização de topologias em redes neurais, entre outros casos (Bonissone *et al.*, 1999; Soh e Y. Yang, 2000).

4.7.1.1 Mecanismo básico de funcionamento

Nos primeiros algoritmos genéticos, os cromossomos eram tipicamente binários. Neste formato, cada *locus* em um cromossomo tem dois alelos possíveis: 0 e 1. Cada cromossomo pode ser visto como um ponto (de dimensão n) no espaço de busca de soluções. Os algoritmos genéticos processam populações de cromossomos, substituindo progressivamente uma

população por outra, potencialmente mais ajustada. Para isto, requerem uma função de *fitness* que assinale um escore a cada cromossomo na população corrente, dependendo de quão bem o cromossomo resolve o problema em questão (Mitchell, 1999).

Um algoritmo genético simples é composto basicamente de três operadores: reprodução (através de seleção dos indivíduos mais ajustados), cruzamento e mutação. A reprodução é um processo no qual os indivíduos são copiados de acordo com os valores das suas funções-objetivo f (funções de *fitness*). A função f pode ser entendida como uma medida de lucro ou utilidade, a ser maximizada. Copiar de acordo com f significa que os indivíduos com maior valor de f têm maior probabilidade de contribuir com um ou mais descendentes na próxima geração. Este operador pode ser implementado de várias formas. Provavelmente a forma mais simples é a roleta, com as casas de cada indivíduo proporcionais ao seu *fitness*. Cada elemento selecionado é copiado para uma população provisória (Goldberg, 1989). A seleção é realizada sobre uma população provisória (P''_t) para gerar a nova população, (P_{t+1}). A probabilidade de seleção dos indivíduos é dada pela proporção de seu ajustamento em relação ao somatório de ajustamentos da população (seleção proporcional), como apontado na Equação 5 (Goldberg, 1989):

$$p(a_i) = \frac{f(a_i)}{\sum_{j=1}^{n} f(a_i)} \qquad \text{(Equação 5)}$$

Esta definição assume valores de *fitness* positivos e uma tarefa de maximização. Os algoritmos genéticos geralmente mantêm população de tamanho constante (isto é, $\mu=\lambda$), da ordem de 50 a 100 indivíduos. Normalmente esta população é inicializada aleatoriamente (Bäck e Kursawe, 1995), mas o analista pode fazer uso do conhecimento anterior na formulação destes indivíduos, o que pode ser apontado como uma das vantagens dos algoritmos genéticos (Cordón *et al.*, 2001).

Após a reprodução, o cruzamento é realizado em duas etapas. Em primeiro lugar, são escolhidos aleatoriamente alguns pares de indivíduos. Em seguida, para cada par é escolhido um número inteiro k, representando a posição de corte, no intervalo [1, n], sendo n o tamanho do cromossomo (quantidade de genes). Os novos indivíduos são gerados (re)combinando as

parcelas [1, k] e [k+1, n] de cada um dos dois indivíduos (Goldberg, 1989).

Bäck e Kursawe (1995) lembram que o mecanismo de cruzamento simples pode ser facilmente estendido para cruzamento múltiplo (vários pontos de corte), e até para cruzamento uniforme, no qual a decisão aleatória sobre a troca de informação entre genes é realizada bit a bit.

A mutação consiste na alteração aleatória de uma parte do indivíduo. Este operador realiza uma função importante, trazendo inovação e diminuindo os riscos de convergência prematura (Goldberg, 1989). A mutação é geralmente interpretada como sendo de importância marginal em algoritmos genéticos. Funciona invertendo aleatoriamente genes isolados, escolhento o gene a ser modificado (gene m) com uma probabilidade pequena. Geralmente p_m assume valores na faixa de 0,01 a 0,001. Investigações mais recentes, todavia, indicam que a mutação pode ser importante em alguns casos e recomendam que p_m fique em torno de $p_m=1/n$, sendo n o número de bits do cromossomo, segundo Bäck e Kursawe (1995).

Outro ponto interessante é a especificação de um determinado número de indivíduos que são selecionados determinísticamente, escolhendo-se os melhores (mais ajustados) em cada geração, antes da seleção aleatória. Esta forma é conhecida como "elitismo". Dois exemplos de operações genéticas são apresentados na Figura 6, a seguir.

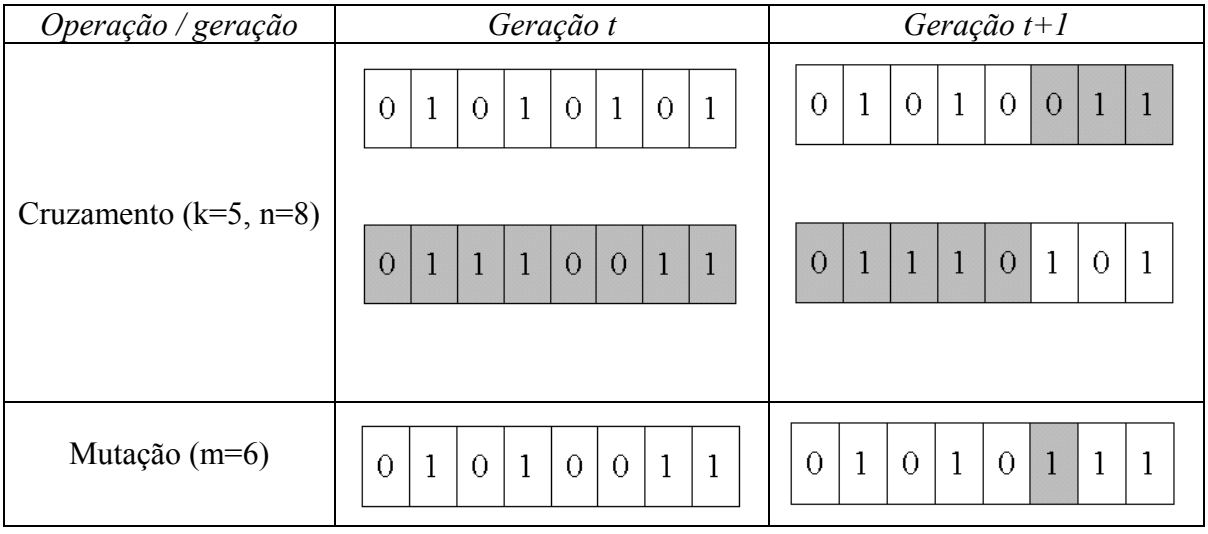


Figura 6: Exemplos de operações genéticas

Outra questão importante é a determinação dos parâmetros das operações. Segundo Bäck e

Kursawe (1995), a probabilidade de cruzamento (p_c) é geralmente da ordem de 0,6. Para Man *et al.* (1999), é comum adotar-se uma taxa de cruzamento (p_c) com valores entre 0,6 e 1,0, enquanto que a probabilidade de mutação é tipicamente menor do que p_m=0,01. Dependendo do domínio, a escolha de p_m e p_c como parâmetros de controle pode ser um problema complexo de otimização não linear. Ademais, dependem criticamente da natureza da função objetivo. De Jong (1975[30], *apud* Grefenstette, 1986) desenvolveu grande estudo exploratório sobre os parâmetros, concluindo pelo uso de populações de 50 a 100 indivíduos, com taxas de cruzamento de 0,6 e de mutação de 0,001. Grefenstette (1986) revisou o trabalho desse pesquisador, indicando populações de 30 indivíduos, cruzamento de 0,95, de mutação de 0,01 e escolha elitista de um indivíduo por geração como valores ótimos. Contudo, há resultados contraditórios. Na aplicação de Sharif e Wardlaw (2000), a aplicação de mais de uma mutação por indivíduo degradou a performance. Este assunto ainda não está resolvido, embora existam algumas sugestões na literatura (Man *et al.*, 1999):

a) populações grandes (100 ou mais indivíduos): p_c=0,6 e p_m=0,001;

b) populações pequenas (até 30 indivíduos): p_c=0,9 e p_m=0,01.

Por outro lado, alguns pesquisadores, como Spears (1993, 1995), apontam para um relativo equilíbrio entre os operadores de cruzamento e mutação. Para esse autor, não há provas da superioridade de um ou de outro, defendendo a análise do desempenho de ambos, nos casos concretos. De qualquer forma, aparentemente os parâmetros são dependentes do problema a ser resolvido, ou seja, é necessário explorar os efeitos da variação das taxas, do tamanho da população e do tempo de treinamento (número de gerações).

A codificação é outro ponto importante na especificação de uma aplicação em algoritmos genéticos. Originalmente, os algoritmos genéticos eram codificados em cromossomos binários. Entretanto Man *et al.* (1999) afirmam que, recentemente, a manipulação direta de cromossomos com valores reais recebeu considerável interesse, especialmente para trabalhar com problemas que têm parâmetros reais. Para estes autores, alguns trabalhos indicam que a representação em ponto flutuante pode ser mais rápida em termos de computação (por exemplo, evitando as conversões no início e no final do processamento) e também mais

[30] DE JONG, K.A. **An analysis of the behavior of a class of genetic adaptive systems**. 1975. PhD Thesis. University of Michigan, Ann Arbor.

consistente. Contudo, nem sempre é superior à codificação binária, segundo estes autores.

No caso de codificações em números reais, o cruzamento é realizado através do cruzamento aritmético, com a combinação linear dos genes correspondentes de um par de indivíduos (Cordón *et al.*, 2001).

4.7.1.2 Fundamentos teóricos dos algoritmos genéticos

A teoria tradicional dos algoritmos genéticos é baseada em Holland (1975[31], *apud* Mitchell, 1999), seu principal criador. Fundamentalmente, os algoritmos genéticos funcionam descobrindo, enfatizando e recombinando bons "blocos de construção" em um esquema altamente paralelo. A idéia é que boas soluções tendem a ser construídas a partir de bons blocos. Holland introduziu o conceito de esquemas (ou "*schemata*") para formalizar a noção informal de "blocos". Um esquema é um conjunto de bits que podem ser 0, 1 ou ***,** sendo que o asterisco representa "qualquer um deles", por exemplo, H=1***1. Os *strings* possíveis para este esquema, tais como 10111 ou 11011, são instâncias de H. O esquema H é dito "de ordem 2", pois tem dois bits definidos. A distância entre os bits definidos mais distantes é seu "tamanho de definição", no caso igual a 4 (Goldberg, 1989; Mitchell, 1999).

Para Man *et al.* (1999), o paralelismo implícito ou intrínseco dos algoritmos genéticos advém do fato que muitos hiperplanos são amostrados quando uma população de *strings* é avaliada. O paralelismo da busca implica em que muitas competições entre hiperplanos são resolvidas simultaneamente (em paralelo). Para estes autores, esta teoria sugere que através do processo de reprodução e recombinação, os esquemas de hiperplanos competidores aumentam ou diminuem sua representação na população, de acordo com o ajustamento relativo dos strings que representam estes hiperplanos.

A "teoria dos esquemas" preocupa-se principalmente com os efeitos destrutivos do cruzamento e da mutação. Contudo, acredita-se que o cruzamento é uma das maiores fontes

[31] HOLLAND, J. H. **Adaptation in natural and artificial systems**. Ann Arbor: University of Michigan Press, 1975.

de poder em algoritmos genéticos, com a habilidade de recombinar instâncias de bons esquemas formando instâncias de esquemas igualmente bons ou potencialmente melhores (Goldberg, 1989).

Outra visão, ligeiramente diferente, é a da "hipótese dos blocos de construção". Neste caso, os blocos são esquemas ajustados de pequeno tamanho de definição (Goldberg, 1989, p.20). Os operadores genéticos de cruzamento e mutação têm a habilidade de gerar, promover e justapor, lado a lado, blocos de construção, para formar os strings ótimos. O cruzamento tende a conservar a informação genética presente nos strings. Assim, quando os *strings* são similares, sua capacidade de gerar novos blocos diminui. Por outro lado, a mutação é capaz de gera blocos radicalmente novos, pois não é um operador conservativo. A seleção dos elementos que devem permanecer para a próxima geração é um procedimento que tende a ser viesado para os blocos que possuem maior ajustamento, garantindo sua representação de geração a geração. Esta hipótese sugere que o problema de codificação é crucial para a performance dos algoritmos genéticos, e que tal codificação deve satisfazer a idéia de blocos pequenos, para ampliar a possibilidade de intercâmbio (Man *et al.*, 1999).

Contudo, Mitchell (1999) afirma que existem controvérsias sobre algumas abordagens, revelando que os algoritmos genéticos não têm uma teoria acabada mas, ao contrário, ainda apresentam questões sem respostas. Segundo esta autora, embora os algoritmos genéticos sejam largamente empregados e sejam simples de descrever e programar, seu comportamento pode ser complexo, não está ainda completamente compreendido seu funcionamento.

Para Illich e Simovic (1998), os algoritmos genéticos não podem ser aplicados a qualquer problema, sua eficiência varia de muito eficiente a ineficiente, em função da complexidade e tamanho do problema, os algoritmos genéticos buscam soluções em todo o espaço, mesmo em regiões de soluções inviáveis, e o esforço de busca pode ser 99% nesta área (perdido), os algoritmos genéticos não levam em conta a forma e o gradiente da função objetivo, que poderiam aumentar as chances de se encontrar um ótimo global e, por fim, os algoritmos genéticos requerem sintonia dos parâmetros de busca.

4.7.2 Estratégias Evolutivas

As estratégias evolutivas surgiram na Alemanha, inicialmente com Rechemberg, aplicando seleção e mutação em uma população de apenas um indivíduo. Schwefel introduziu o conceito de recombinação e ampliou a população para mais de um indivíduo. Um par de pais gera um filho via recombinação, que é posteriormente perturbado via mutação (Bäck *et al.*, 1992; Man *et al.*, 1999). O número de filhos gerados é maior do que N e a sobrevivência é determinística. A sobrevivência ocorre por uma das seguintes formas: (a) sobrevivência dos n melhores filhos, que substituem os pais, ou (b) sobrevivem os n melhores entre pais e filhos. A mutação é o operador mais importante, e cada variável pode ser mudada de acordo com uma distribuição de probabilidade (Herrera *et al.*, 1995). As estratégias evolutivas consistem atualmente de um importante algoritmo para otimização contínua de parâmetros (Bäck e Kursawe, 1995; Bonissone *et al.*, 1999).

Na notação típica de estratégias evolutivas, os símbolos (μ, λ)-ES e $(\mu+\lambda)$-ES indicam os algoritmos nos quais uma população de μ pais gera λ descendentes. Os melhores μ indivíduos são selecionados e mantidos na geração seguinte. No caso de (μ, λ)-ES os pais são excluídos da seleção e não participam da próxima geração, enquanto que em $(\mu+\lambda)$-ES eles participam da seleção. Nas primeiras versões, as estratégias evolutivas utilizavam uma representação contínua e operadores de mutação trabalhando sobre um único indivíduo, isto é, (1+1)-ES. Depois foi acrescentado um operador de recombinação e as estratégias evolutivas foram estendidas para evoluir uma população de indivíduos. Cada indivíduo tem a mesma oportunidade de gerar um descendente e cada um pode ter seu próprio operador de mutação, que pode ser herdado e pode ser diferente para cada parâmetro (Bonissone *et al.*, 1999).

Em contraste com os algoritmos genéticos, as soluções são diretamente representadas por vetores reais e indivíduos que consistem não só destes vetores, mas também dos seus desvios-padrão σ_i (positivos). Estes parâmetros estratégicos σ_i são utilizados pelo operador de mutação para modificar as variáveis x_i correspondentes. A mutação age sobre cada uma das variáveis x_i adicionando números aleatórios Normalmente distribuídos, com média zero e variância σ_i^2 ($N(0, \sigma_i^2)$). Os desvios-padrão σ_i não são constantes e nem são explicitamente considerados (Bäck e Kursawe, 1995).

A mutação de σ_i é baseada em um fator global $\tau'.N(0,1)$ (um único número aleatório é obtido para o indivíduo inteiro) e em um fator local $\tau.N_i(0,1)$ (um número aleatório para cada componente). Os elementos τ representam uma espécie de "taxa de aprendizagem", e devem ser definidos de forma que τ' seja $(2n)^{-0,5}$ e τ seja $(2.n^{0,5})^{-0,5}$, aproximadamente (Bäck e Kursawe, 1995; Bäck *et al.*, 1992).

Este mecanismo de "auto-aprendizagem" dos parâmetros estratégicos auxilia a adaptação destes parâmetros sem a necessidade de encontrar um mecanismo de controle exógeno adequado (sem *fitness*). Além dos desvios-padrão, as covariâncias da distribuição normal generalizada de dimensão n também podem ser usadas na auto-adaptação, introduzindo as "mutações correlacionadas" ao algoritmo. Este mecanismo pode acelerar a busca em casos de topologia local complicada. Embora a mutação seja o operador mais importante, a recombinação dos parâmetros estratégicos e das variáveis é necessária para o processo de auto-adaptação e é geralmente útil no progresso da busca (Bäck e Kursawe, 1995).

Outra função do mecanismo de busca é a transição do tamanho da população, de μ para λ indivíduos (são comuns μ=15 e λ=100). Isto ocorre aplicando-se a recombinação ao nível do indivíduo λ vezes (ao contrário dos algoritmos genéticos, não há uma "taxa de recombinação" menor do que 1, pois a recombinação é sempre aplicada). Finalmente, o operador de seleção é totalmente determinístico e funciona escolhendo os n melhores indivíduos de P"(t) para se tornarem parte da próxima geração. A seleção (μ, λ)-ES é preferida, pois suporta o mecanismo de auto-adaptação (prevenindo uma possível extinção, causada por escolha inadequada dos parâmetros estratégicos) e ajuda a aplicação de estratégias evolutivas no caso de funções objetivo perturbadas ou com variação no tempo. Deve ser usada uma estratégia (μ, λ)-ES com μ não muito pequeno, para que a pressão seletiva não seja muito forte (por exemplo, (15, 100)-ES), e o operador de recombinação deve ser aplicado também nos parâmetros estratégicos (Bäck e Kursawe, 1995; Bäck *et al.*, 1992).

Embora à primeira vista a representação dos indivíduos pareça ser a distinção mais importante, o conceito de auto-adaptação dos parâmetros estratégicos – que não existe nos algoritmos genéticos – é mais significativo. O processo de sintonia dos parâmetros, "à mão", que freqüentemente é demorado em algoritmos genéticos, não é necessário nas estratégias evolutivas. Considerando os operadores genéticos, a mutação tem mais ênfase em estratégias evolutivas, enquanto que a recombinação predomina em algoritmos genéticos, o que é

explicado pela modelagem em nível fenotípico nas estratégias evolutivas e genotípico nos algoritmos genéticos. Por último, a seleção é determinística nas estratégias evolutivas (os melhores) e probabilística nos algoritmos genéticos, que contam com uma chance maior do que zero de escolha de qualquer indivíduo, mesmo para os piores (Bäck e Kursawe, 1995).

4.7.3 Programação Evolutiva

A programação evolutiva foi desenvolvida por Fogel, na década de 60 e tem tradicionalmente utilizado representações específicas dos problemas, de acordo com o domínio. Os algoritmos de programação evolutiva são freqüentemente usados como otimizadores, embora tenham sido propostos como técnicas de aprendizagem de máquina. A forma de mutação é baseada na representação usada, e é geralmente adaptativa. A recombinação não é necessária, pois as formas de mutação usadas são muito flexíveis. Depois da inicialização, todos os N indivíduos são selecionados para serem pais, sendo modificados para produzir N filhos. Estes filhos são avaliados em conjunto com os pais, e N dos 2N indivíduos sobrevivem, usando uma função probabilística baseada na função de adaptação (Bonissone *et al.*, 1999; Herrera *et al.*, 1995).

A programação evolutiva é aplicada em identificação, predição de seqüências, controle automático, reconhecimento de padrões e estratégias ótimas de jogo. Para atingir estas metas, a programação evolutiva evolui máquinas de estado finito que tentam maximizar uma função de lucro. Versões mais recentes têm utilizado otimização contínua de parâmetros e treinamento de redes neurais, tornando-se similares às estratégias evolutivas. A distinção principal é que a programação evolutiva lida com o comportamento de espécies e geralmente não emprega cruzamento, enquanto que a estratégia evolutiva foca nos indivíduos. A programação evolutiva pode ser considerada como um caso especial de $(\mu+\mu)$-ES, no qual a mutação é o único operador usado na busca de novas soluções (Bonissone *et al.*, 1999).

4.7.4 Aplicações de algoritmos evolucionários em Engenharia Civil

Existem diversas aplicações na área da Engenharia Civil, incluindo trabalhos sobre planejamento de canteiros de obras, otimização de estruturas metálicas e também sobre recursos hídricos. Sharif e Wardlaw (2000) apresentaram um sistema de controle do nível de

múltiplos reservatórios em usinas hidroelétricas. Lippai *et al.* (1998) demonstraram o uso de algoritmos genéticos em um projeto de distribuição de água, na otimização do custo do sistema. A população foi composta de 100 indivíduos, com os algoritmos genéticos atingindo os resultados esperados em 460 gerações, usando uma taxa de cruzamento de 50% e mutação de 6%, com o software Evolver rodando por 10h. Illich e Simovic (1998) apresentaram um algoritmo evolucionário para o caso da determinação do custo total de bombeamento de líquidos, cuja abordagem continha várias diferenças em relação aos algoritmos genéticos tradicionais, tais como não ter cruzamento ou mutação. Usaram 1000 indivíduos, e a solução foi atingida em poucas gerações (de 2 a 5, dependendo do exemplo). Eles testaram diversas distribuições de probabilidade para a geração do conjunto inicial.

J. Yang e Soh (1997) apresentaram a otimização de estruturas treliçadas metálicas, com uma abordagem mista de algoritmos genéticos. A população foi de 40 indivíduos, com p_c=0,87 e p_m=0,004. Soh e Y. Yang (2000) usaram programação genética para otimização de estruturas treliçadas, incluindo topologia da estrutura, dimensões e geometria das peças. Encontraram bons resultados, usando p_r=10%, p_c=80% e p_m=10%, com uma população de 500 indivíduos.

4.8 MODELOS ADITIVOS GENERALIZADOS

Os modelos aditivos generalizados foram propostos por Hastie e Tibshirani (1986) e são uma extensão dos modelos lineares, tais como a regressão convencional e os modelos lineares generalizados[32]. Consistem em um modelo estatístico mais flexível, que emprega o somatório de um conjunto de funções, com a forma da Equação 6:

$$Y = f_1(x_1) + f_2(x_2) + ... + f_k(x_k) + e \qquad \text{(Equação 6)}$$

Onde as funções f_i podem assumir qualquer formato. Os modelos lineares são um caso especial, no qual $f_i(x_i)=a_i x_i$. Para Plate *et al.* (2000), se os erros seguem uma distribuição

[32] Nos modelos lineares a variância da variável dependente Y deve ser constante (é a condição de homocedasticidade), enquanto que no modelos lineares generalizados a variância de (Y) é uma função do valor

Normal, a função identidade realmente deve ser a preferida para compor o somatório. Se a variável dependente adotar um formato logarítmico ou exponencial, o modelo assume um formato multiplicativo (Pace, 1998). Geralmente as funções parciais f_i são baseadas em uma única variável, o que facilita a interpretação, geralmente desenvolvida através de gráficos de cada função contra a variável dependente (Hastie e Tibshirani, 1986, 1997).

Os modelos aditivos generalizados são não-parametricos, ou seja, não há condições a serem previamente atendidas pelo modelo, tais como o conhecimento da forma funcional ou do comportamento dos erros (Hand *et al.*, 2001; Hastie e Tibshirani, 1986; Pace, 1998). O ajuste dos coeficientes, todavia, é um dos problemas a serem enfrentados. Se existem diversas variáveis independentes, cada uma delas adotando uma função não-linear distinta, o custo computacional da solução é muito superior ao dos modelos lineares. Diversas soluções foram sugeridas para o cálculo dos coeficientes, utilizando algoritmos de retropropagação, estratégias semi-paramétricas ou técnicas de cálculo numérico tal como o algoritmo de Gauss-Seidel (Anglin e Gençay, 1996; Hastie e Tibshirani, 1986, 1997; Stone *et al.*, 1997). Uma outra alternativa, adotada neste trabalho, são os algoritmos genéticos (ver capítulo 7).

No âmbito do mercado imobiliário, esta técnica foi explorada em alguns trabalhos. Anglin e Gençay (1996) e Pace (1998) realizaram comparações de modelos aditivos generalizados com diversos formatos de regressão, verificando que os modelos aditivos generalizados superavam os modelos de regressão. Por outro lado, Mason e Quigley (1996) encontraram pequenas diferenças em relação ao modelo linear.

4.9 REDES NEURAIS ARTIFICIAIS

As redes neurais artificiais (RNA) são baseadas nos estudos sobre o comportamento do cérebro humano, buscando simular sua forma de processar informações. Não existe uma definição universal para redes neurais, mas é comum encontrar na literatura da área a idéia de que uma RNA é um conjunto de múltiplos processadores (unidades), cada qual tendo uma pequena quantidade de memória, e ligados entre si por canais de comunicação (conectores),

médio de Y (Han e Kamber, 2001).

que têm a capacidade de transportar dados numéricos, codificados em vários formatos (sinais), à semelhança do neurônio humanos. As unidades operam apenas seus dados locais e suas entradas ocorrem pelas conexões. Uma das vantagens das RNA sobre as técnicas estatísticas convencionais (como a análise de regressão) é que as redes não exigem conhecimento prévio sobre o formato dos relacionamentos e não há requisitos especiais quanto aos dados (Kauko, 1997; Portugal e Fernandes, 1996; Rumelhart e MacClelland, 1995; Tay e Ho, 1994).

As RNA podem ser entendidas como mecanismos de reconhecimento de padrões, com habilidade de auto-aprendizagem. Neste sentido, a técnica vem sendo empregada atualmente na identificação de potencialmente bons (ou maus) clientes, na análise de investimentos, no reconhecimento de caracteres, de imagens ou de voz, e também na avaliação de imóveis (Rossini, 1997; Tay e Ho, 1994; Worzala *et al.*, 1995).

É mais correto denominá-las de "redes neurais artificiais", para diferenciar das redes biológicas, embora alguns textos referenciem apenas "redes neurais". Nem todas as RNA são modelos de redes neurais biológicas, embora estas sejam a inspiração costumeira, na busca por mecanismos de reprodução do cérebro humano. As RNA também são chamadas de "modelos conexionistas de computação", em função da questão das conexões entre as unidades, que é fundamental no desempenho das redes (Aubin, 1996; Haykin, 2001; Portugal e Fernandes, 1996).

O elemento fundamental nas redes é a unidade de processamento, denominada "neurônio artificial", que recebe os sinais provenientes das entradas ou de outras unidades, processa as informações e repassa os resultados às unidades seguintes ou às saídas. Uma unidade pode ser representada esquematicamente da seguinte forma (Figura 7):

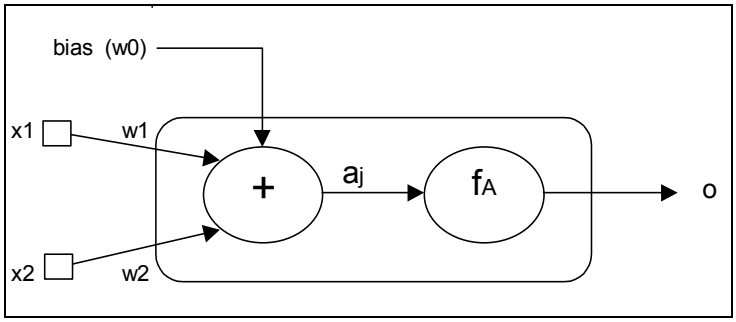


Figura 7: Representação de uma unidade isolada (neurônio)

Basicamente, a unidade realiza a função de receber os sinais ponderados ($w_i x_i$), processá-los inicialmente através de uma função de soma, aplicar uma função de ativação (f_A) e enviá-los aos elementos seguintes da rede. A unidade pode ter várias entradas, mas tem apenas um valor de saída, ainda que com múltiplas cópias. A rede mais simples é uma conjugação de duas ou mais unidades deste tipo. Havendo diversos neurônios artificiais, as possibilidades de diferentes interligações crescem. Conforme a topologia da rede, podem ser reforçadas determinadas habilidades.

O modelo conhecido como Perceptron foi proposto por Rosenblatt, em 1957. Uma rede neural artificial deste tipo pode ser extremamente complexa, conforme a quantidade de neurônios artificiais que contenha, mas apresenta uma configuração básica que reproduz a estrutura da Figura 8. A rede é composta de uma camada de entrada e outra de saída, interligadas por conexões ponderadas (Aubin, 1992; Kovács, 1996; White, 1992).

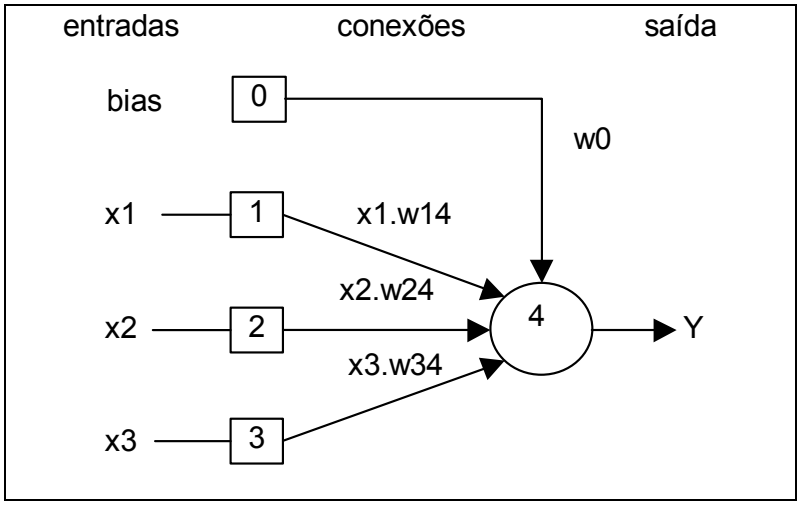


Figura 8: Rede neural artificial de duas camadas (Perceptron)

As unidades de entrada (sensores) enviam os sinais x_i à unidade de saída x_j e os valores transmitidos são ponderados nas conexões, através de um vetor de pesos W. Assim, cada sinal passado de um neurônio i a um neurônio j será reforçado ou atenuado na conexão através de um peso w_{ij}. Definir um peso $w_{ij}=0$ significa desfazer uma conexão. Uma função de saída simples para a rede da Figura 8 poderia assumir uma forma equivalente a $Y=f_A(w_0+w_{14}x_1+w_{24}x_2+w_{34}x_3)$, por exemplo.

Em um caso mais geral, podem ser incluídas unidades intermediárias (chamadas de neurônios

ocultos), que recebem os sinais, processando e encaminhando aos níveis seguintes. Podem existir diversos níveis ocultos, sempre com uma camada de entrada e outra de saída. Já foi demonstrado que uma função contínua qualquer pode ser implementada exatamente em uma rede neural artificial de três camadas, a mais utilizada atualmente. Diversos autores contribuíram para esta demonstração, especialmente Hecht-Nielsen e Cybenko, no final da década de 80, ambos utilizando como base o Teorema da Superposição, de Kolmogorov (Haykin, 2001; Kauko, 1997; Kovács, 1996). Neste caso, a rede assume um aspecto como a rede da Figura 9.

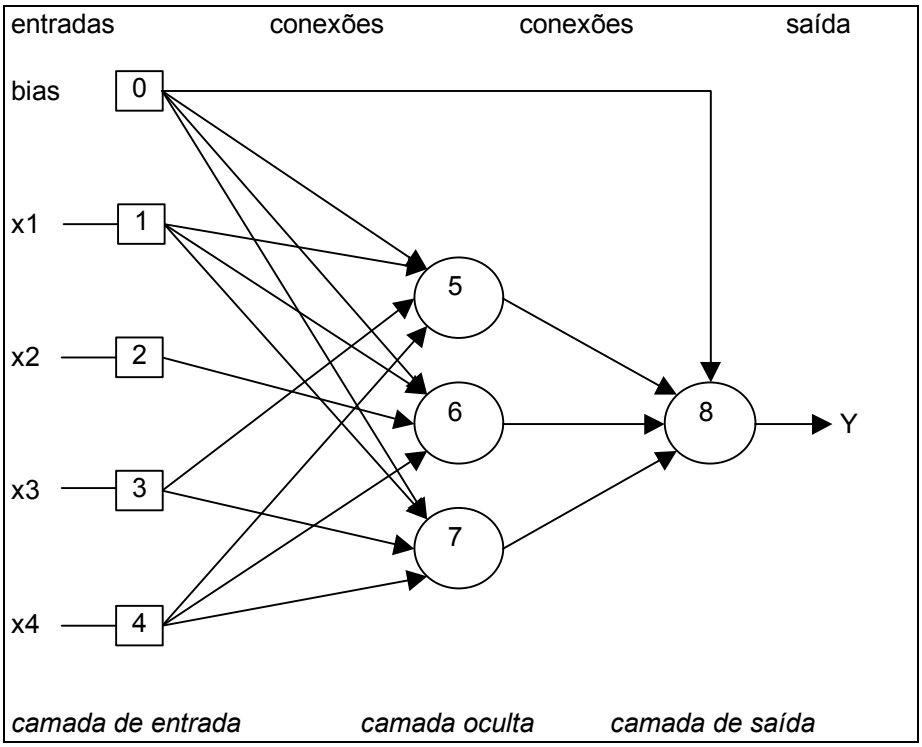


Figura 9: Rede neural de três camadas, com uma camada oculta

Este formato, com múltiplas camadas e contendo uma ou mais camadas ocultas, foi efetivamente implementado em 1974 por Werbos, Parker e Rumelhart, recebendo o nome de "Perceptron Multi-camadas" (Haykin, 2001; Portugal e Fernandes, 1996; Rumelhart e MacClelland, 1995). Neste modelo (Equação 7), o sinal recebido pela função de ativação de um neurônio oculto j é a soma das entradas ponderadas que recebe (White, 1992, p.82).

$$a_j = \sum w_{ij}x_i = w_{0j} + w_{1j}*x_1 + w_{2j}*x_2 + ... + w_{nj}*x_n \quad \text{(Equação 7)}$$

Onde w_{0j} é um termo de *bias*, com função semelhante ao intercepto (constante) da equação de regressão, e os pesos w_{ij} representam a ponderação para a unidade j das entradas recebidas das unidades i (que são todas as unidades conectadas à unidade j). A unidade oculta j produz uma saída $y_j=f_A(a_j)=f_A(\sum w_{ij}x_i)$, onde f_A é uma função de ativação, de formato linear ($y_j=a_j$) ou da família sigmóide ($y_j=(1+e^{-aj})^{-1}$ ou $y_j=2*(1+e^{-aj})^{-1}-1$), geralmente. As unidades ocultas enviam sinais às unidades de saída da mesma forma e a saída é $Y=f_A(\sum w_{jk}x_j)$, ponderando com pesos w_{jk} as entradas recebidas pela camada de saída a partir da saída de cada neurônio j ($x_j=y_j$). Assim, a saída Y será dada pela Equação 8 (White, 1992, p.83).

$$Y = f_A\left[\sum w_{jk}\left(f_A(\sum w_{ij}x_i)\right)\right] = f_A(\sum w_{jk}x_j) \equiv f_A(X, W) \quad \text{(Equação 8)}$$

Onde W é o vetor de ponderações das conexões e X é o vetor de entradas dos neurônios. As conexões e pesos definem o comportamento da rede, que determina a resposta na unidade de saída da rede em situação de equilíbrio. A maioria das RNA tem algum tipo de mecanismo de auto-treinamento, através de ajustes das ponderações nas conexões. Assim, se diz que as redes neurais "aprendem" através de exemplos, e têm alguma capacidade de generalização, após este período. Dada uma rede suficientemente complexa, a regra de aprendizagem deve realizar a tarefa de encontrar os pesos adequados para os dados disponíveis (Haykin, 2001; White, 1992).

A forma mais comum de treinamento é através de exemplos e, segundo Kovács: "por este método, são apresentados exemplos de comportamento à rede ... os exemplos são repassados até que a rede aprenda o comportamento correto", ou seja, até que a rede implemente corretamente a função real para todos os exemplos do conjunto de dados empregado para o treinamento da rede (Kovács, 1996).

O algoritmo *backpropagation* (retropropagação dos erros), embora não seja totalmente preciso, é o mais empregado. Este algoritmo determina os pesos por tentativa e erro, num

processo iterativo de treinamento, a partir das evidências empíricas fornecidas. Geralmente a rede é iniciada com uma configuração mais complexa (mais conexões), sendo simplificada progressivamente, até atingir os resultados esperados (Haykin, 2001).

Em qualquer topologia, entretanto, os dados são apresentados como uma matriz, com um ou mais casos de exemplo, codificados numericamente. Geralmente exige-se uma grande quantidade de dados para o bom funcionamento da RNA. Os dados de entrada assemelham-se aos empregados em outras formas de análise, inclusive estatística inferencial. Geralmente cada informação é chamada de "caso" ou "exemplo".

Na geração das redes neurais, uma amostra é dividida em dois conjuntos, um para treinamento e outro para validação ou teste da rede. O conjunto de treinamento é o conjunto de exemplos usados para aprendizagem (ajustamento dos pesos) e o conjunto de validação é utilizado para verificar o desempenho da rede (capacidade de generalização). O conjunto de dados de treinamento traz para a rede informações de entrada X_t e de saída Y_t, referentes ao caso t. No treinamento através do algoritmo *backpropagation*, a rede é inicializada com um conjunto de pesos aleatórios W_0, que são (retro)ajustados progressivamente pela seguinte relação (Equação 9):

$$W_t = W_{t-1} + \eta \nabla f_A(X_t, W_{t-1}) * (Y_t - f_A(X_t, W_{t-1})) \qquad \text{(Equação 9)}$$

Onde t indica os elementos da amostra, variando de 1 a n, η é a taxa de aprendizagem e ∇f_A é o gradiente de f_A com relação aos pesos W (vetor com as derivadas parciais). Assim, os pesos são reajustados com base nos erros de resposta $[Y_t - f_A(X_t, W_{t-1})]$. A taxa de aprendizagem (η) indica a velocidade desejada para o ajuste. Se for elevada demais, pode haver instabilidade, e eventualmente não ocorrerá a convergência dos parâmetros. Geralmente η é fixa, para os dados de entrada, que é a opção mais apropriada se ocorrem variações nos casos (observações). Entretanto, se há pouca variação de uma observação para outra, pode-se fazer η variar com t (indica-se como η_t, neste caso), para compensar efeitos aleatórios sobre Y_t.

Assim, durante o treinamento, o vetor de pesos é progressivamente ajustado, buscando diminuir a diferença entre as saídas estimadas e as desejadas (valores reais, da amostra), até o nível de precisão desejado. Atingido este ponto, a rede deve ser testada, empregando-se o

conjunto de dados reservado para a validação. Porém, uma das principais dificuldades de uso das RNA é definir o momento de encerrar o treinamento. Se for ultrapassado o limiar ótimo, ocorre o "supertreinamento" (*overfitting*), quando a rede memoriza os dados, caso em que os resultados aplicam-se apenas ao conjunto empregado neste treinamento. Por este motivo, a utilização de um conjunto de dados de validação é fundamental. Contudo, no estágio atual de desenvolvimento desta técnica, o ajuste dos parâmetros, tais como número de nós ou de camadas ocultas ou tempo de processamento, deve ser realizado por tentativa e erro, com auxílio de algumas heurísticas (Haykin, 2001; Worzala *et al.*, 1995).

Há uma certa semelhança entre alguns tipos de redes neurais e a análise de regressão através de modelos generalizados. A função f_A tem a mesma finalidade geral da equação de regressão, bem como as entradas X_t são as variáveis independentes, as saídas Y_t são as variáveis dependentes e os pesos W seriam os parâmetros da equação de regressão. White (1992, p.87) entende que deve-se empregar em conjunto técnicas de inferência e redes neurais, para obter-se mais eficiência no processo de estimação.

Haykin (2001) e Watson (1997), entre outros, afirmam que a desvantagem das RNA é que as mesmas constituem um sistema do tipo "caixa preta". A resposta é uma função ponderada dos vetores, sem uma explicação ou justificativa para o resultado. Por isto, as RNA não podem ser utilizadas em muitos domínios. Na Europa, por exemplo, as instituições financeiras não podem utilizar RNA na análise de crédito, pois há exigências legais de justificativa para eventuais negativas.

Além do Perceptron Multi-camadas, que é o formato mais empregado, há outros tipos de redes neurais, tais como as redes de Hopfield e de função de base radial (*radial basis function*, RBF), e as redes auto-organizáveis, tais como as redes SOM (*self-organizing maps*) também conhecidas como mapas de Kohonen, e as redes ART (*adaptive resonance theory*), com diferentes habilidades (Braga *et al.*, 2000; Haykin, 2001).

4.9.1 Aplicações de redes neurais em avaliações

Na literatura recente podem ser encontrados alguns exemplos de aplicações de RNA na análise de valores de imóveis. Tay e Ho (1994) investigaram a aplicação de redes neurais para

avaliação em massa, usando uma amostra de 833 informações de vendas de apartamentos em Singapura para o treinamento da rede e outros 222 dados para o grupo de controle (validação). A rede foi determinada com o algoritmo de *backpropagation*, com três camadas (de entrada, oculta e de saída), obtendo resultados satisfatórios. Dos imóveis do grupo de validação, apenas 9 apresentaram erros absolutos acima de 50%, em relação ao valor real. Analisando em detalhe, estes autores verificaram que apenas um destes erros não podia ser explicado, sendo que, para os outros, os valores estimados eram até mais adequados que os originalmente coletados.

Evans *et al.* (1995) apresentaram uma aplicação de RNA de três camadas, também determinada com retropropagação dos erros, usando uma amostra de 45 dados de venda de residências, sendo 33 utilizados no treinamento da rede e 12 na validação, concluindo que as RNA parecem ser adequadas para a avaliação de imóveis, mesmo para um conjunto pequeno de dados, embora a precisão seja extremamente dependente do cuidado no tratamento dos dados do conjunto de treino. Dados "com ruído" ou "não apropriados" também afetam sensivelmente o nível de erro, de forma semelhante aos *outliers* na regressão.

Worzala *et al.* (1995), utilizando uma amostra de 288 informações de vendas de residências, divididas em 217 e 71 dados para treinamento e controle, respectivamente, desenvolveram uma rede de três camadas, testando também modelos com amostras parciais. Os resultados não foram considerados satisfatórios, pois foram encontradas diferenças significativas entre os dois *softwares* utilizados, e em repetições com o mesmo *software*. O tempo de processamento também foi considerado longo, atingindo até 25h para estabilizar a rede. Em alguns casos os modelos RNA superavam os resultados da análise de regressão, mas nenhum dos *softwares* demonstrou superioridade.

Rossini (1997) desenvolveu um estudo comparativo entre ARM e RNA, com um conjunto total de 334 informações, testando modelos com 223 elementos e outros com menor quantidade de dados, para simular uma avaliação comercial. Também empregou uma rede neural de três camadas. Seus resultados indicaram superioridade da ARM em relação ao emprego de RNA. O autor relatou duas dificuldades básicas com as RNA: o tempo elevado de processamento e a instabilidade dos resultados. Para a amostra de 223 casos e 42 variáveis o processamento superou 23h, embora o sistema tenha sido rápido para amostras pequenas, com tempos da ordem de um minuto. Os modelos estimados com análise de regressão foram

bastante consistentes, mas os resultados com as RNA variaram entre excelentes e fracos. Porém, Rossini lembra que a ARM é uma técnica antiga, enquanto que as RNA têm um grande campo de desenvolvimento, em termos de utilização e de disponibilidade de *softwares*.

Kauko (1997) apresentou um estudo usando o mapa de Kohonen, utilizando dados de imóveis da Finlândia, de todo o país e da região metropolitana de Helsinque. Para ele, a RNA é meramente uma extensão não linear da ARM, enquanto que o SOM (*self-organizing mapping*) gera um mapa, representando graficamente o fenômeno em análise. Contudo, este autor afirma que a intuição é enfatizada e que não há regras claras para o uso deste método. Por outro lado, em exaustivo levantamento de 3343 *papers* sobre SOM publicados entre 1981 e 1997, Kaski *et al.* (1998) não indicaram nenhum trabalho relacionado com o mercado imobiliário.

Nos trabalhos mais recentes, todavia, há um equilíbrio entre os modelos de regressão e de redes neurais (Cechin *et al.*, 1999, 2000; McCluskey e Borst, 1997; Nguyen e Cripps, 2001). No Brasil, em estudo pioneiro, Guedes (1995) propôs a utilização de RNA para avaliação de imóveis, apresentando um estudo comparativo com ARM. Este autor utilizou uma amostra de 102 informações de imóveis comerciais situados na cidade do Rio de Janeiro para gerar modelos de regressão e redes neurais, reservando outros 11 dados para a fase de testes. Os resultados indicaram que as redes neurais tiveram melhor desempenho, embora este autor ressalve que "a única maneira de testar a precisão de uma rede neural é pelos dados de saída", pois não apresentam uma equação formal, tal como na regressão. Cechin *et al.* (1999, 2000) desenvolveram modelos para dados de aluguel e de venda, comparando regressão linear e redes neurais em diversas topologias, obtendo maior precisão com as redes neurais.

4.10 SISTEMAS BASEADOS EM REGRAS DIFUSAS

Os sistemas baseados em regras difusas (SBRD) tem tido um grande desenvolvimento teórico e em termos de aplicações nos últimos anos, especialmente com sistemas híbridos. Os SBRD consistem em uma base de regras que utilizam a lógica difusa nas partes precedentes ou antecedentes das regras, e uma das características interessantes destes sistemas é a conjugação de efeitos de várias regras para a obtenção do resultado final (Cordón *et al.*, 2001). Antes de

discutir os sistemas de regras, propriamente ditos, é interessante identificar algumas características da lógica difusa.

4.10.1 Lógica difusa

O propósito de um conjunto convencional, no domínio da matemática, é caracterizar precisamente algum conceito. É preciso especificar um universo de discurso, o qual contém todos aqueles elementos que são relevantes para o conceito particular a ser representado (Kacprzyk, 1997). Na teoria clássica dos conjuntos, uma vez que um conjunto tenha sido definido, cada elemento de interesse está incluído ou excluído do conjunto. As expressões "pertence a" ou "é membro de" têm importância essencial na teoria dos conjuntos. Para um dado conjunto A, a expressão "x pertence a A" pode ser escrita como "$x \in A$", bem como o oposto é "$x \notin A$". Esta relação pode ser enunciada como uma função característica, na Equação 10 (Cios *et al.*, 1998; Kacprzyk, 1997):

$$\varphi_A: x \rightarrow \{0,1\} \qquad \text{ou } \varphi_A(x) = \begin{cases} 1, \text{ se } x \in A \\ 0, \text{ se } x \notin A \end{cases} \qquad \text{(Equação 10)}$$

Contudo, este é um sistema dicotômico, que não compreende todas as relações de interesse. Em muitas circunstâncias, existe uma faixa de transição (contínua) entre nula e total pertinência, e a passagem das situações "sim/não" ou "branco/preto" é demasiado abrupta, e não é relevante ou apropriada. Um exemplo deste caso é a representação de relações como "aproximadamente igual a". Segundo Kacprzyk (1997), a teoria dos conjuntos difusos é um meio simples, mas poderoso, efetivo e eficiente de representar e manusear informação imprecisa (conceitos vagos), exemplificada em expressões como "construções altas" ou "grandes números". Para Alcalá *et al.* (2000), a lógica difusa é interessante para sistemas baseados em regras pois, em muitos casos, o raciocínio humano – especialmente o senso comum – apresenta formas aproximadas, por natureza.

Uma das soluções para considerar as medidas intermediárias é a teoria dos conjuntos difusos (*fuzzy-set theory*), que recebeu contribuição fundamental de Lofti Zadeh. Para este autor, afirma que, informalmente, um conjunto difuso pode ser considerado como um tipo de conjunto no qual existe uma progressão gradual de pertinência para não pertinência ou, mais

precisamente, no qual um objeto pode ter um grau de pertinência intermediário entre a unidade (pertinência total) e zero (não pertinência). Neste caso, a função característica é substituída por uma relação de pertinência - $\mu_A(x)$ -, que considera também os valores intermediários, ou seja, inclui os casos em que o objeto pertence parcialmente a dois (ou mais) conjuntos (Zadeh, 1965). Os valores de pertinência expressam os graus nos quais cada objeto é compatível com as propriedades que caracterizam o conjunto em questão (características distintivas da coleção), ou seja, estes valores expressam o grau de associação do elemento com a característica representada pelo conjunto. Um conjunto difuso é caracterizado por uma função de pertinência $\mu_A(x)$, que assume qualquer valor no intervalo [0,1]. Fica claro que um conjunto difuso é uma generalização do conceito clássico de conjuntos, no qual a função característica assumia apenas dois valores, {0,1}. Assim, se A é um conjunto difuso, a relação é a apresentada na Equação 11 (Cios *et al.*, 1998; Kacprzyk, 1997):

$$\mu_A(x) = \begin{cases} 1, \text{ se } x \in \text{ (totalmente) A} \\ \mu_x, \text{ se x pertence parcialmente a A, com } 0<\mu_x<1 \\ 0, \text{ se } x \notin \text{(totalmente) A} \end{cases} \qquad \text{(Equação 11)}$$

Por exemplo, se existem dois conjuntos, denominados "branco" ($\mu_B(x)$) e "preto" ($\mu_P(x)$), e são apresentados elementos intermediários, com quantidade variável de branco e de preto (tons de cinza), os conjuntos difusos poderiam ser formulados, em função do percentual de pontos pretos e brancos do elemento, como na Figura 10:

X	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
$\mu_P(x)$	0,00	0,05	0,10	0,15	0,20	0,25	0,30	0,40	0,50	0,60	0,70	0,75	0,80	0,90	1,00
$\mu_B(x)$	1,00	0,95	0,90	0,85	0,80	0,75	0,70	0,60	0,50	0,40	0,30	0,25	0,20	0,10	0,00

Figura 10: Exemplo de conjuntos difusos

Um elemento qualquer, tal como x=6, terá pertinência parcial de $\mu_p(6)=0,25$ e $\mu_B(6)=0,75$, significando que está mais fortemente relacionado com B. A teoria clássica de conjuntos não permite uma resposta adequada a esta questão.

Fica claro que uma função de pertinência é subjetiva, opondo-se à forma objetiva de uma

função característica, o que é natural, devido à imprecisão que se pretende representar, que depende também do observador, ao contrário das definições precisas dos conjuntos clássicos (Kacprzyk, 1997).

Zadeh (1965) lembra que o significado associado com um valor numérico particular da função de pertinência é puramente subjetivo, por natureza, sendo necessário considerar o contexto da função, ou seja, a relação entre as medidas de pertinência de cada instância. Para Cios *et al.* (1998), a essência dos conjuntos difusos é sua habilidade para lidar com conceitos elásticos ou com a ausência de fronteiras bem definidas entre situações distintas, o que ocorre freqüentemente em situações reais.

4.10.1.1 Notação e terminologia

Em uma linguagem mais formal, pode-se dizer que um conjunto difuso geralmente está contido em um universo de discurso[33] não difuso, que pode ser uma coleção de objetos, conceitos ou construções matemáticas. Por exemplo, um universo de discurso pode ser o conjunto dos números reais, o conjunto de habitantes de uma cidade, o conjunto de objetos em uma sala, etc. Geralmente o universo é indicado pelas letras maiúsculas U, V ou W, com ou sem subscritos ou sobrescritos. Um conjunto difuso em U ou, dito de forma equivalente, um subconjunto difuso de U, é indicado pelas letras A, B, C, D, etc., igualmente com ou sem subscritos ou sobrescritos (Zadeh, 1965).

Um subconjunto difuso A de um universo de discurso U é caracterizado por uma função de pertinência (*membership function*) μ_A: U $\rightarrow$ [0,1], que associa um número real $\mu_A(u)$ no intervalo [0,1] a cada elemento u de U, com o valor $\mu_A(u)$ representando o grau de pertinência (grau de associação) de u com o conjunto difuso A[34]. O suporte de A é o conjunto de pontos em U para os quais $\mu_A(u)$ é positivo. A altura de A é o maior valor de $\mu_A(u)$ em A. O ponto de cruzamento de A é o ponto em U cujo grau de pertinência em A é igual a 0,5. A é normal se

[33] O domínio de discurso é domínio da função A(x), que corresponde ao trecho no qual a função A está definida, ou seja, o conjunto de valores possíveis para X que geram A(x).

[34] Alguns autores, como Cios *et al.* (1998) e Nguyen e Walker (2000), usam A: X $\rightarrow$[0,1], ou A(x). Chao e Skibniewski (1998) indicam de forma simplificada, como f: x$\rightarrow$y.

sua altura é 1 e subnormal se é diferente de 1. Quando $\mu_A(u)$ pode ser interpretado como grau de compatibilidade ou de possibilidade de u dado A, a função μ_A: U $\rightarrow$ [0,1] pode ser chamada de função de compatibilidade (Nguyen e Walker, 2000; Zadeh, 1965). Já Kacprzyk (1997) indica a função de pertinência no formato de pares ordenados: $A=\{(x, \mu_A(x))|x \in X\}$.

Alternativamente, pode ser empregada também uma notação simplificada, como segue. Um conjunto finito não difuso $U=\{u_1, u_2, ..., u_n\}$ pode ser expresso como $U=u_1+u_2+...+u_n$, com "+" funcionando como união, e não como soma aritmética. Um subconjunto difuso finito A é expresso como uma forma linear: $A=\mu_1u_1+\mu_2u_2+...\mu_nu_n$. Quando u_i é numérico, usa-se a notação μ_i/u_i para evitar ambigüidades: $A=\mu_1/u_1+\mu_2/u_2+...\mu_n/u_n$, ou ainda $A=\{\mu_1(u)/u_1; \mu_2(u)/u_2;...;\mu_n(u)/u_n\}$. Se A for infinito, utiliza-se uma integral definida, ao invés do somatório discreto (Cordón *et al.*, 2001; Kacprzyk, 1997; Zadeh, 1965).

4.10.1.2 Tipos de função de pertinência

Para Cios *et al.* (1998), em princípio qualquer função contínua pode ser utilizada para descrever uma função de pertinência associada com um conjunto difuso. Algumas das funções de pertinência mais comuns são as apresentadas a seguir (Equações 12 a 15).

a) Função de pertinência triangular (Equação 12):

$$A(x) = \begin{cases} 0, \text{ se } x <= a \\ (x-a)/(m-a), \text{ se } x \in (a,m) \\ (b-x)/(b-m), \text{ se } x \in [m,b) \\ 0, \text{ se } x >= b \end{cases} \qquad \text{(Equação 12)}$$

b) Função de pertinência S, na qual (a+b)/2 é o ponto de cruzamento da função (Equação 13):

$$A(x) = \begin{cases} 0, \text{ se } x<=a \\ 2.[(x-a)/(b-a)]^2, \text{ se } x \in [a,m] \\ 1-2.[(x-b)/(b-a)]^2, \text{ se } x \in [m,b] \\ 1, \text{ se } x>b \end{cases} \qquad \text{(Equação 13)}$$

c) Função de pertinência trapezoidal (Equação 14):

$$A(x) = \begin{cases} 0, \text{ se } x<a \\ (x-a)/(m-a), \text{ se } x \in [a,m] \\ 1, \text{ se } x \in [m,n] \\ (b-x)/(b-n), \text{ se } x \in [n,b] \\ 0, \text{ se } x>b \end{cases} \qquad \text{(Equação 14)}$$

d) Função de pertinência gaussiana (Equação 15):

$$A(x) = \begin{cases} e^{[-k.(x-m)^2]} \text{ , onde } k>0 \\ \\ \text{ou} \\ \\ e^{[-(x-m)2/k2]} \end{cases} \qquad \text{(Equação 15)}$$

4.10.1.3 Probabilidade e conjuntos difusos

Embora freqüentemente confundidos, probabilidade e conjuntos difusos são conceitos distintos e existem várias diferenças entre eles. A probabilidade é relacionada com a ocorrência de eventos bem definidos, tal como o resultado da extração de um número em uma urna. Podem ser associados níveis de probabilidade a cada um dos eventos. Por outro lado, os conjuntos difusos lidam com problemas de graduação dos conjuntos e descrevem suas fronteiras. Não estão ligados com a freqüência dos eventos, mas com o valor (grau) de associação dos conceitos. Inclusive, pode-se falar em "probabilidade de eventos difusos" (Cios *et al.*, 1998, p.121). Para Zadeh (1965), existem diferentes categorias de imprecisão, e está claro que:

a) a difusibilidade é fundamentalmente diferente da aleatoriedade;

b) a difusibilidade desempenha um papel mais importante na cognição humana do que a aleatoriedade;

c) para considerar efetivamente a difusibilidade é necessário adotar uma postura distinta, que este autor indica como a teoria dos conjuntos difusos.

Segundo Zadeh (1965), para o entendimento das diferenças entre a difusibilidade e a aleatoriedade, é útil interpretar o grau de pertinência em um conjunto difuso como o grau de compatibilidade (ou possibilidade) ao invés de probabilidade.

4.10.2 Sistemas de regras difusas

Uma das aplicações mais importantes da teoria dos conjuntos difusos são os sistemas baseados em regras difusas (SBRD), que são uma extensão dos sistemas baseados em regras clássicas, usando regras difusas de forma a habilitar estes sistemas para aplicações em áreas que apresentam incerteza ou imprecisão (Alcalá *et al.*,2000; Cordón e Herrera, 1999; Cordón *et al*., 2001).

Um SBRD tem dois componentes: (a) o sistema de inferência, que realiza o processo de inferência necessário para gerar uma saída quando uma determinada entrada é especificada, e (b) a base de conhecimento, representando o conhecimento sobre o problema a ser resolvido, constituída de uma coleção de regras, as quais geralmente são obtidas automaticamente a partir de uma coleção de casos (exemplos), em função do conhecido problema de "engarrafamento", apresentado pelos sistemas baseados em conhecimento. Existem vários algoritmos de aprendizagem, tais como métodos *ad hoc*, redes neurais, algoritmos genéticos e *clustering* (Alcalá *et al.*,2000). Uma regra difusa tem o seguinte formato básico (Equação 16):

$$R_i: \text{SE X É } A_i \text{ ENTÃO Y é B} \qquad \text{(Equação 16)}$$

Onde a parte "SE X é A_i" é chamada de antecedente, e a parte "ENTÃO Y é B" é chamada de conseqüente da regra *i*. Um sistema de regras é composto por um conjunto de regras de

formato semelhante. Existem basicamente dois tipos de SBRD: Mamdani e TSK. O primeiro considera variáveis lingüísticas na parte conseqüente, enquanto o tipo TSK utiliza como resultado uma função das entradas (Alcalá *et al.*, 2000).

4.10.2.1 Sistema de regras tipo Mamdani

O primeiro tipo foi proposto por Mamdani e é o mais empregado. Também é conhecido como "sistema de regras difusas com fuzzificador e defuzzificador" ou como "controlador de lógica difusa", ou ainda como "Mamdani lingüístico ou descritivo". Estes sistemas têm quatro componentes: base de conhecimento, sistema de inferência e interfaces de fuzzificação e defuzzificação. A base de conhecimento é composta de uma base de dados (DB) e de uma base de regras, armazenando o conhecimento disponível na forma de regras lingüísticas "SE-ENTÃO". O formato geral é o apresentado na Figura 11 (Cordón *et al.*,2001).

Este tipo de sistema tem algumas características importantes. Pode ser utilizado em aplicações com atributos contínuos, devido ao fato de que lida com números reais nas entradas e saídas. Por outro lado, propicia um ambiente natural para incluir conhecimento especializado na forma de regras lingüísticas e permite combiná-las facilmente com regras geradas semi-automaticamente a partir dos dados de exemplo. Finalmente, tem mais liberdade na escolha dos componentes das interfaces de fuzzificação e defuzzificação, bem como o sistema de inferência, facilitando o ajustamento do sistema de regras ao domínio do problema (Alcalá *et al.*,2000; Cordón *et al*., 2001).

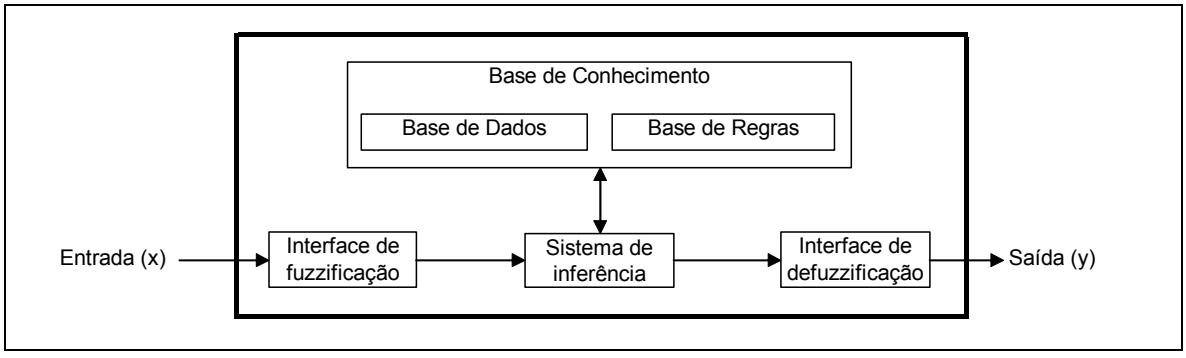


Figura 11: Estrutura de um sistema de regras tipo Mamdani (adaptado de Cordón *et al.*, 2001).

O mecanismo de fuzzificação estabelece um relacionamento entre os valores no domínio de entrada e os conjuntos difusos definidos no mesmo universo de discurso. Para gerar o resultado, após a inferência, o mecanismo de defuzzificação faz o caminho inverso.

Chao e Skibniewski (1998, p.298) exemplificam com um sistema para determinação do tamanho de um bate-estacas. Considerando a altura da estaca e o tipo de solo, uma regra típica seria: "SE o pilar é longo E o solo é duro, ENTÃO use um bate-estacas pesado". As palavras sublinhadas são instâncias dos valores lingüísticos das variáveis difusas. Outra forma de apresentação seria: "[longo, duro]⇒ [pesado]". Um conjunto de regras de mesmo formato, mas com instâncias variáveis, pode ser estabelecido para diferentes condições. Os dados iniciais (entrada) e os finais (saída do sistema) são valores próprios do domínio da aplicação. Por exemplo, o comprimento da estaca em metros ou o peso da máquina em toneladas. Estes valores são as quantidades de suporte nos conjuntos difusos. O processo de obtenção da solução envolve três passos (Chao e Skibniewski, 1998; Cordón *et al.*, 2001):

a) fuzzificar as entradas através das funções de pertinência correspondentes às pré-condições das regras difusas;

b) calcular os valores de disparo das regras (condições-limite) para definir as ponderações das conseqüências de cada regra;

c) defuzzificar as conseqüências agregadas para produzir as saídas.

Dada uma entrada x, existem geralmente duas ou mais funções de pertinência, de uma família de funções difusas, que retornam um valor de pertinência maior do que zero, devido à sobreposição destas funções. As regras que usam estes dados têm valores de disparo diferentes, com conseqüências ponderadas de acordo. A união das conseqüências é calculada para determinar o resultado final ponderado (Cordón *et al.*, 2001).

Cada regra da base de regras é a descrição de uma sentença do tipo "condição-ação" que pode ser claramente interpretada por seres humanos. Uma variação, chamada de DNF (formato normal disjuntivo), trabalha com conjuntos finitos para entradas e saídas. Neste caso, uma regra possível seria a da Equação 17 (Alcalá *et al.*,2000):

R_i: SE X_1 é $\{A_{11}$ ou... ou $A_{1k}\}$ e... e X_n é $\{A_{n1}$ ou... ou $A_{nk}\}$ ENTÃO Y_i é B (Equação 17)

Onde X_j são as variáveis de entrada, A_j e B são conjuntos de variáveis lingüísticas e Y_i é a variável de saída (resultado da regra R_i). Outra alteração, mais recente, adicionando a possibilidade de se ter dois componentes no lado conseqüente, como limites do espaço de solução é apresentada na Equação 18 (Alcalá *et al.*,2000):

R_i: SE X_1 é A_1 e ... e X_n é A_n ENTÃO Y_i está entre B_1 e B_2 (Equação 18)

Uma outra variante, chamada de "Mamdani aproximado", aumenta a precisão à custa da interpretabilidade. A única diferença é que este tipo de regra utiliza diretamente variáveis difusas, e não variáveis lingüísticas, como na Equação 19 (Alcalá *et al.*,2000):

R_i: SE X_1 é A_1 e ... e X_n é A_n ENTÃO Y_i é B (Equação 19)

Onde A_j e B são conjuntos difusos. Não há base de dados, mas somente um conjunto de regras, neste caso.

Após a definição do formato das regras, um dos assuntos mais importantes nos sistemas Mamdani é a definição da fase de defuzzificação, ou seja, a como será realizada a seleção de um elemento de resposta, baseado no conjunto difuso de saída.

Em geral, a saída é um subconjunto difuso, cujos elementos são pares ordenados do tipo μ_1/u_1, onde o vetor u_1 é uma possível decisão final, baseada nos dados de entrada fornecidos ao sistema, e μ_1 é o grau de pertinência de u_1, que determina quão boa é esta solução. A defuzzificação é a operação do tipo $D:[0,1]^u \rightarrow U$, que gera um valor y=a, a ser dado como resposta (saída) do sistema (Nguyen e Walker, 2000). O objetivo de transformar elementos difusos em representações numéricas é permitir que os conjuntos difusos interajam com ambientes numéricos. O processo de raciocínio é abstrato, mas o resultado final deve ser

numérico. Estes mecanismos são freqüentemente referidos como mecanismos de defuzzificação (Cios *et al.*, 1998). Para Alcalá *et al.* (2000), o formato é: $\mu_B(x)=G\{\mu_{B1}(x), \mu_{B2}(x),..., \mu_{Bn}(x)\}$, onde G é um operador de agregação.

Existem vários métodos para o cálculo da solução final, tais como a média dos máximos, o centro de área e o centro de gravidade (Cios *et al.*, 1998; Kacprzyk, 1997). A escolha mais comum é o centro de gravidade, calculado como na Equação 20 (Alcalá *et al.*,2000; Cordón *et al.*, 2001).

$$Y = \frac{\sum_{i=1}^{m} h_i y_i}{\sum_{i=1}^{m} h_i} \qquad \text{(Equação 20)}$$

Onde h_i representa o grau de pertinência do valor de entrada x_i com a parte antecedente da regra i, e y_i é a saída da mesma regra (Alcalá *et al.*,2000; Cordón *et al.*, 2001). Além destes métodos básicos, existem várias outras modificações, tal como a proposta por Kandal e Friedman (1998), baseada no valor mais típico (*most tipical value, MTV*).

4.10.2.2 Sistema de regras tipo TSK

O outro tipo de sistema é conhecido como modelo difuso de Sugeno, ou modelo TSK, iniciais dos autores que desenvolveram este formato (Takagi, Sugeno e Kang). O modelo é composto de antecedente representado por variáveis lingüísticas, sendo o conseqüente uma função das variáveis de entrada, geralmente uma função linear (Cordón *et al.*, 2001), tal como a da Equação 21:

$$R_i: \text{ SE } X_1 \text{ é } A_1 \text{ e } ... \text{ e } X_n \text{ é } A_n \text{ ENTÃO } Y_i = f(X_1, ..., X_n) \qquad \text{(Equação 21)}$$

Ou então como na Equação 22:

$$R_i: \text{ SE } X_1 \text{ é } A_1 \text{ e } ... \text{ e } X_n \text{ é } A_n \text{ ENTÃO } Y_i = p_0 + p_1X_1 + ... + p_nX_n \qquad \text{(Equação 22)}$$

Onde X_j são as variáveis de entrada, A_j são conjuntos difusos especificando seu significado, Y_i é o resultado da regra, e os parâmetros p_k são números reais (Cordón e Herrera, 1999; Cordón *et al.*, 2001). Os A_j podem ser variáveis lingüísticas associadas com conjuntos difusos ou variáveis difusas, diretamente. Estas regras são chamadas de "regras difusas TSK". O resultado, usando uma base de conhecimento e considerando um conjunto de regras, é obtido por uma média ponderada das regras do sistema, através do método do centro de gravidade (ver Equação 20).

A maior vantagem dos sistemas TSK é de que consistem em um conjunto compacto de equações, sendo que os parâmetros p_k são estimados por qualquer processo, tais como regressão múltipla, algoritmos genéticos ou mesmo pela indicação de especialistas. A estrutura básica do sistema é a seguinte - Figura 12 - (Cordón *et al.*,2001):

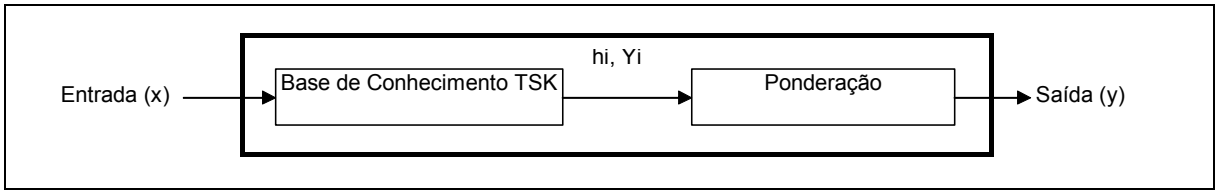


Figura 12: Estrutura de um sistema de regras tipo TSK (adaptado de Cordón *et al.*, 2001).

Os sistemas com regras TSK são baseados em informação numérica, e são especialmente úteis quando não existe ou existe pouco conhecimento disponível. De qualquer forma, o conhecimento especializado é útil para definir os rótulos das variáveis lingüísticas (*labels*), especificar seu significado ou identificar os possíveis estados do sistema (combinações de entradas e saídas). Os conseqüentes nas regras TSK (as equações lineares) geralmente são gerados por meio de um processo automatizado.

O problema é que nem sempre este formato é adequado para representar conhecimento especializado (introduzido diretamente pelo especialista). Entretanto, o conhecimento pode ser introduzido através de uma pequena modificação: quando uma regra é fornecida pelo especialista, com o conseqüente na forma "Y é B", pode-se entender como "$Y=p_0$". Este tipo

de regra é conhecido como “regra TSK simplificada” ou “regra TSK de ordem zero” (Alcalá *et al.*,2000).

4.10.2.3 Desenvolvimento de um SBRD

Basicamente, o desenvolvimento do sistema consiste em encontrar um conjunto finito de regras difusas que estejam aptas a reproduzir o comportamento de entrada e saída do sistema, com base apenas em um conjunto de dados, composto por de vetores de entrada e saída. A modelagem é realizada geralmente em duas etapas. A primeira consiste na geração de uma primeira aproximação para o modelo difuso que descreve o sistema, incluindo a determinação do número de regras necessárias. O resultado desta etapa é uma coleção de regras que pode ser vista como uma estimativa da base final de regras. O segundo passo consiste em sintonizar o conjunto inicial de regras, obtendo a base final (Cordón *et al.* 2001; Gómez-Skarmeta e Jiménez, 1997). Para Cordón *et al.* (2001, p.23), as questões principais na definição de um SBRD são as seguintes:

a) escolha das variáveis relevantes;

b) definição das descrições lingüísticas, incluindo os fatores de escala, a escolha dos termos (rótulos) e a escolha do tipo ou formato das funções de pertinência;

c) derivação das regras aproximadas, incluindo a definição do número de regras e de sua composição (conhecimento especializado) e posterior refinamento, usando algoritmos genéticos ou outro mecanismo.

Há várias formas de obter a base de regras, geralmente contando com o apoio de algoritmos genéticos. Por exemplo, Cordón e Herrera (1999) apresentaram um processo evolucionário em dois estágios para gerar sistemas TSK diretamente dos exemplos, combinando um estágio de geração, no qual as regras com diferentes conseqüentes competem entre elas para formar uma base de conhecimento preliminar, e um estágio de refinamento, no qual as partes antecedente e conseqüente desta base de conhecimento são adaptadas por um processo evolucionário híbrido, composto por um algoritmo genético e uma estratégia evolucionária, obtendo-se a base final. Gómez-Skarmeta e Jiménez (1997) utilizaram clusterização difusa e

algoritmos genéticos para gerar o conjunto inicial de regras e algoritmos genéticos para sintonizar as regras. Também propuseram um algoritmo genético para gerar e sintonizar as regras difusas diretamente dos dados. Já Herrera *et al.* (1995) apresentaram um sistema totalmente gerado por algoritmos genéticos.

4.10.3 Aplicações de sistemas de lógica difusa

A lógica difusa tem emprego freqüente em sistemas de controle e em sistemas de regras difusas. Geralmente é aplicada em conjunto com outras técnicas, tais como redes neurais, raciocínio baseado em caso e algoritmos genéticos, em sistemas híbridos, para proporcionar a aprendizagem da base de regras ou sintonizar o sistema (Bonissone *et al.*, 1999; Cordón *et al.*, 2001). Neste sentido, Bonissone *et al.* (1998, 1999) descrevem diversas alternativas de combinação entre essas técnicas, incluindo uma aplicação no mercado imobiliário. O sistema de avaliações apresentado por esses autores têm dois componentes. No primeiro, a lógica difusa auxilia na definição da arquitetura de redes neurais. No segundo componente, é utilizada na identificação de casos similares para um sub-sistema de raciocínio baseado em casos.

Byrne (1995) propôs a utilização de lógica difusa para levar em conta os elementos de risco e incerteza presentes nas avaliações. Bagnoli e Smith (1998) empregaram lógica difusa para considerar as imprecisões constantes nas medidas dos imóveis, através da fuzzificação das variáveis decorrentes de julgamentos dos avaliadores, numa tentativa de reproduzir o mecanismo de decisão dos agentes, que contam com informação imprecisa.

Em outras áreas da Engenharia Civil, há diversas aplicações. Por exemplo, Leu *et al.* (1999) desenvolveram um sistema para planejamento de obras, utilizando lógica difusa para considerar as incertezas na duração das atividades. A otimização do planejamento em função das restrições de recursos foi realizada com algoritmos genéticos. Karray *et al.* (2000) utilizaram lógica difusa para o planejamento de canteiro de obras, também em um sistema híbrido com algoritmos genéticos. Soh e J. Yang (1996) e Y. Yang e Soh (2000) apresentaram dois sistemas para cálculo de estruturas metálicas, sendo que o primeiro utiliza algoritmos genéticos e o segundo aplica programação genética, ambos incorporando conhecimento especializado no processo de otimização das estruturas.

Em uma abordagem diferente, Chao e Skibniewski (1998) descreveram um sistema para avaliação de alternativas tecnológicas para a construção, baseado na lógica difusa. O exemplo apresentado é sobre métodos alternativos para operação de formas em prédios altos. A motivação para o estudo é que as técnicas formais de decisão, tais como as baseadas na teoria da utilidade, podem ser substituídas com vantagem por uma abordagem de lógica difusa. O *paper* mostra uma abordagem de avaliação de novas tecnologias de construção, de forma a embasar decisões consistentes.

4.11 SISTEMAS HÍBRIDOS

Os sistemas híbridos consistem de aplicações com dois ou mais sistemas, e têm como finalidade compensar as deficiências de cada ferramenta. Nem sempre o sistema resultante tem desempenho melhor do que os sistemas isolados, pois há geralmente alguma perda na conversão e integração das técnicas. Existem várias combinações possíveis, mas as formas de aplicação mais comum envolvem a combinação de regras difusas e algoritmos genéticos e de redes neurais e regras difusas (Braga *et al.*, 2000; Cordón *et al.* 2001).

4.11.1 Sistemas de regras difusas e algoritmos genéticos

A principal motivação para a busca de sistemas híbridos é que os sistemas baseados em regras difusas (SBRD) não são capazes de aprender sozinhos e requerem que a base de conhecimento seja desenvolvida com base em conhecimento especializado ou deduzida a partir dos dados disponíveis. Uma das alternativas mais empregadas é a dedução de um conjunto de regras a partir de um conjunto de exemplos, utilizando redes neurais ou algoritmos genéticos. É comum o uso de sistemas híbridos com soluções evolucionárias para aprender ou sintonizar a base de conhecimento ou a base de regras, sendo este sistema denominado então como um sistema baseado em regras difusas evolucionário (SBRDE).

Os algoritmos genéticos são capazes de explorar um largo espaço de prováveis soluções na busca pelas melhores, inclusive partindo de conhecimento disponível *a priori*. Nos SBRD, o conhecimento disponível *a priori* pode representar variáveis lingüísticas, parâmetros das

funções de pertinência, regras difusas ou mesmo o número de regras a ser obtido. Uma das vantagens dos sistemas de regras é de que pode abrigar regras de diferentes fontes, incluindo as obtidas por sistemas automatizados e as apontadas por especialistas na mesma base, facilitando a atualização ou aperfeiçoamento da base (Cordón *et al.*, 2001).

Os SBRD funcionam através de um raciocínio interpolativo, o qual é conseqüência da cooperação entre as regras que compõem a base de conhecimento, e geralmente mais de uma regra participa da solução apontada pelo sistema, cada uma ponderada por um grau de ajustamento ou pertinência em relação ao caso em análise, ou seja, uma ou mais regras respondem a uma mesma entrada. Já os algoritmos genéticos têm como principal característica a competição entre soluções candidatas. O ajuste de um SBRDE depende da ponderação destes dois aspectos, e o equilíbrio entre os dois é conhecido como "problema da cooperação versus competição" (Cordón *et al.*, 2001).

Existem basicamente duas abordagens para a construção da base de regras. A abordagem Pittsburgh consiste da evolução de uma base de regras inteira simultaneamente (cada indivíduo é um conjunto inteiro de regras), avaliando implicitamente a cooperação entre as regras. As funções de pertinência podem ser sintonizadas em um esquema hierárquico, envolvendo em um primeiro momento a definição dos fatores de escala, em uma sintonia ampla ou global, e em seguida, a sintonia das funções de pertinência em si. Neste formato, o conjunto de regras é evoluído conjuntamente, explorando a cooperação simultânea das regras de cada conjunto. Entretanto, o aumento do número de regras provoca forte aumento no custo computacional, por isto há limitação a um número pequeno de regras, na prática. A outra forma é a abordagem Michigan, que consiste na geração de um conjunto de regras com base na competição entre regras similares na primeira fase, e na cooperação do conjunto de regras na segunda fase, de refinamento (Cordón *et al.*, 2001).

4.11.2 Redes neurais e regras difusas

A outra forma comum de hibridização é a extração de regras a partir de redes neurais treinadas. O maior problema com as redes neurais é que o conhecimento coletado dos dados é armazenado apenas nos pesos da rede, e não tem significância direta para o analista, ao contrário de um conjunto de regras ou dos coeficientes da regressão, por exemplo. Por este

motivo, freqüentemente as redes neurais são chamadas de "*black boxes*". Existe um grande esforço de pesquisa em busca de métodos para explicar as redes neurais, e hoje existem várias formas de interpretar uma rede, usando simulação nas entradas e examinando o comportamento das saídas ou extraindo regras difusas após o treinamento. Por estes métodos é possível compreender, ao menos parcialmente, o comportamento da rede. Por outro lado, a explicação da rede ajuda a aperfeiçoar a própria rede, identificando a relevância dos neurônios ocultos ou de entrada, por exemplo.

Atualmente uma das abordagens mais interessantes é a extração de um conjunto de regras difusas diretamente dos pesos de uma rede treinada. Alguns estudos demonstraram que, sob algumas condições, uma rede neural pode ser aproximada por um sistema difuso, com o grau de precisão desejado, e vice-versa (Benitez *et al.*, 1997).

Redes neuro-difusas têm sido aplicadas com sucesso na extração de conhecimento dos dados na forma de regras difusas, explorando as melhores propriedades das redes neurais e dos sistemas difusos. A fusão destes sistemas é uma opção natural no paradigma da inteligência computacional (*soft computing*), o qual é justamente baseado na sinergia das técnicas em sistemas híbridos (Bonissone *et al.*, 1999).

Há vários métodos de extração de regras a partir de redes treinadas, mas aparentemente poucos podem ser aplicados em avaliação de imóveis, em vista das características dos dados do mercado imobiliário. As restrições encontradas em alguns métodos envolvem a exigência de saídas binárias, entradas com valores discretos e a necessidade de transformações na rede ou limitações a alguns formatos previamente definidos para a arquitetura das redes. Em outros métodos as variáveis de entrada contínuas precisam ser convertidas inicialmente para variáveis difusas, usando um conjunto de termos lingüísticos para converter as entradas em um conjunto de variáveis binárias, com o risco de introduzir tendências nos dados. Ademais, a maioria destes métodos é adequada apenas para tarefas de classificação (Arbatli e Akin, 1997; Bonissone *et al.*, 1998; Cordón *et al.*, 2001; Huang e Xing, 2002; Ishikawa, 2000; Maire, 1999; Setiono, 1997 e 2000; Setiono *et al.*, 1998).

Existem dois métodos que trabalham diretamente com os pesos da rede, sem exigências quanto aos algoritmos de treinamento ou sobre a arquitetura da rede, e que são adequados para estimação de variáveis contínuas, que são os métodos propostos em Benitez *et al.*, (1997)

e estendido em Castro *et al.*, (2002) e FAGNIS (*Fuzzy Automatically Generated Neural Inferred System*), desenvolvido por Cechin (1998).

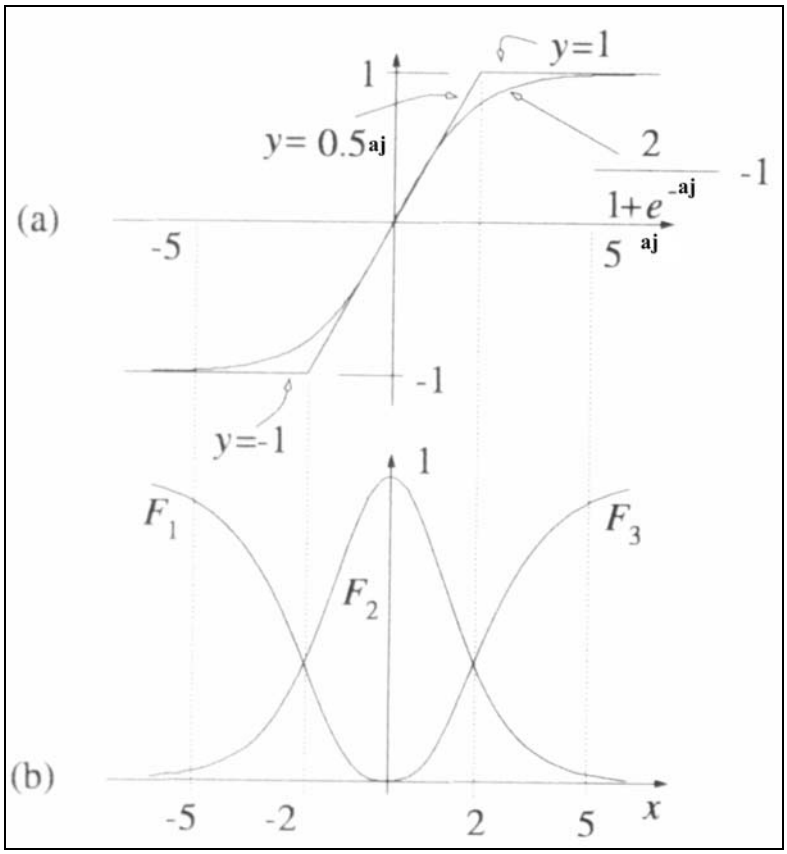
O primeiro é aplicável em avaliações, mas as regras do sistema gerado têm precedentes complexos e a explicação para o usuário não é significativamente melhorada. As regras têm um conector especial introduzido pelos autores (denominado "*interactive-or*"), o qual é um elemento adicional de complexidade (Benítez *et al.*, 1997; Castro *et al.*, 2002).

Por outro lado, o sistema FAGNIS tem potencial para atingir os dois objetivos desejados: razoável precisão do sistema de inferência e simplicidade das regras. Uma breve explanação sobre o sistema é apresentada a seguir, e detalhes podem ser obtidos em Cechin (1998).

A principal dificuldade para compreensão das redes neurais é que as funções de ativação dos neurônios são funções não-lineares, impedindo a composição do conjunto de funções em uma única função. Porém, em virtude das peculiaridades dos dados, geralmente as funções de ativação (f_A) trabalham em uma faixa estreita da possível faixa de variação das entradas. FAGNIS é baseado em uma idéia simples: substituir estas funções de ativação por um conjunto de segmentos lineares (trechos de reta), o que permite simplificar a rede.

O valor de f_A pode ser aproximado por um conjunto de segmentos lineares: $f_A(a_j) \sim \Sigma_i [F_i(a_j)*(p_i*a_j+q_i)]$, onde $f_A(a_j)$ é a função não-linear original, a_j é o sinal de ativação (soma ponderada do vetor de entrada), $F_i(a_j)$ é uma função que relaciona cada valor de a_j ao(s) correspondente(s) segmento(s) linear(es), e $p_i*a_j+q_i$ são o(s) segmento(s) linear(es). Para reduzir as perdas na conversão, $F_i(a_j)$ deve ser um conjunto difuso (Cechin, 1998).

A função de ativação de cada neurônio não-linear é substituída por um sistema de inferência difusa (*Fuzzy Inference System* - FIS) composto por uma regra TSK, na qual a função de pertinência $F_i(a_j)$ é o precedente da regra e o segmento linear $y_i= p_i*a_j+q_i$ é o conseqüente. O analista define o número de segmentos de forma a atingir o nível de erro desejado. Os conjuntos difusos podem ser constantes (definidos igualmente para todas as regras e neurônios) ou variáveis (diferentes entre eles). Por exemplo, se o neurônio tem uma função de ativação sigmóide (tal como $y_j=f_A(a_j)=2*(1+e^{-aj})^{-1}-1$) e o analista deseja utilizar três conjuntos difusos constantes, o FIS pode ser como o apresentado na Figura 13 (Cechin, 1998, p.59).



(a) Função sigmóide e a parte conseqüente das regras; (b) Função de pertinência ideal para cada região da unidade sigmóide

Figura 13: Funções de pertinência para extração de regras (adaptado de Cechin, 1998, p.59)

Uma restrição utilizada para definir as funções de pertinência é que apenas duas funções podem ter pertinência maior do que zero simultaneamente. Assim, para o sistema apresentado na Figura 13, as funções de pertinência podem ter as expressões apresentadas na Equação 23 (Cechin, 1998).

$$F_1(a_j) = \begin{cases} -f_A(a_j) + a_j * f_A'(a_j), a_j < 0 \\ 0, a_j \geq 0 \end{cases} \qquad \text{(Equação 23)}$$

$F_2(a_j) = \{ 2 * f_A'(a_j), \ \forall \ a_j$

$$F_3(a_j) = \begin{cases} f_A(a_j) - a_j * f_A'(a_j), a_j > 0 \\ 0, a_j \leq 0 \end{cases}$$

Os segmentos lineares na parte conseqüente dos FIS são calculados por meio das funções de ativação e suas derivadas no ponto desejado: $p_i = f_A'(a_x)$ e $q_i = f_A(a_x) - p_i.a_x$, onde a_x é escolhido de forma a ser o centro ou a média do intervalo de variação do nível de ativação de cada neurônio. As funções de pertinência associadas a cada segmento linear são relacionadas com a função de ativação: $F_i(a_j) = f_A(a_j)/(p_i * a_j + q_i)$, isto é, o valor de pertinência é o grau de aproximação entre as funções linear e não-linear (Cechin, 1998).

No caso extremo pode existir até uma regra para cada caso da base de treinamento, mas geralmente existem poucas regras por neurônio, em função da variação do sinal de ativação. Mesmo em redes com diversos neurônios não-lineares pode ser gerada uma única função para toda a rede, fazendo a composição entre os neurônios (em paralelo ou em série), pela regra soma-produto. Segundo esta regra, as funções de pertinência são multiplicadas ($G_r(a_j) = F_1(a_j) * F_2(a_j) * ... * F_n(a_j)$) e os segmentos lineares são somados ($y_r = (p_1 + p_2 + ... + p_n).a_j + (q_1 + q_2 + ... + q_n)$).

Finalmente, se existe mais do que uma regra no sistema, as regras aplicáveis devem ser ponderadas pelo grau de ajustamento, ou seja, o valor de pertinência do vetor de entrada em cada regra, usando $Y = (\Sigma_r G_r(X).y_r)/\Sigma_r G_r$, onde Y é a saída do sistema, G_r é a função de pertinência para a regra r, X é o vetor de entrada ($X = \{x_1, x_2, ... x_k\}$), e y_r é a saída da regra r.

4.12 CONSIDERAÇÕES FINAIS

A revisão das características de cada uma das técnicas permite identificar aquelas que podem contribuir mais para uma aplicação no âmbito do mercado imobiliário. As características das técnicas apresentadas estão resumidas na Tabela 1, a seguir.

Tabela 1: Resumo das características das técnicas apresentadas

Técnica	Descrição

Análise de regressão	Gera modelos dos relacionamentos em forma de equações. É flexível e os modelos são facilmente determinados e interpretados, mas é sensível aos dados e exige o conhecimento prévio dos modelos.
Análise de clusterização	Identifica grupos de dados homogêneos. Aceita qualquer tipo de dado numérico e é fácil de aplicar. Porém os grupos gerados podem ser difíceis de interpretar. Exige a definição do número de *clusters* e é sensível à função de distância.
Regras de associação	Indica associações entre os atributos em forma de regras. Os resultados são facilmente interpretados se o número de regras não é grande, mas há dificuldade com itens raros ou imprecisão nos dados.
Árvores de decisão	Indica associações entre os atributos em forma de árvores, as quais geralmente são de fácil interpretação. A estrutura da árvore indica claramente a importância dos atributos. Porém têm dificuldades com variáveis-resposta contínuas e a construção torna-se lenta para problemas com muitos atributos.
Análise fatorial	Gera uma combinação linear dos dados, produzindo fatores não correlacionados. É flexível e de fácil aplicação, mas pode ser difícil de interpretar e exige a definição do número de fatores.
Racicínio baseado em casos	Usa soluções anteriores (experiência) para identificar soluções para novos casos. Os sistemas são compostos por uma base de conhecimento (casos) e mecanismos de organização, seleção e adaptação destes casos. Aceitam dados simbólicos, as soluções são de fácil entendimento e a atualização do sistema é simples. Porém, o sistema torna-se lento para grandes bases e o desempenho depende da função de similaridade.
Algoritmos evolucionários	São mecanismos de busca ou otimização inspirados na genética natural, baseados na sobrevivência das melhores soluções. A seleção é baseada em uma função que mede a aptidão das soluções e a busca utiliza também mecanismos de reprodução e mutação. Os algoritmos evolucionários são robustos e podem ser integrados com outras técnicas. Contudo, podem resultar em soluções que não são as melhores possíveis (ótimos locais).
Modelos aditivos generalizados	Gera modelos em formato de equação, a partir da agregação de funções independentes. São flexíveis, mas os modelos estão vinculados aos formatos escolhidos para as funções.

Tabela 1: Resumo das características das técnicas apresentadas
(continuação)

Redes neurais artificiais	São inspiradas no funcionamento do cérebro humano, com processamento paralelo e distribuído. São flexíveis e não exigem conhecimentos prévios sobre os relacionamentos entre os dados. Entretanto, a arquitetura da rede e os valores dos parâmetros de treinamento precisam ser definidos pelo analista.
Sistemas de regras difusas	Indicam associações entre os dados em forma de regras, usando a lógica difusa para determinar as soluções. São compostos por um mecanismo de inferência e uma base de conhecimento (conjunto de regras). São flexíveis

	e aceitam problemas com incerteza ou imprecisão. Por outro lado, exigem o apoio de outra técnica na geração das regras.
Sistemas híbridos	São baseados na sinergia de duas ou mais técnicas. O desempenho depende das características dos componentes. Para os dois sistemas apresentados: a) Sistemas baseados em regras difusas evolucionários (SBRDE): os algoritmos genéticos são usados para gerar a base de regras, mas existe um aumento de complexidade e de tempo de processamento. b) Extração de regras difusas a partir das redes neurais: a lógica difusa é utilizada para extrair regras que expliquem o funcionamento das RNA, mas igualmente há aumento de complexidade.

As duas tarefas básicas a serem desenvolvidas são a preparação da base de dados, com relevância para a seleção de casos e atributos, e o desenvolvimento de modelos de predição. Há várias alternativas e, em princípio, todas as técnicas apresentadas poderiam ser utilizadas. Entretanto, algumas podem ser mais úteis. Por exemplo, técnicas como árvores de decisão e regras baseadas em lógica clássica provavelmente enfrentarão dificuldades, em função das características de imprecisão presentes no mercado imobiliário. As técnicas para aprendizagem não supervisionada, tais como clusterização e análise fatorial, são mais flexíveis e adequadas para lidar com relacionamentos não previamente conhecidos.

Em termos de geração de modelos preditivos, ressaltam-se os modelos aditivos generalizados, as redes neurais e os sistemas de regras difusas. Os dois primeiros são adequados para ambientes em que estão presentes relacionamentos não lineares e nos quais há falta de um modelo prévio dos relacionamentos, e os sistemas baseados em regras difusas podem considerar as imprecisões típicas do mercado imobiliário. Face às habilidades e deficiências que cada técnica apresenta, a aplicação em sistemas híbridos é recomendável, como será detalhado no próximo capítulo.

5 PROPOSTA DE UMA NOVA ABORDAGEM PARA A AVALIAÇÃO DE IMÓVEIS

5.1 CONSIDERAÇÕES INICIAIS

A proposta de uma nova abordagem para as avaliações deve levar em conta alguns elementos essenciais sobre o funcionamento do mercado e sobre a tarefa de avaliação de imóveis. A avaliação (estimação do valor de mercado) é a busca do preço mais provável que um imóvel pode atingir em condições normais, e o mercado imobiliário tem algumas características peculiares que afetam as estimativas.

Em primeiro lugar, o mercado imobiliário é de concorrência imperfeita. Neste ambiente econômico, as informações que podem ser obtidas no mercado imobiliário são os preços pagos, os quais são apenas *proxy* dos valores de mercado dos imóveis, ou seja, há um erro de medida quanto ao elemento principal da análise. Assume-se geralmente que este erro é aleatório e segue uma distribuição Normal, ao menos aproximadamente. Portanto, a média dos preços deve igualar-se à media dos valores em grandes amostras. Esta hipótese indica a conveniência de utilizar médias ao invés de casos isolados como base para as estimativas. Em outras palavras, a média é uma forma de minimizar o erro das avaliações.

Além desta questão, como os imóveis são bens heterogêneos, é preciso obter uma média ponderada, considerando as principais características diferenciadoras. Os efeitos simultâneos de várias características provocam dificuldades na comparação direta entre os casos e, portanto, há necessidade de obtenção de um modelo numérico para avaliação destas médias. O ajustamento é geralmente desenvolvido através de modelos hedônicos de preços (modelos microeconômicos de formação de preços), estimados por regressão múltipla. Quanto mais similares os elementos de comparação, em termos de características próprias das unidades e características de localização, menos ajustes serão necessários[35].

[35] Mas ainda será necessário obter a média dos preços destes imóveis. A simples identificação de casos similares quanto aos atributos não é suficiente, pois não se dispõe do valor de mercado para estes casos.

Também é importante analisar as peculiaridades decorrentes da finalidade do trabalho avaliatório. Nos três formatos apontados - modelos gerais de mercado, para avaliação em massa e para avaliação individual - há necessidade de precisão e explicabilidade dos resultados, embora em graus distintos. O tempo e os recursos disponíveis dificultam o uso de modelos detalhados em todas as ocasiões e a própria necessidade de uma análise ampla pode exigir o desenvolvimento de modelos gerais, que contemplem uma ampla área geográfica ou diversos tipos de imóveis em um único modelo. Os três modelos podem ser necessários em uma mesma empresa ou órgão público, e um sistema de avaliações deve atender a todos com a mesma base de dados, por questões de consistência entre os modelos e de facilidade e custos de manutenção da base. Assim, é necessário adotar uma base de dados consistente, preparada antecipadamente, que sirva de fonte para todos os modelos, e a proposta consiste de duas etapas, basicamente:

(1) Geração de uma base de dados consistente e capaz de embasar uma visão geral quantitativa do mercado, envolvendo a preparação dos dados e a extração (a partir destes dados) de conhecimento sobre localização, casos e atributos relevantes.

(2) Geração de modelos de maior precisão, através de técnicas alternativas, utilizando técnicas isoladas ou em abordagens híbridas, com a previsão de técnicas para subdivisão dos dados para os modelos de massa e individual.

Não há alterações no método de avaliação em si (comparação de dados de mercado), mas na forma ou procedimento de execução deste método.

5.2 GERAÇÃO DA BASE DE DADOS

A abordagem proposta foi inspirada no processo de descobrimento de conhecimento em bases de dados (DCBD), especialmente na fase de preparação dos dados. As peculiaridades dos dados do mercado imobiliário, com freqüentes erros e omissões, exigem uma cuidadosa preparação nos dados, tendo em vista a construção de uma base de dados que possa apoiar as avaliações. O desenvolvimento destas tarefas de forma objetiva também pode contribuir para o aperfeiçoamento da modelagem, através da redução dos problemas dos dados e da própria redução da subjetividade na análise. Embora não tenham sido encontrados trabalhos que

explicitassem a preparação de dados de mercado imobiliário dentro do processo de DCBD, exemplos de aplicações em áreas próximas são animadores (Berry e Linoff, 1997, 2000; Cabena *et al.*, 1997; Han e Kamber, 2001; Soibelman e Kim, 2002; Weiss e Indurkhya, 1998; Westphal e Blaxton, 1998).

O desenvolvimento da base inicia pela preparação dos dados, utilizando um conjunto de técnicas. A análise inicial é baseada na visualização dos atributos, explorando o comportamento dos atributos. Através dos gráficos é relativamente fácil identificar casos com problemas, bem como ampliar o conhecimento sobre os próprios atributos (Berry e Linoff, 2000; Hair *et al.*, 1998; Kohavi, 2000; Pyle, 1999).

É comum a ocorrência de dados com falhas ou omissões na área do mercado imobiliário. O preenchimento de elementos omitidos é geralmente desenvolvido através de correlação, vizinhança próxima ou regressão (Hair *et al.*, 1998; Pyle, 1999; Weiss e Indurkhya, 1998). O algoritmo de vizinhança próxima já foi utilizado em avaliação de imóveis, como auxiliar na seleção de casos em RBC (O'Roarty *et al.*, 1997a, 1997b), na seleção de casos e estimação de valores (McCluskey *et al.*, 1997; McCluskey e Anand, 1999) ou diretamente na estimação de valores (Isakson, 1986), e é uma técnica útil para esta etapa, embora utilizado em uma função auxiliar, neste caso.

Na fase de enriquecimento devem ser coletados atributos úteis, considerando especialmente aqueles baseados em uma visão objetiva, tais como medidas de distância a pólos valorizantes, ou baseados nos dados disponíveis, utilizando medidas de localização baseadas nos erros cometidos pelos modelos (Dubin, 1988; Gallimore *et al.*, 1996; González *et al.*, 2002a; Lang e Jones, 1975; McCluskey *et al.*, 2000; Ward *et al.*, 1999).

A seleção de atributos e casos é uma das etapas mais importantes e algumas das técnicas utilizadas nesta tarefa são Box-Cox, análise de correlação, componentes principais, regressão e redes neurais. O procedimento de Box-Cox é freqüentemente citado como alternativa para consideração de efeitos não lineares na especificação da forma funcional, inclusive com aplicações no mercado imobiliário (Blackley *et al.*, 1984; Dantas e Cordeiro, 2001; Kang e Reichert, 1987; Milon *et al.*, 1984). A matriz de correlações é um instrumento simples, mas que permite identificar pares de atributos com relacionamento forte (Diaz, 2000; Neter *et al.*, 1990). Já a análise de componentes principais é útil para casos em que existem diversos

relacionamentos entre os atributos (multicolinearidade), e tem sido utilizada para identificar os relacionamentos ou obter um conjunto de novos atributos, não colineares, com bons resultados (González, 1993; Kain e Quigley, 1970; Morton, 1977; Wilkinson e Archer, 1973). A identificação dos atributos mais relevantes para os modelos preditivos, em uma abordagem tipo envoltório, utiliza as próprias técnicas de estimação, tais como regressão, nos formatos tradicional e com superfícies, e redes neurais. Ao final desta etapa, devem ser examinados os pressupostos da regressão (Gujarati, 2000; Neter *et al.*, 1990).

5.3 GERAÇÃO DE MODELOS

A outra parte da solução consiste na substituição da regressão múltipla por outras técnicas, menos sensíveis aos problemas nos dados, especialmente usando técnicas não-paramétricas. A utilização de uma base de dados geral deve ser complementada por técnicas para seleção dos dados, conforme o tipo de avaliação. Apresenta-se os mecanismos de seleção e a seguir, as técnicas testadas.

5.3.1 Seleção dos casos para a modelagem

O modelo geral (modelo de "Mercado") baseia-se na totalidade da base, excluindo apenas os casos com erros ou identificados como *outliers*. Este modelo fornece indicações para o desenvolvimento dos outros dois formatos, para os quais foram utilizados critérios específicos de seleção ou subdivisão.

Nos modelos de avaliação em massa buscou-se o aperfeiçoamento pela segmentação dos dados, gerando diversos sub-modelos. A divisão foi realizada com clusterização, usando o algoritmo *k-means*. Esta técnica já foi utilizada por alguns autores para identificação de sub-mercados (Bourassa *et al.*, 1999; Bourassa e Hoesli, 1999). Com esta ferramenta, bastante conhecida e empregada no descobrimento de conhecimento, podem ser identificados grupos de imóveis relativamente homogêneos (Berry e Linoff, 2000; Westphal e Blaxton, 1998).

Entretanto, na clusterização o número de grupos deve ser determinado pelo analista, exigindo a exploração deste parâmetro, o que dificultaria a utilização desta técnica para a identificação de dados para a avaliação individual. Para este formato de avaliações foi proposto um ciclo

avaliatório distinto, mais ágil, baseado nos conceitos do raciocínio baseado em casos (RBC), consistindo na seleção de uma amostra de casos similares ao caso dado (avaliando) na fase de recuperação com a geração de modelos com técnicas auxiliares na fase de adaptação dos casos (Figura 14). Havendo a necessidade de detalhamento dos modelos com vistorias, estas são desenvolvidas antes da modelagem. Os modelos gerais de mercado podem contribuir gerando estimativas expeditas do valor, importantes para orientar o analista. Após a geração dos modelos e a estimativa dos valores de mercado, o restante do ciclo é similar ao RBC tradicional (conforme Aamodt e Plaza, 1994). A verificação das soluções é difícil, no mercado imobiliário, pois há grande espaço para a negociação entre as partes, mas pode ser realizada através dos intervalos de confiança (IC), aceitando-se a solução proposta quando o preço praticado estiver inserido neste intervalo. Neste caso, ocorre a retenção do caso (aprendizagem).

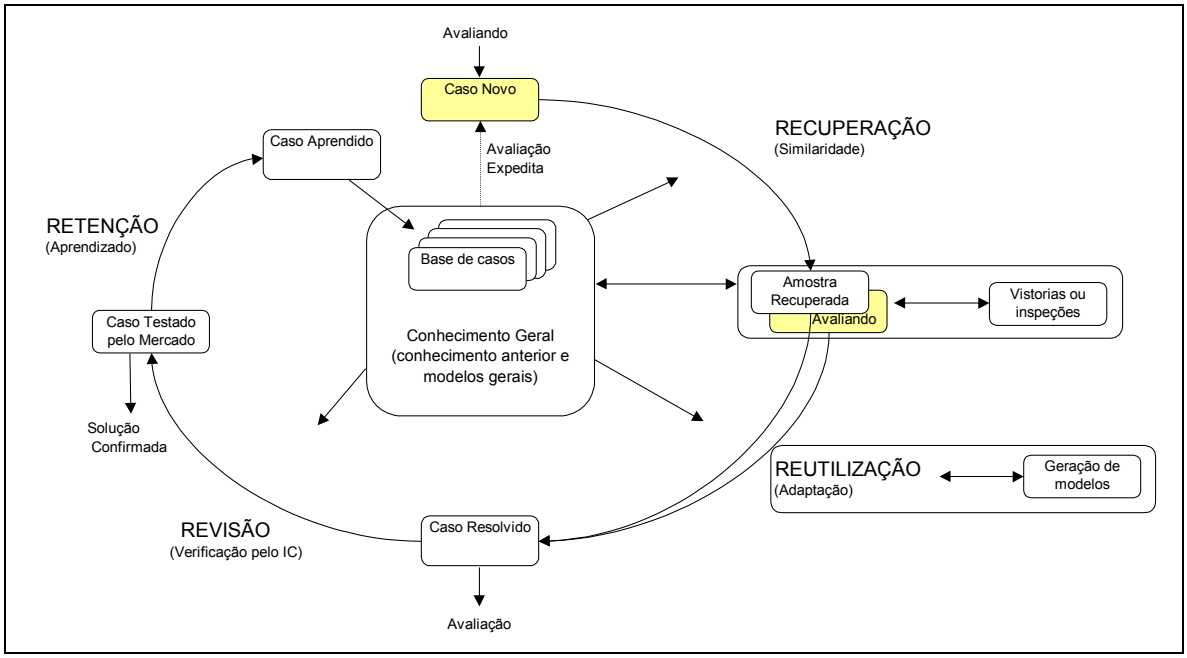


Figura 14: Ciclo proposto para avaliação individual (adaptado de Aamodt e Plaza, 1994).

As aplicações de RBC no mercado imobiliário geralmente são projetadas para a seleção e posterior adaptação de um pequeno número de casos (tipicamente 3 casos), usando o algoritmo de vizinhança próxima, com os pesos determinados *a priori* por especialistas ou através de regressão, redes neurais ou algoritmos genéticos (González e Laureano-Ortiz,

1992; O'Roarty *et al.*, 1997a, 1997b; Pacharavanich *et al.* 2000; Ribeiro, 1999). Contudo, como foi visto, há um componente aleatório nos preços, os quais não são totalmente explicados através dos atributos normalmente coletados, e a média baseada em poucos casos não é uma medida conveniente. Ademais, o desenvolvimento de um modelo específico, gerado para os dados da amostra, poderá aperfeiçoar a estimativa em relação ao resultado obtido com o algoritmo de vizinhança próxima com pesos genéricos.

Desta forma, a principal diferença do sistema proposto para o mecanismo tradicional do RBC é que o sistema deve selecionar uma amostra ampla, e não apenas alguns casos. Foi empregado um mecanismo de vizinhança próxima para a identificação da similaridade, com pesos baseados nos modelos gerais, agregando uma ponderação exponencial com a distância entre os imóveis.

No mercado imobiliário, assumem importância a localização e diversos outros atributos. A localização é um atributo que não pode ser mensurado diretamente, mas sabe-se que dois imóveis próximos tendem a dispor da mesma acessibilidade e da mesma qualidade de vizinhança, diminuindo a semelhança a medida em que a distância entre eles aumenta. Assim, a similaridade de localização pode ser medida usando a distância (em linha reta) entre eles. Em relação aos demais atributos, as diferenças entre os imóveis podem ser determinadas pelas diferenças encontradas em cada atributo relevante, ponderadas pela importância relativa destes atributos. O formato do índice de similaridade adotado consiste da seguinte relação (Equação 24):

$$SIM_{a,b} = 1 / [(1+k_0*d_{a,b}^2)*(1+k_1*\Sigma_i (a_i*|X_{ib}–X_{ia}|^{ei}))] *100\% \qquad \text{(Equação 24)}$$

Onde $SIM_{a,b}$ é a medida de similaridade entre o avaliando (a) e o caso da base (b), $d_{a,b}$ é a distância entre eles (em quilômetros) e a segunda parcela é a medida das diferenças entre os dois, identificadas através dos seus conjuntos de atributos X_{ia} e X_{ib}, respectivamente, e ponderadas através dos preços hedônicos a_i. Adicionalmente, podem ser empregados reforços na ponderação usando expoentes distintos para cada atributo (e_i). Foram utilizadas duas constantes (k_0, k_1) para calibrar a medida, em função da dimensionalidade dos termos.

A similaridade máxima é atingida quando os dois componentes são anulados

simultaneamente, ou seja, quando o caso da base é um imóvel com características iguais e situado no mesmo prédio do avaliando. Obtida a amostra de casos similares (casos relevantes), podem ser realizadas complementações com dados de campo e um modelo preditivo específico é determinado, finalmente gerando a estimativa.

5.3.2 Técnicas para o desenvolvimento de modelos

Em todos os casos citados, a tarefa de modelagem a ser desenvolvida consiste na estimação de modelos preditivos do tipo regressão, com foco em uma variável contínua (no caso, o valor do imóvel). A revisão bibliográfica sobre técnicas de inteligência artificial indicou que algumas técnicas são mais adequadas para resolver este tipo de tarefa, tais como redes neurais e regras difusas. Ademais, há uma tendência de utilização de sistemas híbridos nos últimos anos, com duas ou mais técnicas associadas de forma a obter melhores resultados do que seriam obtidos isoladamente. Considerando estes elementos, as alternativas empregadas para a estimação dos valores incluem algumas técnicas de inteligência artificial e a regressão múltipla:

a) análise de regressão tradicional;

b) regressão com superfícies matemáticas;

c) redes neurais com explicação por regras difusas;

d) modelos aditivos generalizados estimados com algoritmos genéticos;

e) sistemas de regras difusas, obtidas com algoritmos genéticos.

A primeira alternativa envolve os tradicionais modelos hedônicos estimados por análise de regressão múltipla. Em função do largo uso desta ferramenta na área de avaliações, a análise sobre as diferenças entre os modelos utilizou os modelos de regressão como referência.

A segunda alternativa baseia-se em superfícies de resposta, utilizando a técnica de *Trend Surface Analysis*, a qual incorpora às variáveis normais polinômios com composições das coordenadas dos imóveis, considerando a localização através de equações ou modelos, incluindo termos com as coordenadas planas (X, Y) em vários graus. A finalidade desta abordagem é compensar as dificuldades decorrentes do desconhecimento relativo sobre os

reais valores de localização. Pode ser empregada na ausência de conhecimento especializado, para verificação ou detalhamento de algumas regiões, para atualização de trechos com elevada atividade imobiliária ou em transição (para os quais os especialistas ainda não reúnem conhecimento suficiente), ou mesmo para a atualização intermediária entre apreciações de comissões de especialistas (nas revisões anuais das plantas genéricas de valores, por exemplo). As superfícies podem ser estimadas com regressão, com redes neurais ou com algoritmos genéticos. A primeira opção foi escolhida, em função da disponibilidade de testes estatísticos para seleção das parcelas (González *et al.*, 2002a).

As redes neurais reúnem habilidades para estimação em domínios não lineares ou com desconhecimento inicial dos modelos, uma das dificuldades apontadas com os modelos hedônicos tradicionais. As dificuldades das redes neurais residem na falta de um modelo explícito, o qual pode ser exigido para tributação ou avaliações judiciais. Para obter estes modelos, explicando o comportamento das redes neurais, foram empregadas regras difusas extraídas a partir das redes treinadas (González *et al.*, 2002b).

O quarto tipo apresentado consiste de Modelos Aditivos Generalizados, os quais são basicamente modelos hedônicos não lineares. Também podem ser considerados como uma forma de regressão não paramétrica, ajustada por algoritmos genéticos. A finalidade é considerar relacionamentos não-lineares existentes nos dados, como ocorre com as redes neurais, porém através de um modelo hedônico explícito (Mason e Quigley, 1996; Pace, 1998).

Por fim, foram utilizados modelos baseados em sistemas de regras difusas. Outra dificuldade apontada nos modelos convencionais é a segmentação de parcelas homogêneas do mercado, considerando características físicas dos imóveis ou de localização. Além dos próprios critérios de segmentação, surgem problemas na definição das fronteiras entre dois grupos, as quais normalmente revelam diferenças abruptas. A lógica difusa permite considerar transições mais adequadas, mantendo a continuidade. Foram desenvolvidos dois sistemas de regras difusas, um baseado no tamanho, utilizando a área total, e outro baseado na localização. Um das vantagens deste formato é a fácil atualização ou aperfeiçoamento da base, através da inclusão de novas regras.

Os modelos de avaliação em massa, nos quais existem conjuntos de modelos baseados em subgrupos de imóveis homogêneos, podem ser considerados como árvores de regressão ou

sistemas de regras baseadas na lógica clássica, no caso das quatro primeiras soluções, enquanto que os sistemas baseados em regras difusas consistem em um sistema único. A principal diferença entre eles situa-se em relação ao disparo de cada regra ou modelo, que nas regras lógicas é {0,1}, ou seja, aplica-se apenas uma regra para cada caso (imóvel), e nas difusas é um valor no intervalo [0,1], com a possível utilização de mais de uma regra difusa no cômputo das estimativas.

5.4 IMPLEMENTAÇÃO

Foi gerada uma aplicação do sistema proposto, gerando inicialmente a base e em seguida aplicando as técnicas para gerar modelos preditivos. Conforme o fluxo apresentado na Figura 1, o início da análise consiste na coleta de conhecimento do domínio, que no caso versa sobre o mercado imobiliário em termos gerais e especificamente sobre o mercado imobiliário de Porto Alegre, com a identificação do problema e das possíveis soluções, tarefa que foi desenvolvida nos quatro capítulos anteriores.

Para o desenvolvimento da parte empírica do trabalho, buscou-se uma base de dados relativamente grande, em virtude do uso de técnicas da área de inteligência artificial, as quais geralmente oferecem melhores resultados com bases de dados maiores. Ademais, o exame de modelos para avaliação em massa também exigia que os dados fossem em quantidade e disseminados pela área da cidade. Em função dos prazos da parte empírica deste trabalho, era inviável a coleta tradicional, com busca exaustiva em imobiliárias e vistorias in loco (como realizado em Franchi, 1991 e González, 1993, por exemplo), e buscou-se uma fonte que disponibilizasse diretamente uma grande quantidade de dados. Não há instituições que coletem e disponibilizem comercialmente dados do mercado imobiliário em Porto Alegre. Tendo em vista os objetivos propostos, optou-se pela utilização de dados do setor de tributos da Prefeitura Municipal de Porto Alegre, consistindo de declarações dos contribuintes no pagamento do imposto de transmissão dos imóveis. Como a participação dos apartamentos no estoque de imóveis da cidade é de aproximadamente 56% dos imóveis registrados no cadastro municipal (De Cesare, 1998), este tipo de imóvel foi escolhido para a análise.

A preparação dos dados desenvolvida está descrita no capítulo 6, bem como a mineração e

análise dos dados estão detalhadas no capítulo 7.

6 PREPARAÇÃO DOS DADOS

6.1 COLETA DOS DADOS

Os dados brutos foram obtidos junto à Secretaria da Fazenda do Município de Porto Alegre em agosto de 2001. Foi solicitada uma autorização especial ao Secretário da Fazenda, e os dados foram liberados com o compromisso de não serem divulgados individualmente e utilizados apenas para finalidades acadêmicas. A Secretaria armazena as declarações em meio eletrônico, em sistema próprio. Foram recebidos os casos de transmissão com pagamento de tributo intervivos, de competência municipal, reunindo informações sobre 31.277 declarações de transações de apartamentos ocorridas no intervalo de 07/08/1998 (início do registro completo das guias no sistema da Prefeitura)[36] a 01/08/2001 (data da extração dos dados). Os dados recebidos correspondem a todos os casos registrados de declarações de transmissão de apartamentos no período citado (incluem apenas os casos com efetivo pagamento do tributo).

6.1.1 Reconhecimento dos dados – significado, limitações e escopo

Este tipo de dado já foi utilizado anteriormente, no âmbito de outras pesquisas (Furtado, 1993; González e Formoso, 1995a, 1995b; Melazzo, 1993; Smolka *et al.*, 1989). Há peculiaridades dos dados provenientes de declarações de contribuintes. Por exemplo, não estão incluídos os casos de compra e venda de imóveis financiados sem a efetiva transferência, para evitar o refinanciamento do saldo devedor (conhecidos como "contratos de gaveta"). Também há peculiaridades na informação sobre os valores dos imóveis, pois os dados obtidos possuem algumas limitações, especialmente quanto aos valores declarados pelos contribuintes e mesmo quanto às estimativas fiscais (valores estimados para os imóveis), pelas dificuldades em

[36] Até essa data, eram registradas no sistema apenas informações básicas, tais como o número da guia, a data da declaração, os valores declarado e estimado e o valor do tributo pago.

avaliar grandes quantidades de imóveis semanalmente.

Para evitar a quebra do sigilo fiscal, a Secretaria da Fazenda não forneceu a identificação completa do imóvel, omitindo o número do apartamento. Assim, torna-se difícil a correção dos eventuais erros nos dados através da consulta ao cadastro municipal, além da própria dificuldade de acesso ao cadastro (permissão para consulta). Desta forma, o procedimento empregado foi de corrigir ou complementar os valores omissos e excluir os dados com erros, após um exame cuidadoso. Em aplicações normais de empresas ou avaliadores também não há acesso fácil aos cadastros municipais, reforçando a opção por esta solução. O volume de dados obtido e as finalidades do estudo (de caráter exploratório) também permitem a adoção deste caminho.

6.1.2 Dimensionamento e extração

Verificou-se que a quantidade de dados disponível era adequada, sendo de tamanho viável para a análise pretendida. Em virtude da disponibilidade eletrônica, a amostra obtida abrangeu todos os dados do período, para o tipo de imóvel "apartamento". Neste período, o cadastro municipal continha cerca de 275 mil apartamentos, portanto a amostra coletada representa aproximadamente 11% do total de imóveis no município, tendo em vista que este tipo de imóvel dificilmente escapa ao cadastramento público e que a dupla venda no período de 3 anos provavelmente atingiria uma reduzida parte da amostra.

6.1.3 Conversão de arquivos

Os dados foram recebidos em oito arquivos de formato padrão Microsoft Excel (tipo "xls"). A primeira atividade foi a verificação da integridade dos arquivos, com a consolidação em um arquivo único com os dados ordenados cronologicamente (ordem original), gerando-se um código geral de identificação do caso (um número inteiro de 1 a 31.277). Os dados foram transferidos para o pacote estatístico SPSS[37]. O exame geral dos dados brutos, com a

[37] Foi utilizada a versão 9.0 do *Statistical Package for Social Sciences* – http://www.spss.com (SPSS, 1999)

compreensão do significado dos campos, foi realizado neste *software*, em virtude da facilidade de geração de gráficos, bem como determinação dos parâmetros estatísticos. Foi utilizado também como base geral para os dados, em função das facilidades de tratamento e de conversão para outros formatos.

6.2 EXPLORAÇÃO E LIMPEZA INICIAIS

Após a obtenção dos dados, foi realizado um exame geral verificando-se o comportamento das variáveis. Como são dados que exigem autorização para consulta, nestes casos a remoção ou correção através de técnicas específicas foi a melhor alternativa, tendo em vista ainda a disponibilidade de uma grande quantidade de dados. Se a amostra fosse pequena, para um dado tipo de imóvel ou região com problemas, poderia ser realizada uma verificação nas guias originais ou a busca no cadastro, visando corrigi-los. Naturalmente, é muito útil o conhecimento anterior sobre a cidade e sobre a fonte de informação (dados fiscais).

6.2.1 Identificação dos campos (nomes e significados)

As variáveis recebidas foram as relacionadas a seguir. Os títulos originais dos campos foram mantidos para facilitar a comunicação com o pessoal do setor em eventuais contatos futuros. As observações que acompanham a descrição resultam de conhecimento anterior, da análise dos dados e de informações obtidas em contatos com funcionários do setor durante o período da pesquisa.

a) Data: dia, mês e ano da declaração;

b) Logradouro e imóvel: endereço (rua e número do prédio);

c) Área total do terreno (Artoterr) e área transmitida do terreno (Artrterr): indicam a superfície total e a parcela transmitida juntamente com o apartamento;

d) Área total de construção (Artocons) e área construída privativa (Arcopriv): estas variáveis indicam a área total e a área privativa, respectivamente.

e) Padrão construtivo (Tipo): esta variável indica o tipo construtivo, conforme

padrões internos da secretaria de obras do município. O imóvel é examinado quando da vistoria de licenciamento ("Habite-se"), e a categoria atribuída é registrada no cadastro municipal;

f) Ano da construção (Ano.Const): trata-se do ano de construção do prédio, indicado pelo ano em que foi realizada a vistoria de licenciamento. Pode ser nominalmente distinto para apartamentos no mesmo prédio, em função de peculiaridades como interrupções da obra, limitações de financiamento apenas para imóveis novos, reformas, problemas para aprovar alterações de projeto e outros casos;

g) Valor declarado (Vl.Declarado): valor da transação, segundo declaração do contribuinte. Geralmente é indicado o preço praticado. Em muitos casos este valor é subestimado, na tentativa de reduzir o imposto a ser pago. Em outros casos é lançado um valor histórico (do contrato de promessa de compra e venda, por exemplo);

h) Valor atribuído (Vl.Atribuido): é a avaliação do fiscal de tributos para o imóvel. Em princípio, representa o valor de mercado para o imóvel, mas o fiscal geralmente adota o valor declarado, se este é superior, e também comete erros de avaliação, até em função do volume de estimativas. Pode ser um número sem significado prático, se a transação é isenta de imposto. Também é comum que os fiscais usem tabelas ou modelos simplificados na avaliação.

6.2.2 Análise do comportamento das variáveis (estatística descritiva e visualização)

Inicialmente foi utilizada a estatística descritiva para a exploração dos limites do conteúdo das variáveis numéricas. As medidas estatísticas da base de dados estão apresentadas na Tabela 2. Mesmo nesta análise simplificada, existem indicações de possíveis erros, tais como os valores máximos para declarações e estimativas fiscais e as áreas mínimas de terrenos e construções, além de 1.765 casos de área privativa com valor nulo.

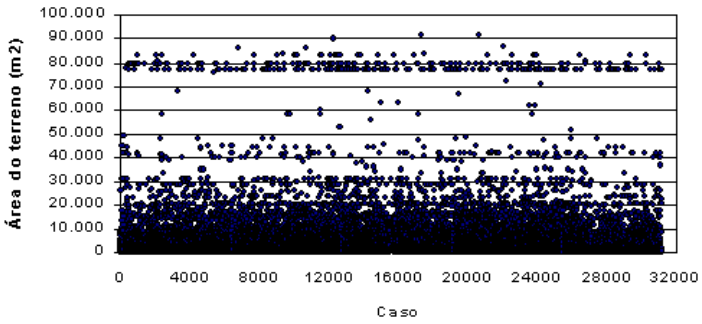
Tabela 2: Características das variáveis originais

Variável	Tipo	Unidade	Mínimo	Máximo	Média	Desvio-padrão
Artoterr	Contínua	m^2	1,00	91.416,00	4.516,52	10.160,85
Artrterr	Contínua	m^2	0,01	941,40	40,89	33,99
Artocons	Contínua	m^2	1,00	2.306,00	93,47	74,82
Arcopriv	Contínua	m^2	0,00	760,52	70,11	51,38
Tipo	Categórica	-	-	-	-	-
Ano.Const	Discreta	ano	0	2001	1979,98	55,63
Vl.Declarado	Contínua	R$	0,00	18.172.078,00	51.598,04	122.413,13
Vl.Atribuído	Contínua	R$	0,01	18.172.078,00	65.885,69	131.462,01

Além das variáveis descritas na Tabela 2, duas outras colunas constantes do arquivo original, denominadas de "Estacionamento Coberto" e "Estacionamento Descoberto", deveriam conter variáveis binárias indicando o tipo de estacionamento, mas vieram sem nenhuma informação (vazias), e foram eliminadas da análise posterior.

A análise das variáveis através de gráficos complementa a análise numérica, permitindo o exame da distribuição relativa dos dados, que pode indicar se os elementos extremos estão deslocados dos demais ou se há continuidade nos valores. Geralmente é possível identificar o caso no próprio gráfico. Os gráficos apresentados na Figura 15 são gráficos de dispersão, exceto para a variável Tipo, que é categórica, sendo apresentado um gráfico de barras, indicando a quantidade de casos em cada categoria. Nestas figuras (exceto para a Figura 15e), o eixo horizontal indica o número do caso, o qual cresce com o tempo, portanto também podem ser utilizados para verificar o comportamento temporal das variáveis, embora o período abrangido seja pequeno (37 meses).

(a) Artoterr



(b) Artrterr

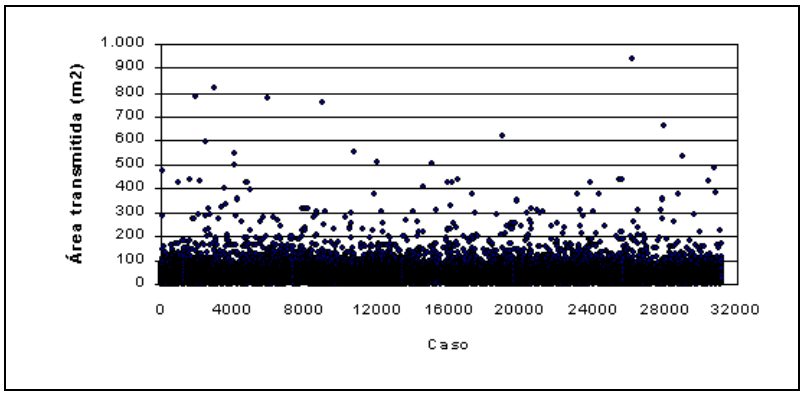
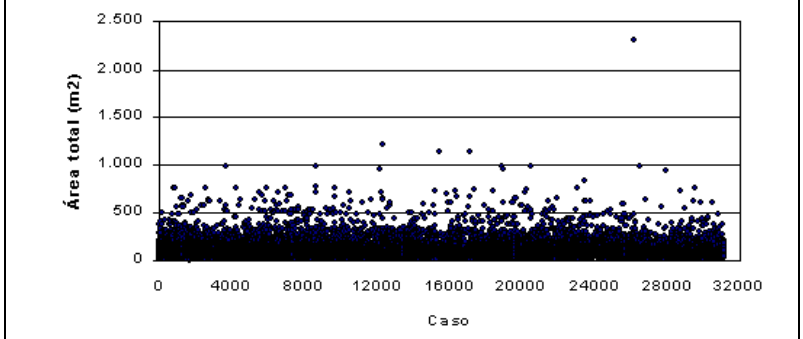
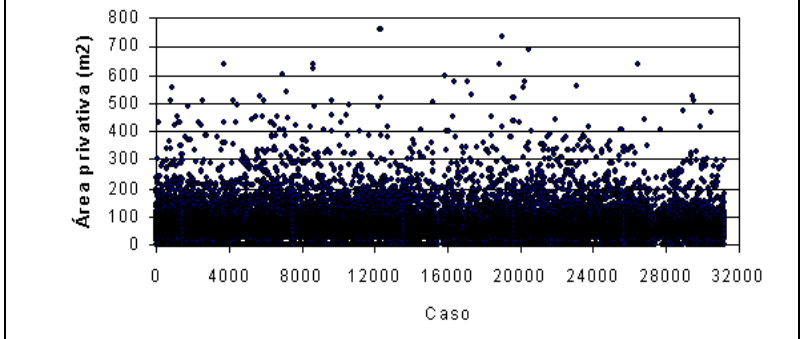


Figura 15: Gráficos de dispersão das variáveis originais

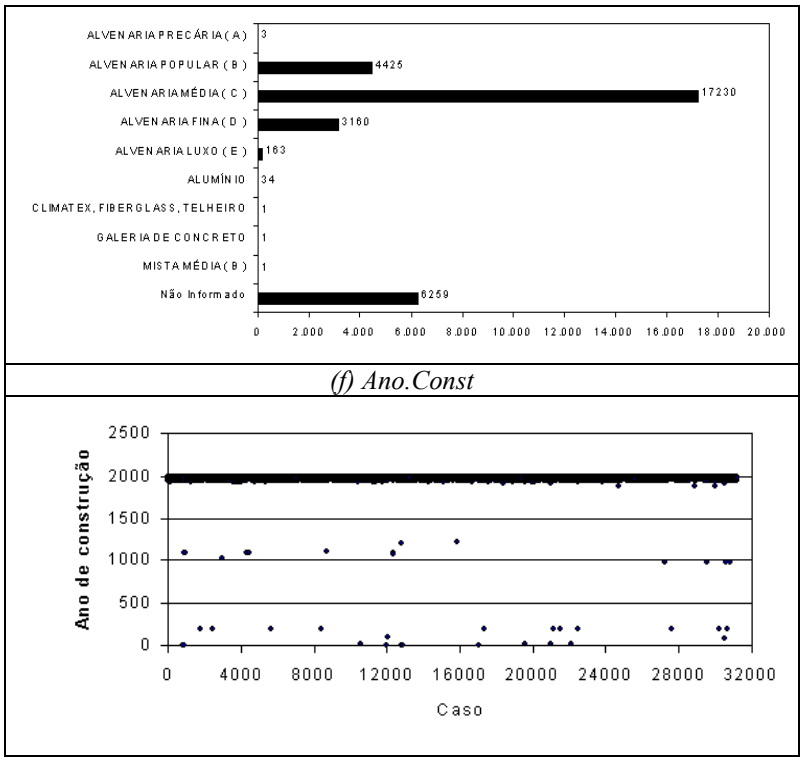
(c) Artocons



(d) Arcopriv



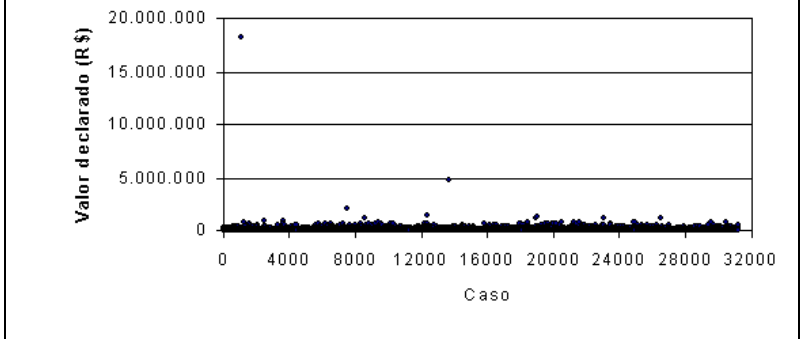
(e) Tipo



(f) Ano.Const

Figura 15: Gráficos de dispersão das variáveis originais (continuação)

(g) Vl.Declarado



(h) Vl.Atribuído

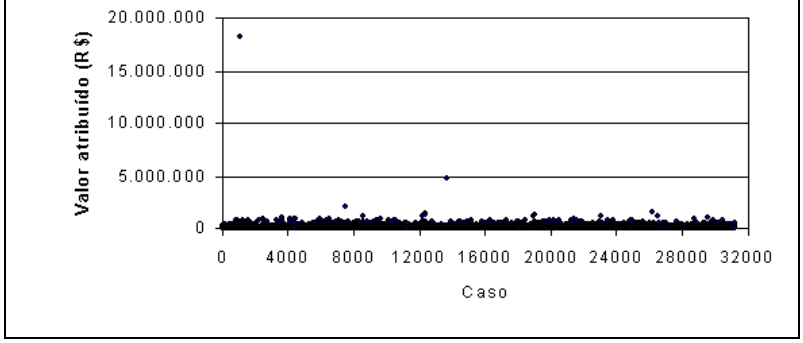


Figura 15: Gráficos de dispersão das variáveis originais (continuação)

O exame das variáveis indicou a presença de alguns erros. Para algumas variáveis, o conhecimento disponível permite a obtenção de soluções para correção ou preenchimento de valores omitidos através de outros apartamentos no mesmo prédio ou em prédios vizinhos, conforme será explanado na seção seguinte. Por exemplo, para um mesmo prédio a área total do terreno deve ser igual para todos os apartamentos. Por outro lado, outras variáveis apresentaram casos com grande afastamento dos demais, os quais podem ser erros, *outliers* ou simplesmente outros tipos de imóveis (outras categorias), mas que provavelmente não serão úteis na modelagem. As variáveis com casos removidos nesta etapa foram a área total e o valor dos imóveis:

a) Área total construída (variável Artocons): a análise através de dois gráficos, um em escala linear (Figura 15) e outro em escala logarítmica (Figura 16) permite verificar os extremos superior e inferior, respectivamente. No caso, foram identificados dois imóveis com claro afastamento dos demais. No extremo superior, há um imóvel com área de 2.306 m^2, que pode ser considerada grande demais para um apartamento, no contexto de Porto Alegre. Verificando a localização e o preço, conclui-se que provavelmente se trata da transferência de um prédio inteiro, classificado inadequadamente como um apartamento. No extremo inferior, há um imóvel com área de 1 m^2, com outras informações incorretas, e que não também não deve ser considerado. Existem diversos imóveis (cerca de 100 casos) com áreas entre 10 m^2 e 20 m^2, os quais podem ser garagens, transacionadas individualmente, ou mesmo apartamentos compactos, tipo "quarto e sala" (conhecidos em Porto Alegre como "JK"), e foram mantidos para a análise posterior, pela dificuldade de identificação da real situação e pela relativa continuidade das áreas no extremo inferior.

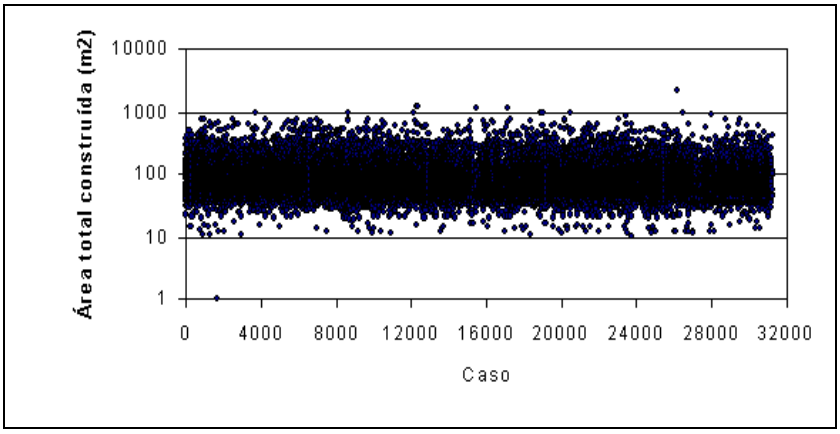


Figura 16– Gráfico logarítmico da área total (Artocons)

b) Valores dos imóveis (variáveis Vl.Declarado, Vl.Atribuído, Valor e Unitário): em função do valor ser a variável de maior interesse e tratar-se justamente da variável a ser explicada, não foram realizadas correções, nos casos em que há dúvidas sobre os valores, mas apenas exclusões. Os valores extremos encontrados não são razoáveis: R$ 0,00 e R$ 18 milhões por si só indicam problemas (ver Tabela 2). Examinando os dados, verificou-se que alguns erros aparentemente decorriam de erros de digitação, enquanto que outros podem ser explicados por peculiaridades da fonte dos dados. No caso dos valores declarados, é comum ocorrerem casos de sub-declaração, quando o contribuinte indica o valor nominal do contrato, que é de outra época (por exemplo, no caso de financiamentos possivelmente 15 ou 20 anos antes), ou propositalmente indica um valor inferior, tentando pagar uma quantia menor de imposto de transmissão. Mas não se espera que o fiscal aceite esta declaração, como em 11 casos em que o valor atribuído também está abaixo de R$ 1,00. Nestes casos, verificou-se, em contato com funcionários da Prefeitura, que se trata de transações em que o contribuinte está isento de pagamento, o que diminui o interesse do fiscal. A guia deve ser preenchida e incluída no sistema, mas não há conferência ou avaliação sobre os dados constantes da mesma.

Percebe-se que tanto o valor declarado pelo contribuinte quanto o valor estimado pelo fiscal apresentam restrições em relação aos preços praticados pelo mercado. Entretanto, a análise indicou que os valores declarados continham muitos casos com valores irrisórios, bem como outros erros, conforme já detectado em pesquisas anteriores (González e Formoso, 1995a). Assim, em função da maior consistência, foi escolhido o valor estimado como variável a ser analisada.

Para realizar a limpeza dos dados, é interessante inicialmente considerar algumas transformações nos dados, tal como a determinação da correção monetária sobre os valores declarados, pois a utilização dos valores nominais pode causar distorções na análise, especialmente em caso de largos intervalos de tempo nos dados coletados ou presença de inflação[38]. Assim, a análise foi realizada sobre os valores em moeda aproximadamente constante, usando-se os valores atribuídos totais (Valor) e unitários (Unitário), ambos

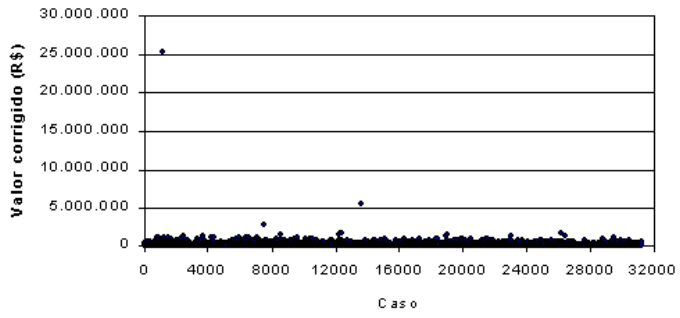
[38] O exame de alguns dados pode ser aprimorado com a inclusão de variáveis auxiliares. Esta é uma etapa de enriquecimento, pois inclui informações úteis para a análise desejada, e que está posicionada após o exame inicial, conforme o fluxo apresentado no Capítulo 3. Entretanto, como foi visto, a preparação dos dados não é realizada em etapas estanques, sendo resultado de diversas ações de limpeza e enriquecimento, que se sucedem e se complementam mutuamente, sendo difícil executar as etapas em uma seqüência estanque.

corrigidos monetariamente[39]. Os gráficos destas variáveis, apresentados a seguir, indicam o descolamento de algumas transações em relação às demais. Os gráficos em escala linear (Figuras 17a e 17b) permitem investigar apenas o extremo superior, pois há uma concentração maior de valores no outro extremo, dificultando o exame visual. Para verificar o extremo inferior foram construídos gráficos em escala logarítmica (Figuras 18a e 18b).

Verifica-se que os imóveis com valor total superior a R$ 5 milhões ou inferior a R$ 1.000,00 são casos muito discordantes dos demais, e não podem ser utilizados (Figuras 17a e 18a). A investigação dos valores unitários confirmou estes problemas. O exame dos valores unitários, já excluindo esses dados, revelou outros limites (Figuras 17b e 18b), indicando valores máximos de R$ 4.000/m^2 e mínimos de R$ 10/m^2.

Na análise é essencial verificar também os preços dos imóveis vizinhos, verificando o contexto local de preços, pois geralmente há uma variação apreciável de preços dentro da área urbana. Mesmo assim, diversos casos foram marcados (identificados como erros), conforme apontado na Tabela 3.

(a) Valor



(b) Unitário

[39] Os valores declarados foram corrigidos pelo IGP-DI, conforme descrito adiante, na seção sobre enriquecimento.

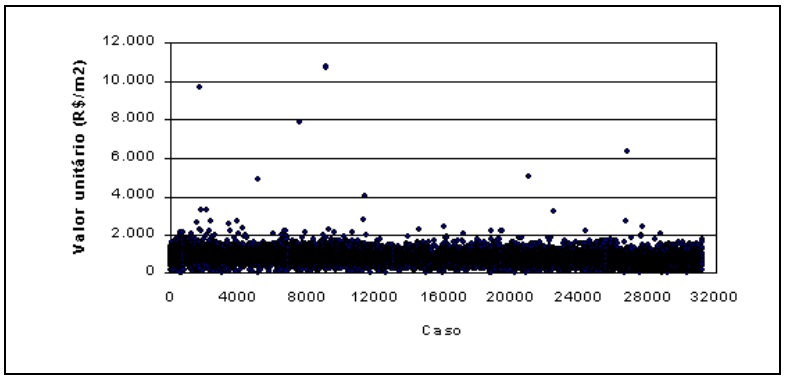


Figura 17 – Gráficos lineares das variáveis Valor e Unitário

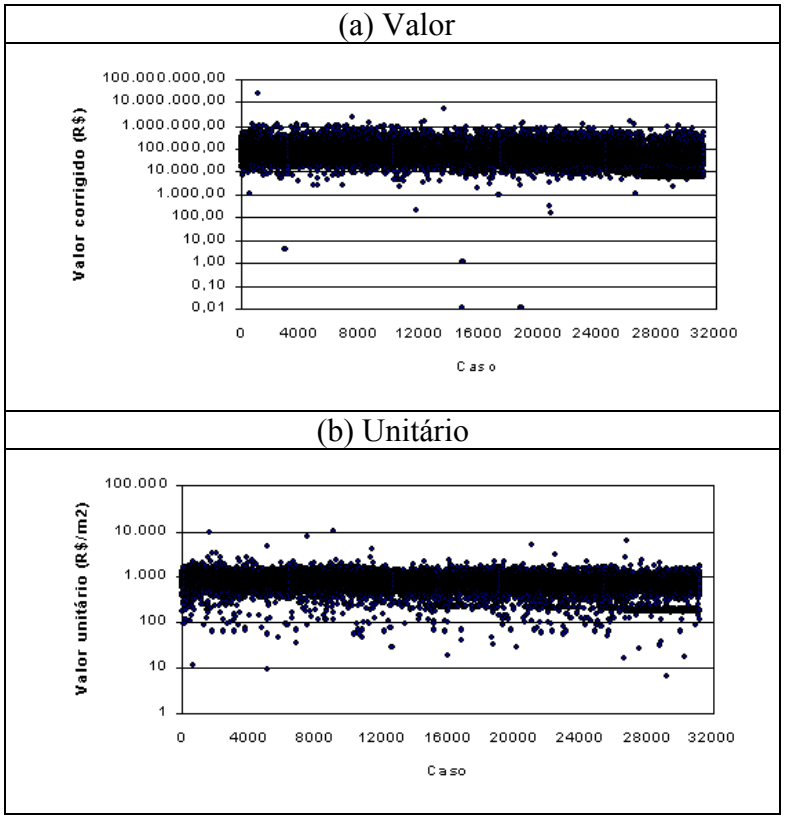


Figura 18 – Gráficos logarítmicos de Valor e Unitário

Os valores encontrados no extremo superior são tão elevados que podem indicar erros de digitação ou casos de super-declaração, os quais são suspeitos de serem casos de "lavagem de dinheiro". Como os valores declarados pelos contribuintes e os atribuídos pelos fiscais são idênticos, possivelmente não são casos de erro de entrada. Não são conhecidos casos de apartamentos com valores próximos de R$ 18 milhões, nem são esperados valores unitários acima de R$ 10 mil/m^2 em Porto Alegre. Quando a declaração do contribuinte é superior à estimativa fiscal, normalmente o fiscal deve aceitar o valor declarado como base para o

tributo. Este procedimento não é correto, pois o texto da lei indica o "valor venal" como base de cálculo[40], ou seja, o valor de mercado, o qual nem sempre coincide com o valor da transação (preço), que provavelmente foi o valor declarado pelo contribuinte. Em todos os casos, o tributo deveria ser baseado na estimativa do valor de mercado. Por outro lado, estas transações "suspeitas" devem ser informadas à Receita Federal, para que sejam posteriormente investigadas. De qualquer forma, não são transações normais, e os casos devem ser removidos da análise posterior.

6.2.3 Limpeza inicial (identificação de erros grosseiros em casos ou variáveis)

A estratégia adotada foi de não remover efetivamente os casos suspeitos, apenas marcar os mesmos através de uma variável identificadora. Assim, foi incluída uma nova variável, denominada Excluídos, com valor zero para os casos normais e com a indicação do motivo para exclusão através de um código, variando de 1 a 5 para os demais. Com este procedimento, pode-se facilmente desenvolver a modelagem com ou sem os dados suspeitos, permitindo também a eventual correção futura (Tabela 3).

Tabela 3: Número de casos e de erros detectados

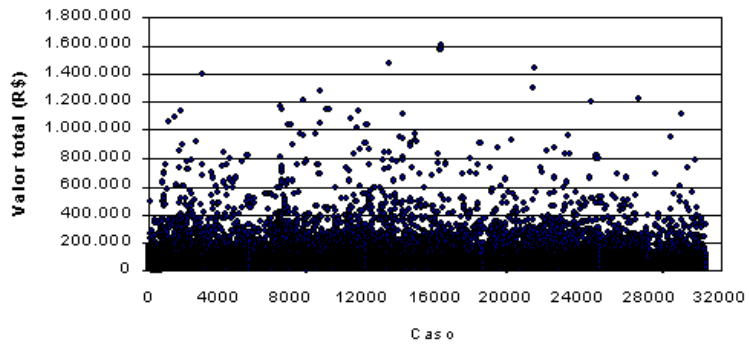
Situação	Código	Número de casos
Valor muito grande (>R$ 5milhões ou >R$ 4.000/m^2)	1	9
Valor muito pequeno (<R$ 1.000 ou <R$ 10/m^2)	2	35
Área total muito grande (>1.500 m^2)	3	1
Área total muito pequena (≤1 m2)	4	1
Diversos erros (mais de um dos anteriores)	5	10
Dados aparentemente corretos (mantidos)	0	31.221
Total	-	*31.277*

O resultado final para as variáveis afetadas pelas operações de limpeza pode ser verificado adiante, na Tabela 9. Após a limpeza, nova etapa de visualização deve ser desenvolvida, pois a remoção dos erros permite maior clareza no exame. Por exemplo, para as variáveis Valor e Artocons, os novos gráficos estão apresentados, respectivamente, nas Figuras 19a e 19b, após

[40] De acordo com a Lei Municipal que regulamenta o imposto de transmissão em Porto Alegre (LC 197/89, art. 11). Esta expressão pode ser encontrada na legislação de outras cidades brasileiras.

a remoção dos 56 casos com problemas[41].

(a) Valor



(b) Artocons

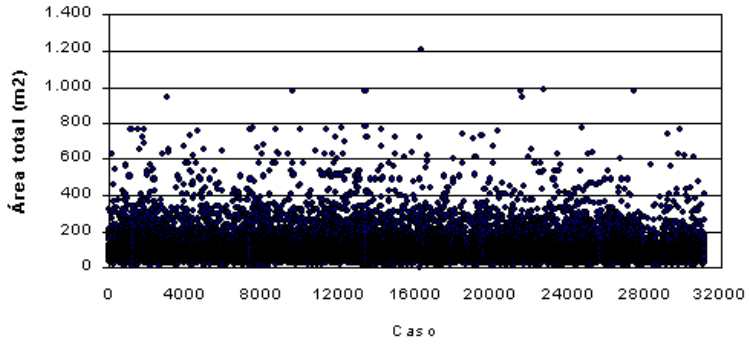
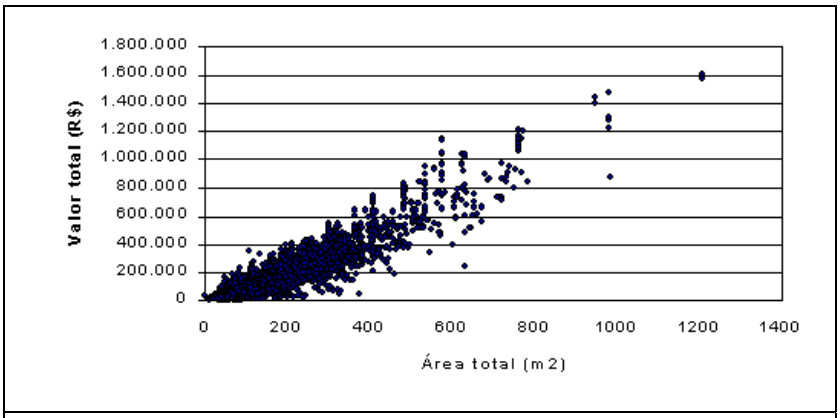


Figura 19 – Gráficos de Valor e Artocons, após a remoção de casos com problemas

Também pode ser desenvolvida uma análise bivariada, verificando o comportamento de pares de variáveis. Por exemplo, para a área total (Artocons) e os valores corrigidos (Valor e Unitário), os gráficos (Figuras 20a e 20b) indicam relacionamentos típicos, com forte relação da área com o valor total e com a diminuição dos valores unitários à medida que a área cresce.

(a) Artocons x Valor

[41] As figuras e tabelas apresentadas a partir deste item contam com 31.221 casos (excluindo os casos com erros).



(b) Artocons x Unitário

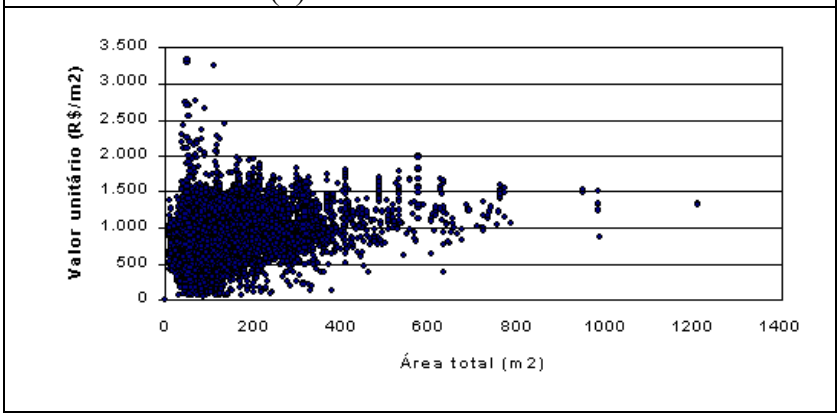


Figura 20 – Gráficos de Artocons x Valor e Artocons x Unitário

6.3 CORREÇÃO E COMPLEMENTAÇÃO DOS DADOS

Nesta etapa buscou-se corrigir alguns erros e complementar dados omitidos. Foram desenvolvidas rotinas para manipulação do banco de dados, visando corrigir ou amenizar problemas nas variáveis importantes, com os resultados descritos a seguir. O procedimento adotado foi o de manter os valores originais, gerando uma nova variável para receber os dados "corrigidos", visando permitir o teste posterior da melhoria nos dados. Após a correção também deve ser efetuada uma nova etapa de visualização dos dados, verificando os resultados (não apresentada para evitar repetição).

Tendo em vista a natureza espacial do mercado imobiliário, a estratégia principal de correção dos dados omitidos ou incorretos utilizou o algoritmo de vizinhança próxima (*k-nearest neighbors,* k-NN), em duas etapas. Quando existiam na amostra outros apartamentos do

mesmo prédio (distância zero) estes foram utilizados na estimação. Em alguns casos existiam diferentes medidas em parâmetros que deveriam ser iguais para todo o prédio, tal como a área do terreno ou o padrão de construção. Nestes casos foi realizada uma equalização, buscando o valor mais comum em cada imóvel. Não havendo outros casos no mesmo prédio, as estimativas foram obtidas pela média dos parâmetros dos 10 apartamentos mais próximos, usando o inverso da distância entre eles como elemento de ponderação[42].

Esta alternativa baseia-se no conhecimento geral sobre o fenômeno da estruturação intra-urbana: o ambiente construído influi nas decisões sobre as novas construções e pode-se presumir que haja similaridade entre imóveis construídos em locais próximos e com pequena diferença de tempo. As variáveis alteradas foram as seguintes:

a) Área do terreno (variáveis Artoterr e Artrterr): há vários casos de valores diferentes de área total do terreno para o mesmo prédio. Foi realizada uma equalização, adotando o tamanho mais comum para cada endereço, lançando os valores em uma nova variável, Artoterr2. A área transmitida do terreno foi ajustada na mesma proporção da variação de Artoterr para Artoterr2, gerando Artrterr2.

b) Área construída (variáveis Artocons e Arcopriv): o exame da área total não indicou outros problemas além dos relatados acima. Porém, existem 1.765 casos em que a variável Arcopriv tem o valor "Nulo", e outros 221 casos em que tem valor menor que 1 (somando mais de 6% do total de dados), além de outras distorções, ou seja, existem 29.291 casos válidos. Como se trata de variável importante, os valores correspondentes foram estimados também através do algoritmo k-NN, em duas etapas, como explanado acima. Esta forma de correção baseia-se na hipótese de que os imóveis em geral estão de acordo com a norma brasileira que regula a incorporação em condomínios, a NBR-12.721 (ABNT, 1992). Esta norma indica o rateio das áreas de uso comum proporcionalmente às áreas de uso privativo, com poucas exceções. Como a norma original (NB-140) é de 1966 (ABNT, 1966), para imóveis construídos até esta data esta regra pode não ser verdadeira (na amostra coletada existem 5.246 unidades construídas anteriormente, cerca de 17% da base), embora seja razoável para o mercado em questão (ABNT, 1966, 1992). Com este procedimento, todos os casos foram suplementados,

[42] Para tanto, inicialmente foi incluído o posicionamento espacial (determinando as coordenadas dos imóveis, X, Y), em outra operação de enriquecimento, descrita na seção seguinte.

gerando-se uma nova variável, Arcopriv2.

Outra forma de cálculo que poderia ser utilizada é uma estimação simplificada, adotando a relação média entre área total e área privativa calculada para os casos válidos como elemento de correção (Hair *et al.*, 1998). Por exemplo, para os 29.291 casos com área privativa maior do que $1m^2$, a média Arcopriv/Artocons é 0,785. Substituindo os 1.986 casos com problemas por 0,785*Artocons, pode-se obter uma correção alternativa. As correlações das variáveis em questão com o valor corrigido monetariamente (Valor) estão apresentadas na Tabela 4, a seguir.

Tabela 4: Comparação das alternativas de correção para a área privativa (Arcopriv)

Variáveis	Correlações	Observações
Valor x Artocons	0,931	-
Valor x Arcopriv	0,853	Original, com 29.291 casos
Valor x Arcopriv2	0,897	Corrigida usando k-NN
Valor x Arcopriv	0,889	Corrigida usando a média

Percebe-se que esta segunda estimativa fornece resultados semelhantes à anterior, sendo ainda de implementação muito mais rápida e simples. Se os erros forem distribuídos aleatoriamente, pode-se esperar que a técnica de correção ou preenchimento de casos omitidos não altere as médias e desvios-padrão dos dados. Porém, a utilização desta forma de correção em modelos de estimação multivariados poderá resultar em tendências, pela existência de correlações entre esta variável e outras variáveis explicativas, e a utilização de uma forma mais complexa, baseada em conhecimento do domínio, pode trazer melhores resultados. Desta forma, adotou-se a correção baseada em k-NN, que gerou Arcopriv2.

c) Padrão construtivo (variável Tipo): esta informação é preenchida pelo contribuinte, sem a conferência do fiscal de tributos. Verificou-se um número elevado de dados incorretos, tais como a indicação de tipo construtivo "Alumínio" e a omissão ou indicação de valores diferentes para o mesmo prédio (mais de 6.000 casos). A correção foi realizada com o procedimento descrito acima: equalização inicial para obter o valor típico de cada prédio usando as informações do mesmo prédio e a estimação com base nos prédios próximos para os prédios sem informação ou com informação incoerente (ver Tabela 5). Isso ocorreu em

1.935 casos, nos quais o apartamento foi o único transacionado no prédio dentro do período pesquisado, sem a possibilidade de utilizar a informação do "mesmo prédio", sendo então estimado pela vizinhança. Após, as categorias foram convertidas em valores numéricos, conforme a escala apresentada na Tabela 5 (ver coluna "Código"). A codificação linear é arbitrária, mas é coerente com a prática dos fiscais e aparentemente de acordo com o mercado. Foram geradas a variável Tipo2, contendo os códigos corrigidos, e um conjunto de variáveis binárias, destinadas a investigar a não linearidade de escalonamento das categorias, todas com valores {0,1}: Pop, Med, Fin e Lux.

Tabela 5: Variáveis relacionadas com o padrão de construção

Categoria	Código	Tipo (Número inicial de casos)	Tipo2 (Número de casos após a correção)	Variáveis Binárias
Alvenaria Precária (A)	1	3	0	-
Alvenaria Popular (B)	2	4.425	4.512	Pop
Alvenaria Média (C)	3	17.230	22.914	Med
Alvenaria Fina (D)	4	3.160	3.649	Fin
Alvenaria Luxo (E)	5	163	146	Lux
Alumínio	-	34	0	-
Climatex, Fiberglass, Telheiro	-	1	0	-
Galeria de Concreto	-	1	0	-
Mista Média (B)	-	1	0	-
Não Informado	-	6.259	0	-
Totais		*31.277*	*31.221*	

d) Ano da construção (variável Ano.Const): os casos claramente incorretos são aqueles com data anterior a 1900 (em função do processo de urbanização da cidade, com verticalização apenas a partir do início do Século XX) ou com valor 0, 19, 78, etc. (ver Figura 15f). Os problemas foram solucionados pela comparação com outros imóveis do mesmo prédio (endereço igual). Em apenas 3 casos com erros não existia outro imóvel no mesmo prédio, sendo utilizados dados da vizinhança para estimação da correção. Foi gerada uma nova variável, denominada Ano.Const2. Esta estratégia pode causar distorções, caso a região tenha uma ocupação heterogênea.

Ao final desta etapa os dados resultantes para as variáveis alteradas foram novamente examinados, verificando-se que as variáveis alteradas tiveram uma diminuição de desvio-padrão, em relação aos dados originais (ver Tabelas 2 e 7).

6.4 ENRIQUECIMENTO

Nesta etapa foram incluídas algumas variáveis sem custo de coleta, mas que geralmente são úteis, algumas decorrentes de transformações na base, outras utilizando indicações da literatura, como no caso da medição de distâncias a alguns pontos.

a) Data (variável Mês): indica o momento da declaração, considerando o início da série no mês de agosto de 1998 (Mês=1) e o final em agosto de 2001 (Mês=37). Para investigar a influência de sazonalidade, foi gerado um conjunto de doze variáveis binárias (Jan={0,1}, ... , Dez={0,1}, com o número de casos correspondentes apresentados na Figura 21a), e para analisar a hipótese de não linearidade da variação temporal, foi gerado um conjunto de 37 variáveis binárias, conforme o mês da transação (M_1={0,1}, ..., M_{37}={0,1}, idem, na Figura 21b).

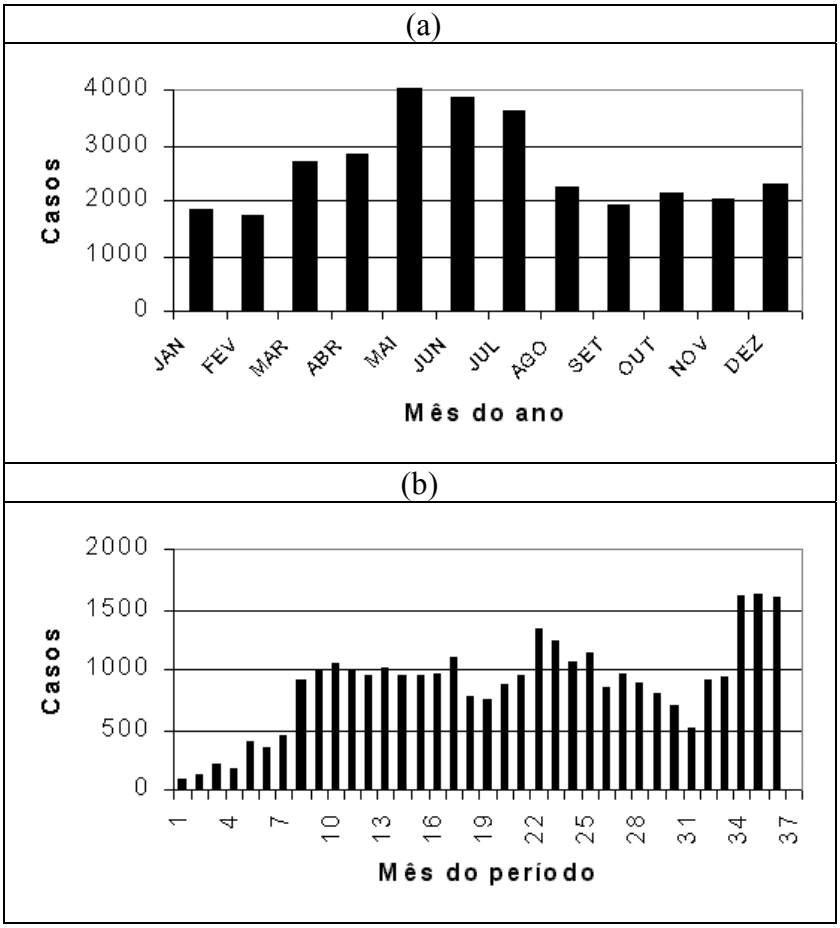


Figura 21 - Distribuição dos casos nos meses do ano (a) e no período da amostra (b)

Percebe-se uma razoável regularidade ao longo dos três anos de dados pesquisados, com incremento no número de dados ao final do período, ao contrário da distribuição pelos meses do ano, que demonstra concentração maior nos meses de outono e inverno (Figuras 21a e 21b). Examinando a base de dados, verificou-se que existe uma grande concentração de imóveis dos conjuntos habitacionais construídos entre 1980 e 1982 nestes períodos de acréscimo, o que decorre do encerramento dos financiamentos, e não de características do mercado, ou seja, parte da aparente sazonalidade deve-se ao final dos prazos de financiamento de um grande contingente de imóveis dos conjuntos habitacionais.

b) Medidas de localização: dada a natureza espacial do mercado imobiliário, é de fundamental importância a consideração da posição de cada um dos imóveis na malha urbana. Esta característica não estava incluída nos dados brutos recebidos. No cadastro municipal existem variáveis que indicam a posição agregada, tais como "SETOR", que vincula o imóvel a uma malha de quadrantes, tipicamente de 1 km x 1 km. Porém, esta informação não é contemplada nas guias de ITBI, sendo incluído apenas o endereço do imóvel.

Uma das formas de estimar os efeitos de localização é através da medição das distâncias dos imóveis a alguns pontos de provável interesse da população, seguindo indicações da literatura. Existem diversos elementos atrativos em uma área urbana, tais como os citados a seguir.

1. Coordenadas do imóvel (variáveis X, Y): para medir a distância aos pólos de interesse inicialmente é necessário identificar a posição dos imóveis no espaço. Com base no endereço, foram determinadas as coordenadas para cada um dos imóveis, usando como referência o prédio da Prefeitura Municipal de Porto Alegre (latitude: 30º1'30'' e longitude: 51º13'40''). Este ponto foi tomado como origem de um eixo cartesiano, sendo a variável X medida no eixo O-L e Y no eixo S-N, ambas em km. Esta foi uma informação que exigiu razoável esforço de coleta, em termos de tempo de análise e custo computacional. Foi utilizada uma base de dados desenvolvida anteriormente que consiste basicamente na descrição das vias urbanas como trechos de retas, utilizando o número do prédio como referência para a interpolação (González, 1996). A utilização de um Sistema de Informações Geográficas exigiria a digitalização de todo o mapa da cidade e posteriormente a identificação da localização dos dados da amostra, o que não era necessário para o restante da análise. Em Porto Alegre, adota-se a numeração dos prédios conforme a distância (em metros) do início da via, atribuindo-se numeração ímpar à

esquerda e par no lado direito da via (regra de origem portuguesa). A distribuição dos imóveis pode ser visualizada rapidamente através de um gráfico tal como o da Figura 22. Verifica-se que aparentemente a distribuição espacial é coerente com o mercado imobiliário, existindo imóveis em todos os bairros, com concentração nas áreas que possuem grandes conjuntos habitacionais e nos bairros com maior verticalização. A área dos círculos indica a quantidade de imóveis em uma mesma localização (mesmo endereço), e serve como indicação do grau de atividade do mercado em cada local. De posse das coordenadas, é fácil calcular distâncias a quaisquer pontos de interesse, tais como *shopping centers* ou áreas de lazer, usando a fórmula euclidiana de distância no plano (distância=$((x_1-x_2)^2+(y_1-y_2)^2)^{0.5}$).

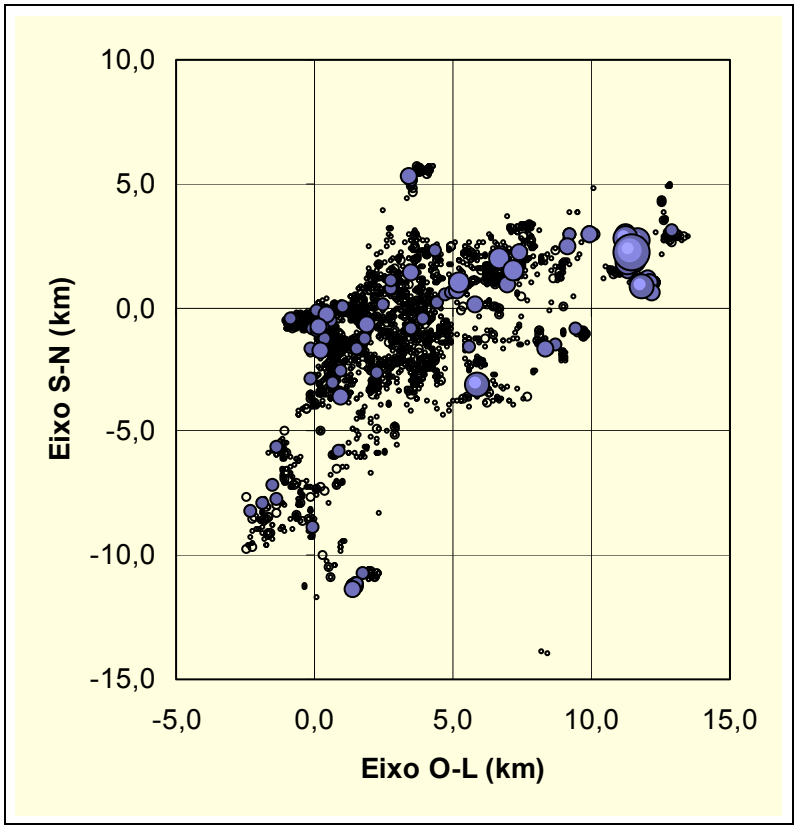


Figura 22 - Posição dos imóveis da amostra - a origem (0,0) corresponde ao prédio da Prefeitura (“centro da cidade”, com latitude 30°1’30’’ e longitude 51°13’40’’).

2. Classificação para a localização (variável Bairro): inicialmente foi identificado o bairro em que se situa cada imóvel, conforme limites definidos por legislação municipal,

utilizando seu posicionamento espacial. Em seguida, foi gerada uma classificação para os bairros, visando estabelecer uma primeira aproximação para as diferenças espaciais de preços (Bairro). Esta medida é qualitativa, sendo definida subjetivamente pelo autor, com base em um conjunto critérios, tais como qualidade de vizinhança, padrão construtivo médio, acessibilidade, transporte e padrão sócio-econômico dos residentes. A classificação e a distribuição dos imóveis por bairro são as seguintes (Tabela 6).

Tabela 6: Distribuição dos imóveis por bairro e classificação dos bairros (variável Bairro)

Local	Casos	Bairro	Local	Casos	Bairro
Aberta dos Morros	475	19	Mont Serrat	628	56
Agronomia	1	5	Navegantes	105	18
Auxiliadora	473	51	Nonoai	393	19
Azenha	611	33	Parque São Sebastião	248	33
Bela Vista	644	100	Partenon	912	26
Boa Vista	505	56	Passo da Areia	751	26
Bom Fim	700	43	Passo das Pedras	5	17
Bom Jesus	171	22	Petrópolis	1590	58
Camaquã	285	24	Praia de Belas	104	59
Cascata	5	15	Protásio Alves	446	20
Cavalhada	366	26	Restinga	39	30
Cel. Aparício Borges	17	23	Rio Branco	1007	52
Centro	3295	34	Rubem Berta	2931	15
Chácara das Pedras	134	43	Santa Cecília	333	40
Cidade Baixa	1230	37	Santa Maria Goretti	86	17
Cristal	487	23	Santa Teresa	233	19
Cristo Redentor	770	27	Santana	955	40
Farrapos	133	44	Santo Antônio	462	25
Farroupilha	48	41	São Geraldo	258	26
Floresta	708	37	São João	600	37
Glória	101	31	Sarandi	540	25
Higienópolis	311	41	Teresópolis	186	27
Humaita	352	19	Três Figueiras	46	89
Independência	451	62	Tristeza	349	41
Ipanema	29	37	Vila Assunção	17	76
Jardim Botânico	505	35	Vila Floresta	35	17
Jardim Carvalho	203	20	Vila Ipiranga	493	22
Jardim do Salso	258	22	Vila Jardim	69	16
Jardim Itu	602	24	Vila João Pessoa	77	23
Jardim Lindoia	290	51	Vila Nova	692	18
Medianeira	306	28	Vila São José	53	19
Menino Deus	1555	37	Vila São Pedro	83	24
Moinhos de Vento	474	89			

3. Centro da cidade (variável Centro): na literatura, é comum a referência ao CBD (*Central Business District*), ou seja, o centro histórico-comercial da cidade, que representaria o ponto de concentração de empregos, comércio e serviços. Em uma cidade mais desenvolvida, tal como Porto Alegre, podem existir vários pólos importantes. De qualquer forma, é uma medida importante, representada aqui pela distância dos imóveis até a origem das coordenadas (ver Figura 22).

4. Variável Comércio: como elementos adicionais de comércio, além do centro da cidade, foram utilizadas as distâncias aos dois *shopping centers* e aos seis hipermercados existente em Porto Alegre à época da coleta dos dados, adotando-se a menor distância, ou seja, a distância ao *shopping* ou ao hipermercado mais próximo de cada imóvel. Na Figura 23a são indicados os posicionamentos relativos destes pólos, com os pontos de área maior indicando os dois *shopping centers*. Foram considerados os seguintes pontos (Tabela 7).

Tabela 7: Pólos de comércio considerados

Denominação	Tipo	Coordenadas	
		X	Y
SC Iguatemi	*Shopping center*	6,500	0,500
SC Praia de Belas	*Shopping center*	0,010	-2,250
Bourbon Assis Brasil	Hipermercado	4,200	2,500
Bourbon Country	Hipermercado	6,350	0,900
Carrefour Partenon	Hipermercado	6,150	-3,550
Carrefour Passo da Areia	Hipermercado	5,500	1,700
BIG Sarandi	Hipermercado	8,150	3,650
BIG Cristal	Hipermercado	-1,500	-6,000

5. Variável Lazer: foram considerados os parques urbanos públicos e uma praia tradicional na cidade. Não foram incluídos clubes privados, por serem de acesso restrito, embora possam ter efeitos quanto à paisagem e qualidade do ar. Os pontos de referência para estes elementos foram os apontados na Tabela 8, os quais podem ser visualizados na Figura 23b (a praia está indicada pelo ponto de menor dimensão).

Tabela 8: Pólos de lazer considerados

Denominação	Tipo	Coordenadas X	Coordenadas Y
Jardim Botânico	Área verde	4,850	-2,300
Parque Moinhos de Vento (Parcão)	Área verde	2,750	0,200
Parque Farroupilha (Redenção)	Área verde	1,100	-0,900
Parque Chico Mendes	Área verde	11,250	0,500
Parque da Harmonia (Maurício S. Sobrinho)	Área verde	-0,700	1,250
Parque Marinha do Brasil	Área verde	-0,300	-3,000
Parque Mascarenhas de Morais	Área verde	3,850	5,200
Praia de Ipanema	Praia	0,100	-12,000

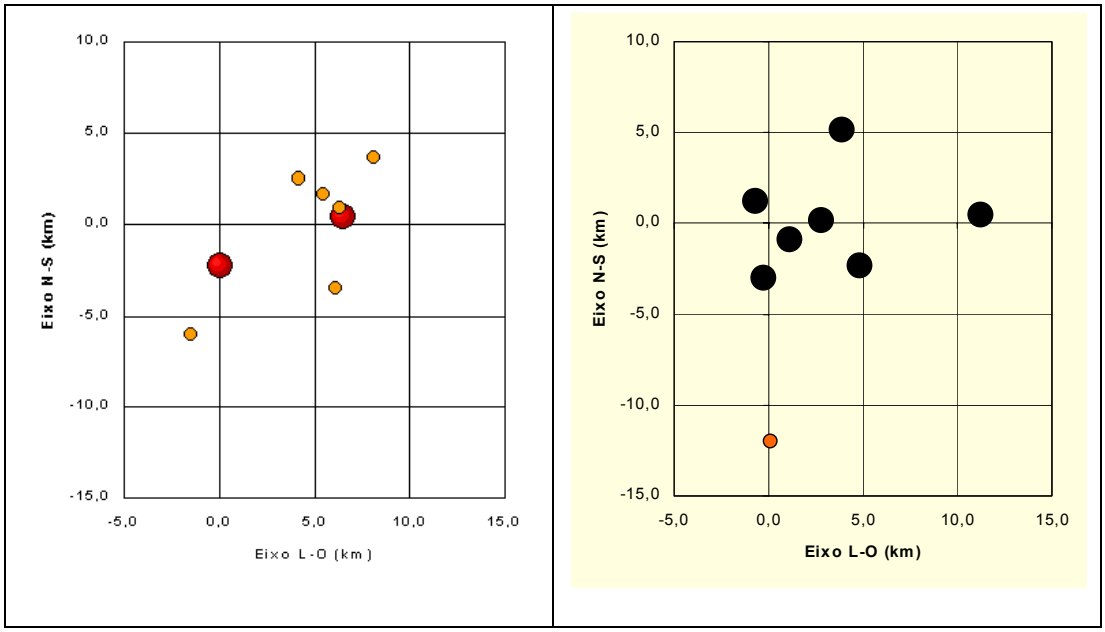


Figura 23 - Posicionamento das referências de (a) Comércio e (b) Lazer

6. Vista panorâmica (variável Vista): neste grupo também foi incluída uma variável destinada a verificar a influência da vista panorâmica para o Rio Guaíba, que é considerado um símbolo de Porto Alegre. Nota-se uma valorização elevada para os imóveis que possuem esta característica. Utilizou-se uma variável binária (Vista={0,1}), pois neste caso a distância é menos importante. A estimação desta variável foi realizada de forma bastante simplificada: em vistoria de campo na área próxima ao Rio, foram identificados alguns prédios, adotando-se valor 1 para todos os apartamentos do prédio que estavam na amostra, englobando 49 casos. Um procedimento mais preciso seria de

avaliar caso a caso, pois no mesmo prédio podem existir imóveis voltados para o lado oposto. Porém esta análise dependeria da consulta aos projetos, utilizando ainda o número do apartamento, o qual não estava disponível na amostra obtida, e também não seria viável para grande quantidade de dados. A variável foi incluída para evidenciar a necessidade de consideração desta questão nas análises sobre o mercado imobiliário. Por outro lado, embora o efeito de uma vista panorâmica sobre os preços seja conhecido (Bond *et al.*, 2002), há o problema da dificuldade de justificação, no caso de avaliação para tributação, por ser um elemento de julgamento subjetivo.

7. Localização baseada em medidas de erro (variável Local) - foi determinada uma medida baseada nos erros apurados em modelos desenvolvidos adiante, na fase de seleção de variáveis e de casos, usando modelos construídos sem as variáveis que medem a localização, de forma semelhante ao explorado por Gallimore *et al.* (1996), González *et al.* (2002a), McCluskey *et al.* (2000). As medidas para cada imóvel foram ajustadas por superfícies de resposta, sendo finalmente convertidas para uma escala [0,100].

c) Relação entre a área total e a parcela transferida do terreno (variável Fração): existe uma grande variação da área total do terreno (de 73,92 m^2 a 86.970,27 m^2). Examinando a posição dos imóveis, percebe-se que os imóveis situados em terrenos pequenos tendem a situar-se em locais próximos ao centro da cidade, onde a urbanização é antiga e os terrenos já eram ou tornaram-se pequenos ao longo do tempo, por causa da grande procura e valorização. Por outro lado, terrenos grandes provavelmente estão relacionados a grandes projetos de urbanização (habitação de massa), longe do centro da cidade, o que os associa a aspectos desvalorizantes (em termos de padrão construtivo e sócio econômico do entorno ou mesmo distância de transporte), mas há também uma relação possível com a qualidade de vida, através de questões como afastamento lateral, ventilação e insolação. No caso, além da área do terreno, adotou-se a razão entre a área total do terreno e a área transmitida com o apartamento, conhecida no mercado como "fração ideal". Espera-se que esta relação seja constante dentro do prédio em função da prática e da norma que regula o setor (ABNT, 1992). A variável foi calculada a seguinte relação (Equação 25):

Fração=Artrterr2/Artoterr2 (Equação 25)

d) Relação entre as áreas privativa e total (variável Razão): no caso de Artocons e Arcopriv, há uma elevada colinearidade (r=0,972). A diferença entre elas representa a área de uso comum correspondente ao apartamento. Embora sejam fortemente correlacionadas, ambas são interessantes, pois em alguns casos a área de uso comum pode variar bastante, conforme a região da cidade, a época da construção, a disponibilidade de equipamentos de lazer no prédio, etc. A área condominial provavelmente não é valorizada da mesma forma que a área privativa. Para investigar este efeito, foram geradas duas variáveis, a primeira com a diferença e a segunda com a relação entre elas, ou seja, conforme as relações das Equações 26 e 27:

Condomínio=(Artocons–Arcopriv2) (Equação 26)

Razão=Arcopriv2/Artocons. (Equação 27)

e) Refinamento da variável Tipo2: as variáveis qualitativas discretas geralmente utilizam escalas arbitrárias, como é o caso de Tipo2, por exemplo (ver Tabela 5). Entretanto, a exploração de modelos demonstrou que a escala adotada para Tipo2 não era adequada. A conjugação com as variáveis binárias (Pop,...,Lux) indicou que a escala (1,...,5) provocava uma distorção em relação à categoria superior (imóveis com padrão "Alvenaria Luxo"), embora sendo adequada para as demais categorias. Para corrigir este problema, foi gerada uma nova variável (Tipo3), na qual o padrão Luxo é representado pelo valor 6, passando a variável a contar com as categorias (1,2,3,4,6).

f) Variável Idade: é diferença entre o ano da transação (obtido da variável Data) e o ano da construção do imóvel (Ano.Const2), lembrando que neste caso o ano de construção é representado pelo ano do licenciamento, nem sempre coincidente com o término da construção, identificando na realidade a "idade fiscal" do imóvel. A distribuição das duas variáveis é a apresentada na Figura 24. Percebe-se a concentração dos casos próximo aos 20 anos de idade, o que reflete os registros de transmissão decorrentes do encerramento do

período de financiamento, em diversos conjuntos habitacionais (construídos no final dos anos 70 e no início dos anos 80). A participação expressiva dos imóveis de conjuntos habitacionais (cerca de 25% do total) pode provocar tendências nos modelos construídos, se a fração no mercado não for correspondente.

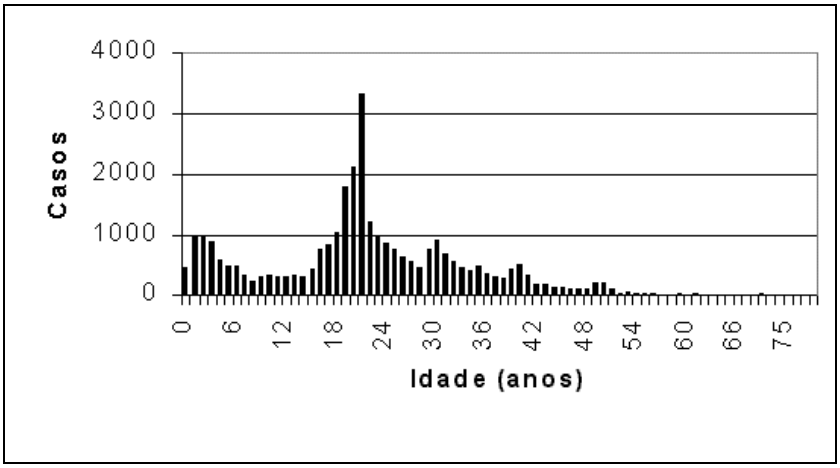


Figura 24 - Distribuição dos imóveis segundo a Idade

g) Valores das imóveis (variáveis Valor e Unitário): são os valores corrigidos monetariamente, resultantes da atualização dos valores indicados pelo fiscal (Vl.Atribuido), visando a comparabilidade entre os valores ocorridos em momentos diferentes. Pindyck e Rubinfeld (2002) afirmam que é necessário corrigir os preços quando há inflação, trabalhando com preços reais. Não existem índices específicos de correção monetária para o mercado imobiliário de Porto Alegre, sendo então utilizado o Índice Geral de Preços – Disponibilidade Interna (IGP-DI), calculado pela Fundação Getúlio Vargas, corrigindo os valores da data da declaração até agosto de 2001 e gerando a variável Valor. Em seguida, foi calculado o valor unitário dos imóveis usando a área total dos apartamentos: Unitário=Valor/Artocons (ver Figuras 17 a 20).

6.5 SELEÇÃO DE DADOS RELEVANTES (REDUÇÃO DA BASE DE DADOS)

Gerada a base de dados inicial, com limpeza dos dados com erros grosseiros, preenchimento de valores omitidos e enriquecimento com informações importantes, ainda é necessária uma

outra etapa, de análise mais detalhada e de aperfeiçoamento da base de dados, antes da aplicação dos procedimentos de geração de modelos ou estimação.

Esta etapa envolve a análise do desempenho ou relevância das variáveis e dos casos em conjunto. Nas fases anteriores a utilidade das informações incluídas para a solução do problema não é verificada diretamente, sendo que o processo é guiado pelo conhecimento anterior (conhecimento do domínio e indicações de importância pelos resultados de outras aplicações ou recomendações fundamentadas na teoria), geralmente com a análise isolada (univariada) das variáveis. As variáveis incluídas na análise seguinte são as decorrentes do processamento anterior (Tabela 9).

Tabela 9: Variáveis da base de dados após tratamento das fases anteriores

Variável	Tipo	Unidade	Mínimo	Máximo	Média	Desvio-padrão
Artoterr	Contínua	m^2	1,00	91.416,00	4.516,52	10.160,85
Artoterr2	Contínua	m^2	73,92	86.970,27	4.514,35	10.142,45
Artrterr	Contínua	m^2	0,01	941,40	40,89	33,99
Artrterr2	Contínua	m^2	0,05	823,68	40,88	33,20
Fração	Contínua	-	$7,50*10^{-5}$	1,00	$4,40*10^{-2}$	$6,32*10^{-2}$
Artocons	Contínua	m^2	10,59	1.208,84	93,11	72,60
Arcopriv	Contínua	m^2	0,00	760,52	70,11	51,38
Arcopriv2	Contínua	m^2	5,67	760,52	70,41	50,40
Condomínio	Contínua	m^2	0	448,32	22,70	26,44
Razão	Contínua	-	0,08	1,00	0,78	0,11
Tipo2	Discreta	-	2	5	2,98	0,53
Tipo3	Discreta	-	2	6	2,99	0,55
Pop	Binárias	-	0	1	0,114	0,35
Méd	Binárias	-	0	1	0,734	0,44
Fin	Binárias	-	0	1	0,117	0,31
Lux	Binárias	-	0	1	0,005	$6,5*10^{-2}$
Ano.Const	Discreta	Ano	0	2001	1979,98	55,63
Ano.Const2	Discreta	Ano	1920	2001	1978,29	12,07
Idade	Discreta	Anos	0	81	21,56	12,10
X	Contínua	km	-2,95	13,49	3,74	3,70
Y	Contínua	km	-14,00	5,74	-1,13	3,34
Bairro	Discreta	-	5	100	35,83	17,35
Centro	Contínua	km	0,02	16,28	5,24	3,55
Comércio	Contínua	km	0,09	10,64	2,27	1,30
Lazer	Contínua	km	0,15	8,53	1,75	1,06
Vista	Binária	-	0	1	$1,57*10^{-3}$	$3,96*10^{-2}$
Local	Contínua	-	1	100	48,64	8,32
Mês	Discreta	mês	1	37	21,57	9,39

Jan, ..., Dez	Binárias	-	0	1	-	-
Mês1, ..., Mês37	Binárias	-	0	1	-	-
Valor	Contínua	R$	2.271,61	1.603.531,26	76.374,64	91.397,31
Unitário	Contínua	R$/m^2	52,84	3.327,32	728,99	285,55

Como referido anteriormente, a remoção de dados irrelevantes, que não acrescentam informação significativa ou qualidade à análise, pode reduzir o custo computacional e o custo de modelagem (tempo do analista) e facilitar o entendimento dos modelos obtidos, pois modelos mais compactos provavelmente serão mais facilmente compreendidos pelos usuários.

Há três formas de seleção ou redução de dados, envolvendo: (1) variáveis (redução horizontal, nas colunas da planilha), realizando classificação de importância das variáveis, composição e remoção das redundantes ou irrelevantes; (2) casos (redução vertical, nas linhas da planilha), com a identificação de elementos que aparentemente são válidos, se analisados frente ao todo, mas que vistos em detalhe revelam-se incorretos (possíveis *outliers*), e remoção de casos redundantes; e (3) redução da variação dos valores das variáveis (suavização ou discretização), especialmente útil para os métodos lógicos (tais como sistemas de regras clássicas ou árvores de decisão). No caso, como a tarefa é de predição ou estimação de valores e as técnicas a serem utilizadas não exigem a redução, não foram realizadas operações de suavização ou discretização (exceto as transformações de variáveis discretas em conjuntos de variáveis binárias que, em última análise, pode ser considerada uma forma de discretização).

A seleção de atributos e de casos é um processo iterativo, pois o conhecimento reunido durante a análise das variáveis permite classificar mais adequadamente os erros (ou explicar melhor os casos), e a remoção de erros permite identificar mais claramente a importância das variáveis.

Na abordagem por envoltório, tanto para variáveis quanto para casos, foram utilizadas as técnicas de regressão (com modelos hedônicos tradicionais e na forma de superfícies de resposta) e redes neurais, pois as peculiaridades de desempenho de cada técnica podem influir na seleção do melhor subconjunto de variáveis[43]. As superfícies de resposta foram incluídas

[43] As redes neurais foram ajustadas com o *Stuttgart Neural Network Simulator* (SNNS), desenvolvido pela

em função do desconhecimento relativo (nesta fase da análise) sobre a relevância ou correção das variáveis de localização incluídas.

6.5.1 Seleção dos atributos (redução horizontal)

A classificação da potencialidade das variáveis foi realizada inicialmente com o mecanismo de filtro. Uma das formas mais comuns de filtro é a matriz de correlações, utilizada para identificar os relacionamentos das variáveis independentes com as dependentes e as variáveis independentes aproximadamente colineares (variáveis que contém aproximadamente a mesma informação). A matriz de correlações das variáveis independentes com as dependentes está na Tabela 10.

Tabela 10: Matriz de correlações das variáveis independentes com as dependentes[44]

Variáveis	Valor	Unitário	Variáveis	Valor	Unitário
Artoterr	-0,116	-0,113	Tipo3	0,623	0,673
Artoterr2	-0,116	-0,112	Pop, ..., Lux*	0,702	0,681
Artrterr	0,427	0,081	X	-0,102	-0,291
Artrterr2	0,427	0,081	Y	0,071	0,065
Fração	0,233	0,099	Bairro	0,545	0,560
Artocons	0,931	0,410	Centro	-0,218	-0,445
Arcopriv	0,853	0,374	Comércio	-0,098	-0,378
Arcopriv2	0,897	0,381	Lazer	-0,127	-0,115
Condomínio	0,846	0,400	Vista	0,060	0,115
Razão	-0,287	-0,275	Local	0,214	0,342
Ano.Const	0,072	0,077	Mês	-0,152	-0,330
Ano.Const2	0,338	0,303	Jan, ..., Dez*	0,103	0,227
Idade	-0,347	-0,325	Mês1, ..., Mês37*	0,181	0,410
Tipo2	0,600	0,680			

Percebe-se que existem algumas variáveis independentes com correlação maior com uma ou com outra variável dependente, tais como Artocons e Centro, enquanto outras têm correlação

Universidade de Stuttgart, que é um *software* de uso livre, disponível na internet em ftp://ftp.informatik.uni-stuttgart.de/pub/SNNS.

[44] Para as variáveis marcadas com * trata-se de correlação múltipla, excluindo a primeira variável binária em cada caso (correlação em módulo).

semelhante com todas as opções de variável dependente, como é o caso da Idade.

No caso de conjuntos de variáveis binárias, emprega-se a correlação múltipla para comparar o conjunto com a correlação da variável de origem. Em vista das características deste tipo de conjunto, uma das categorias deve ser removida, para evitar problemas estatísticos (combinação linear exata). Por exemplo, na comparação entre Tipo2, Tipo3 e Pop,...,Lux, a variável binária Pop não foi incluída e as correlações com a variável Valor foram, respectivamente, 0.600, 0.623 e 0.702, indicando que a transformação desenvolvida em Tipo2 (gerando Tipo3) foi benéfica e que o conjunto de binárias tem relação mais forte com os dois formatos do valor do que as duas variáveis discretas. Para as demais opções de variáveis dependentes os resultados foram semelhantes, neste caso. As demais variáveis alteradas nas fases anteriores também apresentaram desempenho igual ou superior às originais (Tabela 10).

Em alguns casos, as correlações podem indicar a opção entre variáveis concorrentes, como nos casos das variáveis corrigidas ou complementadas em relação às originais ou de variáveis aproximadamente colineares. Estas variáveis podem ser entendidas como mutuamente excludentes, pois representam medidas alternativas para o mesmo atributo.

As correlações entre variáveis independentes indicam algumas combinações interessantes (Tabela 11). Na análise de regressão, no caso de correlação forte entre as variáveis independentes, geralmente exclui-se uma delas, permanecendo as variáveis que apresentam maior correlação com a variável dependente (ou maior valor da estatística t). Deve-se especificar um limite de correlação mínimo para indicar um relacionamento importante. Na etapa de investigação da relevância dos atributos, os modelos de regressão podem incluir variáveis com correlações maiores, em função do caráter exploratório e foi utilizado um limite de $|r|=0,8$, ou seja, apenas as variáveis independentes com correlações maiores do que 0,8 não foram incluídas simultaneamente.

Outra forma de redução de dimensionalidade pode ser desenvolvida através da análise de componentes principais. Em função das correlações verificadas entre algumas variáveis independentes, há o risco de multicolinearidade (ver Tabela 11), e a aplicação de componentes principais permite obter um novo conjunto de variáveis, as quais não são correlacionadas entre si (ou seja, a correlação de qualquer combinação de fatores é nula). O método foi aplicado sobre as variáveis apresentadas na Tabela 11, com exceção dos três

conjuntos de variáveis binárias (para Tipo3 e Mês). A inclusão destes conjuntos de variáveis binárias ampliou o número de fatores, sem vantagens na análise posterior. Foi aplicada rotação Varimax, visando facilitar a interpretação dos fatores.

Tabela 11: Matriz de correlações entre as variáveis independentes[45]

Variáveis	Artocons	Razão	Tipo3	Idade	Bairro	Centro
Artoterr2	-0,124	0,114	-0,084	-0,061	-0,231	0,278
Artrterr2	0,497	0,015	0,128	-0,218	0,089	0,208
Fracao	0,314	-0,042	0,116	-0,022	0,214	-0,180
Artocons	1,000	-0,303	0,543	-0,295	0,490	-0,189
Arcopriv2	0,972	-0,111	0,497	-0,226	0,458	-0,180
Condominio	0,893	-0,620	0,543	-0,379	0,472	-0,177
Razão	-0,303	1,000	-0,341	0,342	-0,285	0,208
Tipo3	0,543	-0,341	1,000	-0,266	0,533	-0,393
Pop, ...,Lux*	0,614	0,352	1,000	0,382	0,545	0,476
Ano.Const2	0,288	-0,335	0,252	-0,998	0,071	0,300
Idade	-0,295	0,342	-0,266	1,000	-0,081	-0,288
X	-0,082	0,155	-0,230	-0,237	-0,326	0,730
Y	0,073	0,013	0,100	0,024	0,092	-0,093
Bairro	0,490	-0,285	0,533	-0,081	1,000	-0,530
Centro	-0,189	0,208	-0,393	-0,288	-0,530	1,000
Comércio	-0,076	0,161	-0,320	-0,062	-0,157	0,519
Lazer	-0,097	0,015	-0,048	-0,159	-0,350	0,458
Vista	0,022	-0,001	0,064	-0,048	0,042	-0,023
Local	0,176	-0,124	0,198	-0,050	0,388	-0,364
Mês	-0,102	0,126	-0,214	0,069	-0,142	0,162
Jan, ..., Dez*	0,085	0,106	0,204	0,205	0,151	0,211
Mês1, ..., Mês37*	0,133	0,187	0,296	0,096	0,210	0,281

A interpretação dos fatores é baseada nas "cargas", que indicam a intensidade dos relacionamentos das variáveis com cada um dos fatores (Tabela 12). O exame do conjunto de variáveis com as maiores cargas em cada fator (em módulo) indica o provável significado do fator. Por exemplo, o Fator1 é fortemente relacionado com as medidas de área, enquanto que o Fator3 está vinculado às medidas de localização. Entretanto outros fatores, tal como o Fator5, não têm uma vinculação clara com as variáveis de origem.

Foi gerado um conjunto de 8 novas variáveis, correspondentes ao produto dos fatores pelas

[45] Idem nota anterior.

variáveis, sendo nomeadas como Fator1,...,Fator8, as quais foram investigadas na abordagem envoltório. A dificuldade com esta técnica é a perda da explicabilidade das variáveis, em relação ao usuário, ou seja, os fatores obtidos não têm vinculação clara com a realidade, como têm área, idade ou outros elementos convencionais. Para aplicações em que a precisão é critério prioritário, porém, podem ser interessantes, e devem ser testadas.

Tabela 12: Fatores obtidos por componentes principais, com rotação Varimax

Variáveis	Fator 1	Fator 2	Fator 3	Fator 4	Fator 5	Fator 6	Fator 7	Fator 8
Artoterr2	0,403	0,109	0,282	0,078	-0,011	0,737	-0,004	0,031
Artrterr2	0,007	-0,008	0,752	-0,093	-0,185	-0,128	-0,067	0,125
Fração	0,146	-0,019	-0,300	-0,042	-0,006	0,827	-0,014	-0,076
Artocons	0,929	0,132	-0,051	-0,016	0,026	0,244	-0,005	0,000
Arcopriv2	0,887	0,024	-0,022	0,020	0,035	0,314	-0,009	0,054
Condomínio	0,859	0,317	-0,098	-0,081	0,005	0,072	0,005	-0,103
Razão	-0,325	-0,491	0,231	0,274	0,093	0,232	-0,007	0,238
Tipo3	0,634	0,173	-0,150	-0,433	0,008	-0,091	-0,033	0,091
Ano.Const2	0,163	0,947	0,051	0,054	0,032	0,058	-0,009	0,065
Idade	-0,171	-0,945	-0,056	-0,019	-0,020	-0,056	0,006	-0,058
X	-0,105	0,243	0,393	0,354	0,736	0,059	0,108	-0,037
Y	0,104	-0,088	-0,118	-0,123	0,907	-0,044	-0,115	-0,007
Bairro	0,611	-0,019	-0,499	-0,120	-0,075	-0,072	-0,009	0,050
Centro	-0,253	0,344	0,621	0,467	0,199	0,110	0,176	-0,021
Comércio	0,012	0,055	0,273	0,788	-0,275	-0,064	0,175	-0,111
Lazer	-0,202	0,220	0,537	-0,359	0,225	0,183	0,430	-0,021
Vista	0,024	0,051	-0,050	0,013	-0,020	-0,045	0,013	0,928
Local	0,127	0,081	-0,552	-0,162	-0,094	0,042	0,001	0,168
Mês	-0,134	-0,019	-0,099	0,518	0,173	0,033	-0,051	0,093
Soma	*3,717*	*1,107*	*1,144*	*1,116*	*1,753*	*2,409*	*0,592*	*1,408*

A etapa seguinte da seleção de atributos é a análise utilizando a abordagem envoltório. Em face da quantidade de variáveis, existe uma profusão de modelos alternativos, inviabilizando o teste de todos. A inclusão de todas as variáveis em um modelo inicial, removendo progressivamente as de menor desempenho (estratégia *backward*) também não é conveniente, pela complexidade do modelo inicial, pelo acréscimo de tempo de processamento e pelas interações existentes entre as variáveis. Por exemplo, sabe-se antecipadamente que algumas variáveis não devem participar conjuntamente do mesmo modelo de regressão, em virtude de elevada colinearidade entre elas. A estratégia adotada foi de definir um subconjunto inicial de

variáveis com base nas informações disponíveis, testando progressivamente a inclusão das demais e convergindo para o grupo final de variáveis através da abordagem envoltório. As correlações com as variáveis independentes e os modelos exploratórios testados indicaram que a variável dependente com o valor total (Valor) gerava modelos com melhor desempenho do que os modelos com o valor unitário (Unitário), para as três técnicas.

A seleção das variávies dependentes pode ser afetada pela escala dos valores, pois os valores totais são naturalmente maiores do que os unitários, provocando um problema de geometria nas medidas de erro. É interessante adotar mais de um indicador de erro, para diminuir o risco de erro na escolha das variáveis. Por outro lado, nas redes neurais, que exigem a normalização dos dados, as duas variáveis foram convertidas para um intervalo similar (de 0 a 1), e os resultados foram similares ao dos modelos de regressão. De qualquer forma, a maioria dos trabalhos consultados na literatura de mercado imobiliário utiliza valores totais, e aparentemente os agentes do mercado imobiliário raciocinam também em termos de valores totais, reforçando a opção pelos valores totais.

A seleção das variáveis independentes decorreu do teste de quatro estratégias de seleção, cada uma delas apontando um modelo hedônico de preços. As alternativas testadas foram as seguintes (Equações 28 a 31):

a) Seleção pelo conhecimento anterior: Valor=f(Artocons, Condomínio, Tipo3, Idade, Bairro, Mês) (Equação 28)

b) Seleção por filtro 1 - variáveis com maior correlação com a variável dependente: Valor=f(Artrterr2, Artocons, Condomínio, Tipo3, Idade, Bairro) (Equação 29)

c) Seleção por filtro 2 (componentes principais): Valor=f(Fator1, ...,Fator8) (Equação 30)

...,Fator8)

d) Seleção utilizando o algoritmo *stepwise* (usando regressão): (Equação 31)

Valor=*f*(Artocons, Razão, Tipo3, Idade, Bairro, Mês)

Investigando estes modelos alternativos com as ferramentas de estimação, verificou-se que o desempenho dos modelos (28) a (31) foi semelhante para as três ferramentas, com pequena vantagem para os modelos estimados com redes neurais (Tabela 13). A análise levou em conta o desempenho geral dos modelos através do exame do coeficiente de determinação (R^2_a e R^2) e do erro padrão (EP) dos modelos e as indicações de importância de cada variável independente (usando a estatística *t* ou o índice GI, conforme o caso). Em geral, nos modelos com superfícies de resposta as coordenadas tiveram menor importância (menor significação estatística) do que as variáveis incluídas nos modelos alternativos. Nas redes neurais foram utilizadas arquiteturas 6:6:1 com os modelos das equações (28), (29) e (31) e 8:8:1 quando foram testados os fatores (equação 30)[46].

Tabela 13: Investigação dos modelos alternativos - variável dependente: Valor

Modelo	Regressão hedônica convencional		Regressão hedônica com superfícies		Redes neurais[47]	
	R^2_a	EP	R^2_a	EP	R^2	EP
Equação 28	0,894	29.748	0,896	29.440	0,925	25.112
Equação 29	0,894	29.823	0,896	29.478	0,924	25.207
Equação 30	0,871	32.821	0,884	31.173	0,911	27.564
Equação 31	0,897	29.361	0,901	28.759	0,921	24.985

Embora exista um razoável equilíbrio entre as alternativas, os resultados indicam um

[46] Nas redes neurais, a indicação de um modelo I:J:K indica uma rede composta por três camadas, com I neurônios na camada de entrada, J neurônios ocultos e K neurônios na camada de saída.

[47] Para as redes neurais o coeficiente de determinação (R^2) foi calculado através do quadrado da correlação entre a variável dependente original e os valores estimados pela rede.

desempenho ligeiramente superior do modelo da equação (31). Assim, este foi o modelo escolhido para ser a base dos testes seguintes. A partir deste modelo verificou-se a influência da inclusão das demais variáveis e da alteração de formato das variáveis iniciais.

As variáveis dependentes foram investigadas na forma linear, bem como com a transformação através do procedimento de Box-Cox. Trata-se de uma primeira etapa de modelagem, mas deve ser buscado um modelo de razoável ajustamento para então aplicar esta técnica. A investigação das variáveis Valor e Unitário indicou respectivamente transformações com expoentes 0.75 e 0.5, ou seja, a substituição por $Valor^{3/4}$ e $Unitário^{0,5}$.

A seleção das outras variáveis independentes foi realizada incrementalmente, testando-se a inclusão de cada uma das variáveis que não participaram do modelo (31), inclusive o efeito de transformações numéricas sobre elas (usando Box-Cox). As variáveis com transformações não demonstraram desempenho superior às variáveis originais e não justificam o aumento de complexidade causado pela adição de um elemento não linear no modelo. No caso das variáveis binárias, verificou-se o desempenho dos modelos com a inclusão do conjunto (sempre excluindo-se uma das variáveis binárias do conjunto). Por exemplo, para a comparação de Tipo3 e Pop,...,Lux, partindo do modelo (31), já utilizando a transformação na variável dependente, foram desenvolvidos os modelos apresentados na Tabela 14[48]. Neste caso verifica-se que o conjunto de variáveis binárias apresenta um desempenho ligeiramente superior nas três ferramentas, indicando a substituição para os modelos seguintes.

Tabela 14: Exemplo de seleção de variáveis - variável dependente: $Valor^{3/4}$

Modelo	Regressão hedônica convencional		Regressão hedônica com superfícies		Redes neurais	
	R^2_a	EP	R^2_a	EP	R^2	EP
Equação 31 com Tipo3	0,917	25.743	0,921	25.525	0,925	25.549
Eq.31 com Med, Fin, Lux	0,918	25.573	0,922	25.218	0,918	26.306

Repetindo este procedimento, o modelo linear inicial foi aprimorado incrementalmente, com

[48] Quando existe uma transformação na variável dependente, é necessário calcular a transformação inversa das estimativas antes de calcular os erros, visando manter a comparabilidade quanto à variável dependente. Nos resultados apresentados na Tabela 14 e nos modelos seguintes foi realizado este cálculo.

revisão constante dos resultados. Em geral, o desempenho das três técnicas foi semelhante, em termos de seleção dos atributos. Assim, e visando à facilidade de comparação dos modelos a serem desenvolvidos posteriormente (capítulo 7), buscou-se selecionar um único subconjunto de atributos para todas as ferramentas. A seleção de atributos evoluiu continuamente, inclusive durante a seleção de casos. Os resultados obtidos indicam o seguinte conjunto de variáveis para a análise posterior, ou seja, o modelo final foi o seguinte (Equação 32):

$$\text{Valor}^{3/4} = f(\text{Artocons, Razão, Med, Fin, Lux, Idade, Bairro, Centro, Comércio, Vista, Mês}) \quad \text{(Equação 32)}$$

6.5.2 Seleção de casos (redução vertical)

A seleção de casos foi desenvolvida em paralelo com o aperfeiçoamento ou teste de variáveis (evoluindo conjuntamente com os modelos apresentados na seleção de atributos). Esta é a segunda etapa de limpeza e detecção de elementos discordantes, buscando agora o ajuste fino. O volume de dados impede uma análise detalhada, caso a caso, que permita identificar individualmente as causas dos erros, como usualmente se realiza nas amostras pequenas. O processo deve ser ágil, porém garantindo que não sejam removidos dados apenas por serem diferentes dos demais. Levado ao extremo, este processo poderia resultar em uma amostra mínima, altamente homogênea, mas com pouca capacidade de generalização.

O exame dos dados precisa ser multivariado, pois os preços imobiliários são afetados por diversas características (os imóveis são bens compostos). Assim, o exame dos casos através de filtros é difícil, pois exige a pré-definição de limites (no caso, limites para os valores dos imóveis), indicando como mais favorável uma abordagem tipo envoltório. A relevância dos casos em relação ao mercado, investigada através dos modelos construídos, foi avaliada através da sua adaptação aos modelos provisórios, melhorados incrementalmente com o aumento de conhecimento sobre os dados e sobre os próprios modelos. Naturalmente, há influência do conhecimento anterior (modelos ajustados em outros trabalhos empíricos e

teoria), durante a investigação.

As técnicas utilizadas para examinar os erros foram os erros padronizados, erros absolutos e a medida D2 de Mahalanobis. Esta última é uma medida multidimensional, que indica a distribuição de cada observação ao centro médio das observações, adotando-se o nível de 0,001 para o teste de significância.

Foram utilizados os modelos provisórios para verificar imóveis com valor potencialmente discrepante das condições normais de mercado, com a exclusão apenas dos dados marcados na primeira fase (erros grosseiros). Como o ajustamento do modelo pode ser afetado pela presença de dados com erros, a investigação foi realizada em etapas, excluindo progressivamente os erros que atingiam o limite de 4 desvios-padrão (aumentando o ajustamento do modelo diminui o desvio-padrão dos erros, atingindo casos que não estavam além do limite na iteração anterior, mas estavam no limiar), em duas ou três fases, geralmente.

Durante o processo de desenvolvimento da seleção de atributos, os modelos podem apresentar como problemas (indicados por erros grandes) alguns casos que na verdade são normais mas que não foram bem compreendidos pelos modelos, possivelmente ainda incompletos. Por este motivo, os dados foram apenas marcados, sem a exclusão, sendo re-testados periodicamente.

A seleção de casos relevantes é afetada pela ferramenta de aprendizagem empregada na abordagem envoltório. Verificou-se que alguns dos elementos marcados como discordantes (potenciais *outliers*) foram apontados pelas três ferramentas, enquanto outros apareceram em apenas uma ou duas indicações. As quantidades de casos em cada situação foram as seguintes (Tabela 15).

Tabela 15: Número de *outliers* detectados

Situação	Casos
Pré-selecionados (erros removidos na limpeza)	56
Regressão com o modelo (32)	461
Regressão – modelo (32) com superfícies de resposta	486
Redes Neurais com o modelo (32)	441

Uma das alternativas, neste caso, é de remover apenas os casos que forem discordantes em dois ou mesmo nos três modelos (só excluindo com a certeza da discordância), ou seja,

eliminar apenas os que atingiram 4 desvios-padrão com as três ferramentas. Outra opção é remover todos os casos indicados por ao menos uma ferramenta, visando maior segurança quanto aos erros. Em função da ampla disponibilidade de dados e do nível de tolerância adotado (4 desvios-padrão) esta segunda opção foi adotada.

Os dados selecionados para posterior análise foram 30.363 casos, ou seja, foram marcados 858 casos, além dos 56 já marcados anteriormente. Tendo em vista o caráter exploratório do estudo, e visando ainda à comparabilidade dos resultados, bem como facilitar o desenvolvimento da parte empírica, optou-se por utilizar uma base única. A adoção de arquivos distintos de treinamento e teste para cada ferramenta provocaria uma profusão de arquivos, com pequena vantagem para os objetivos do trabalho. Entretanto, provavelmente todas as técnicas seriam beneficiadas, em termos de desempenho, pelo uso de conjuntos de variáveis e de casos particularizados. Ademais, o limite de 4 desvios-padrão foi escolhido também para manter parte da variabilidade dos casos, permitindo investigar a sensibilidade das ferramentas a esta variabilidade.

6.6 INVESTIGAÇÃO DOS PRESSUPOSTOS BÁSICOS DA REGRESSÃO

Nas etapas apresentadas os dados foram preparados, obtendo informações sobre o formato do modelo e identificando casos problemáticos (com erros ou potenciais *outliers*). Ainda devem ser investigados os demais pressupostos da análise de regressão: heterocedasticidade (variância do erro inconstante), normalidade e independência dos erros. Estes elementos foram investigados através de gráficos, construídos com os erros calculados com o modelo 16 (ver Tabela 16, modelo e)[49].

A Figura 25a apresenta a comparação entre valores observados e estimados. A distribuição revela razoável aproximação entre os dois. Na Figura 25b, apresenta-se a distribuição do erro padronizado em função dos valores estimados padronizados. Percebe-se uma ligeira tendência de heterocedasticidade, com crescimento dos erros simultaneamente ao crescimento dos valores estimados.

[49] Estes gráficos foram construídos com 30.363 casos (excluindo os *outliers*).

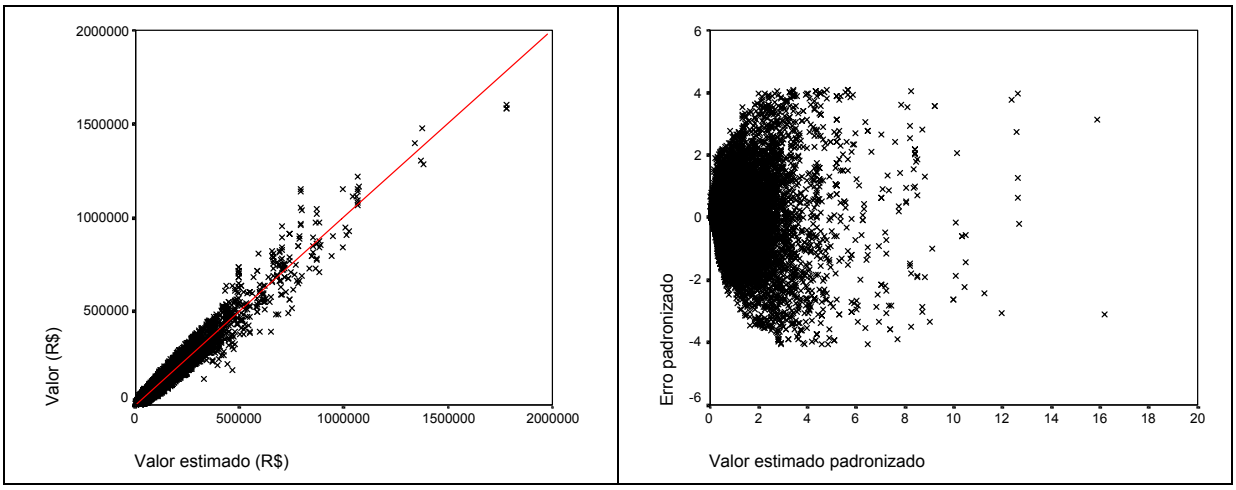


Figura 25: Distribuição dos valores estimados: (a) comparado com Valor e (b) comparado com o erro padronizado

Em seguida foram construídos os gráficos de investigação da normalidade, consistindo de histograma e gráfico de probabilidade Normal acumulada (Figuras 26a e 26b). A análise destes gráficos indica que os erros têm uma distribuição aproximadamente Normal. Há sinais de fuga à Normalidade, mas são toleráveis, em função do tamanho da amostra e da variabilidade associada com ela (diferenças entre os tipos de imóveis). Pode-se concluir que os dados, após a preparação, permitem a construção de modelos adequados aos pressupostos estatísticos.

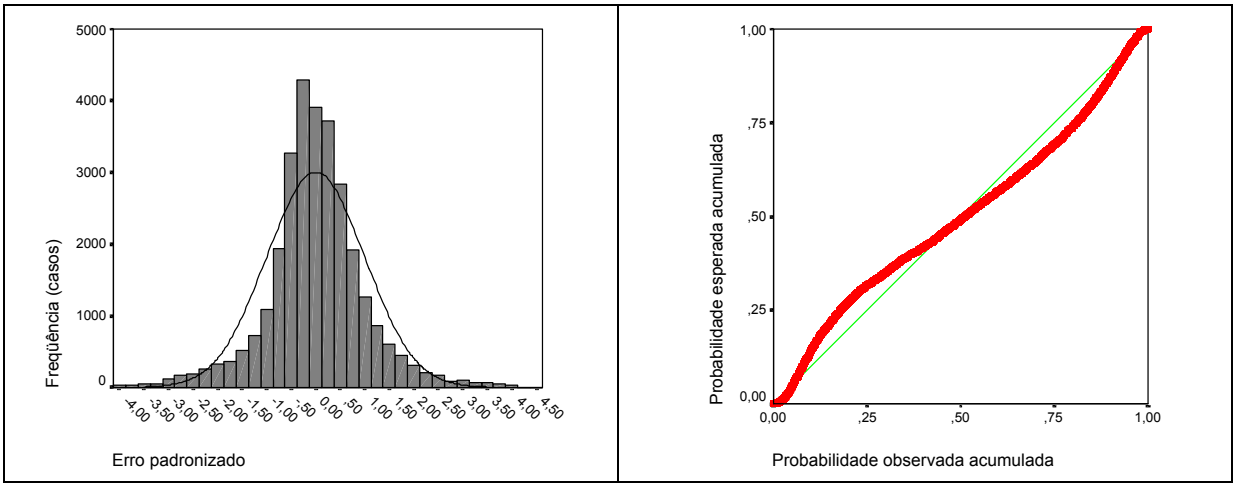


Figura 26: Gráficos de investigação da normalidade: (a) Histograma e (b) Probabilidade acumulada

6.7 PROGRESSOS OBTIDOS NA PREPARAÇÃO DOS DADOS

Os dados obtidos continham expressiva parcela de erros. Como demonstração dos progressos obtidos na fase de preparação dos dados, apresenta-se na Tabela 16 os resultados obtidos para os modelos de regressão hedônica tradicional, incluindo a descrição dos modelos utilizados, o número de casos (n) e de variáveis independentes (k), e os dois parâmetros convencionais de verificação de resultados: o coeficiente de determinação ajustado (R^2_a) e o erro padrão (EP). A Figura 27 apresenta graficamente a evolução destes dois parâmetros de ajustamento. O erro padrão sofreu uma redução de cerca de 70% na primeira etapa de limpeza, e nova redução, de cerca de 35%, na seleção de informações relevantes, especialmente seleção de casos. O crescimento do coeficiente de determinação ajustado seguiu padrões similares. Os modelos apresentados foram desenvolvidos levando em conta as questões convencionais de multicolinearidade, significação dos coeficientes (sinais e valores absolutos) e análise dos gráficos de erros, bem como limites máximos de significância das variáveis, de 5%, conforme a norma de avaliações (ABNT, 1989), ou seja, os modelos apresentados podem ser considerados os modelos de melhor ajuste em cada momento ou condição da base, ao final de cada etapa.

Tabela 16: Evolução dos modelos gerais durante a preparação dos dados

Modelo	Número de casos (n)	Número de variáveis (k)	Ajustamento (R^2_a)	Erro padrão (EP)
Base de dados original [a]	31.277	4	0,307	109.454
Base após limpeza inicial [b]	31.221	4	0,858	31.897
Base após correção e complementação das variáveis [c]	31.221	5	0,887	30.754
Base após enriquecimento [d]	31.221	10	0,896	29.413
Base após seleção de variáveis e eliminação de *outliers* [e]	30.363	11	0,951	19.018

Modelos:

a) Vl_Atribuido=f(Artoterr,Artrterr,Artocons,Ano.Const)

b) Valor=f(Artoterr,Artrterr,Artocons,Ano.Const)

c) Valor=f(Artoterr2,Artrterr2,Artocons, Tipo2,Ano.Const2)

d) Valor=f(Artoterr2,Artrterr2,Artocons, Razão, Tipo2, Idade, Bairro, Lazer, Vista, Mês)

e) Valor=f(Artocons, Razão, Med, Fin, Lux, Idade, Bairro,Centro, Comércio, Vista, Mês)$^{4/3}$

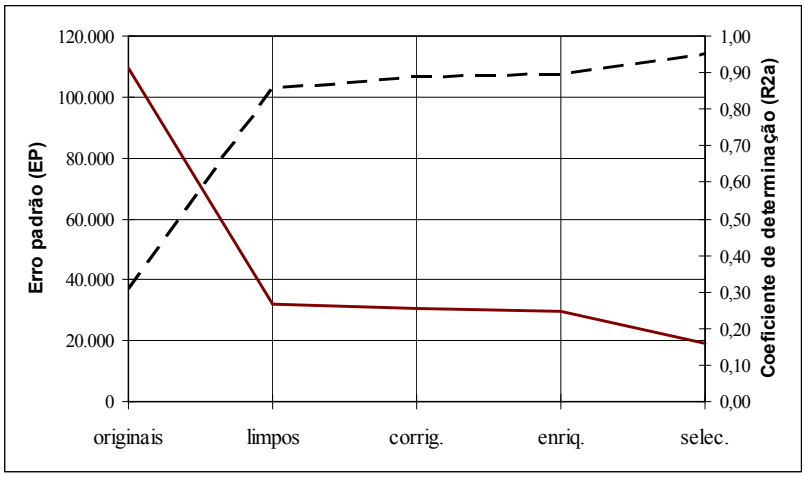


Figura 27: Evolução do erro padrão e do coeficiente de determinação ajustado dos modelos gerais

É importante acompanhar a evolução dos indicadores de desempenho dos modelos. Pode ocorrer que, mesmo com as operações de tratamento dos dados, os modelos não apresentem evoluções significativas. Neste caso, podem existir problemas nos dados, tais como multicolinearidade ou omissão de variáveis importantes, não coletadas ou substituídas por variáveis *proxy* com fraco desempenho. Para os modelos apresentados, verifica-se que há grande evolução dos parâmetros apresentados (R^2_a e EP), evidenciando a importância da preparação dos dados. As estatísticas do conjunto final de dados são as apresentadas na Tabela 17, a seguir.

Tabela 17: Características das variáveis após a preparação dos dados (com base em 30.363 casos)

Variável	Unidade	Mínimo	Máximo	Média	Desvio padrão
Valor	R$	2.271,61	1.603.531,00	74.310,64	86.765,11
Artocons	m^2	10,59	1208,84	90,74	69,30
Razão	-	0,08	1,00	0,78	0,11
Tipo2	-	2	6	2,98	0,54
Pop	-	0	1	0,15	0,35
Med	-	0	1	0,74	0,44
Fin	-	0	1	0,11	0,31
Lux	-	0	1	$4{,}25*10^{-3}$	$6{,}50*10^{-2}$
Idade	anos	0	81	21,65	12,03
X	km	-2,95	13,49	3,76	3,72
Y	km	-14,00	5,74	-1,14	3,37
Bairro	-	4	88	31,21	14,96

Centro	km	0,02	16,28	5,28	3,57
Comércio	km	0,09	10,64	2,2753	1,3120
Vista	-	0	1	$1,25*10^{-3}$	$3,54*10^{-2}$
Mês	mês	1	37	21,67	9,36

7 GERAÇÃO DE MODELOS PARA AVALIAÇÃO INDIVIDUAL E COLETIVA

7.1 CONSIDERAÇÕES INICIAIS

A partir dos dados preparados conforme exposto no Capítulo 6, foram explorados modelos para as três situações típicas de avaliação, identificados adiante como modelos gerais ou "de mercado", de avaliação em massa e de avaliação individual, sendo os dois primeiros considerados como modelos de avaliação coletiva. Como foi visto no Capítulo 5, os modelos de predição gerados pertencem a cinco categorias, em termos das técnicas empregadas na modelagem: regressão tradicional e com superfícies matemáticas, redes neurais com explicação por regras geradas com lógica difusa, modelos aditivos generalizados com

parâmetros determinados por algoritmos genéticos e extração de regras difusas da base de dados, também utilizando algoritmos genéticos.

A comparação entre os modelos gerados para o tipo Mercado e para os dados agrupados por área foi desenvolvida pela união das estimativas de cada sub-modelo em uma série única, compondo o que seria uma planta de valores (modelo de avaliação em massa) e permitindo a comparação destes dois modelos, exceto no caso do sistema de regras difusas, o qual gera um modelo único.

A análise do desempenho dos modelos durante a fase de desenvolvimento dos modelos foi baseada no erro cometido, sendo este erro medido pela diferença entre as estimativas dos modelos e os valores observados (coletados). Em função de serem adotadas diversas técnicas, buscou-se indicadores comuns a todas, usando mais de um indicador de erro. Os indicadores utilizados no treinamento (desenvolvimento dos modelos) foram o erro padrão da estimativa (EP) e o erro absoluto percentual médio (EA). Para comparação dos modelos finais, posteriormente foram calculados os coeficientes de dispersão (COD) e de correlação entre os valores observados e os estimados pelos modelos ($r_{y.y}^{h}$) e a quantidade de erros menores do que 5% e 10% (como indicação de erros pequenos) e maiores do que 50% (erros grandes)[50]. Os indicadores escolhidos são encontrados normalmente nos trabalhos empíricos que envolvem o mercado imobiliário e o exame de técnicas de modelagem. Todos os erros foram calculados sobre o valor final, após as transformações inversas na variável dependente, ou seja, sobre o valor que seria apresentado ao usuário final, em Reais. As comparações de modelos foram realizadas com os resultados do conjunto de teste, em todos os casos.

7.2 MODELOS PARA AVALIAÇÃO COLETIVA

Os dados para o modelo de Mercado resultam diretamente dos trabalhos desenvolvidos

[50] O erro padrão é calculado por $EP=[(\Sigma(Y-Y^h)^2/(n-2)]^{0,5}$. Com o crescimento da amostra, o erro padrão aproxima-se da raiz do erro quadrado médio ($RMSE=[(\Sigma(Y-Y^h)^2/n]^{0,5}$) e também do desvio-padrão dos erros amostrais ($s=[(\Sigma(\varepsilon-\bar{\varepsilon})^2/(n-1)]^{0,5}$), onde Y é o preço observado, Y^h é a estimativa dos modelos, n é o tamanho da amostra, ε é o erro ($\varepsilon=Y-Y^h$) e $\bar{\varepsilon}$ é o erro médio. O erro absoluto percentual médio é calculado por $EA=|(Y-Y^h)/Y|/n*100\%$ e o coeficiente de dispersão é o desvio percentual médio absoluto das razões Y^h/Y em relação à mediana das razões: $COD=(\Sigma|Y^h/Y-m|/Y)/n*100\%$, onde m é a mediana das razões.

durante a preparação dos dados, conforme descrito no capítulo anterior. Para o modelo de avaliação em massa, foi utilizada a segmentação em sub-mercados através de clusterização.

As indicações dos modelos iniciais e da bibliografia levaram à segmentação da base de dados em parcelas (relativamente) mais homogêneas através da clusterização usando a área total (Artocons), em vista de sua elevada correlação com a variável dependente (r>0,90). Verificou-se a influência da subdivisão usando outros atributos, ou combinações de atributos, sem vantagens significativas sobre o critério adotado. Também foram testadas alternativas para a quantidade de subdivisões, testando 3, 4 e 5 parcelas. Foi preferida a divisão em três parcelas, pelos resultados similares às outras e pela vantagem da simplicidade (menor quantidade de modelos). Como ocorre em geral, neste caso há um dilema entre simplicidade e precisão.

A base foi dividida em três categorias, reunindo os imóveis em grupos com área total de [10,59; 138,07), [138,07; 355] e (355; 1.208,84]. Estes grupos foram denominados de Pequenos, Médios e Grandes, em função dos tamanhos dos apartamentos incluídos em cada um deles (ver Tabela 18).

A segmentação também é útil para evitar heterocedasticidade no caso da regressão, pois diminui a variação total da variável dependente em cada sub-modelo (por exemplo, para o conjunto total de dados, esta variação é de R$ 2.271,61 a R$ 1.603.531,26 - ver Tabela 17, item 6.7), além de permitir modelos potencialmente mais ajustados aos dados.

Em seguida foi realizada uma revisão dos casos excluídos como *outliers* durante a preparação dos dados, investigando se os modelos compostos com as parcelas da base geral proporcionavam explicação melhor para estes casos. Verificou-se que a divisão em parcelas relativamente mais homogêneas permitiu especificar modelos mais ajustados aos dados, mas sem evolução quanto aos casos inicialmente marcados como *outliers*. O conjunto de dados não foi alterado, permanecendo com 30.363 casos válidos para análise e os dados foram divididos em três grupos e em conjuntos de treinamento e teste, conforme exposto na Tabela 18. A coluna identificada como “Excluídos” indica os casos com erros não recuperáveis, eliminados na primeira fase de limpeza (ver item 6.2).

Tabela 18: Distribuição dos dados após a clusterização

Grupo	Área total	Total	Excluídos	*Outliers*	Treino	Teste	%teste
MERCADO	10,59–1.208,84 m^2	31.277	56	858	24.289	6.074	20,00
MASSA							
Pequenos	<138,07 m^2	26.928	37	455	21.148	5.288	20,00
Médio	138,07–355 m^2	3.923	8	367	2.838	710	20,01
Grandes	>355 m^2	426	11	36	303	76	20,05

A união dos modelos construídos para os grupos de imóveis Pequenos, Médios e Grandes gera o modelo de avaliação em massa. Os modelos investigados têm o formato de Modelos Hedônicos de Preços (MHP), com o formato geral apresentado na Equação 1, incluindo atributos que representam as características do próprio imóvel (área, idade, etc.), outros descrevendo a localização, através de medidas qualitativas para a vizinhança e de medidas de distância para a acessibilidade, e ainda sobre a data da transação. Os atributos incluídos no modelo (Equação 33) são os descritos e selecionados no capítulo 6.

$$\text{Valor} = (a_0+a_1*\text{Artocons}+a_2*\text{Razão}+a_3*\text{Med}+a_4*\text{Fin}+a_5*\text{Lux}+a_6*\text{Idade}+ \quad \text{(Equação 33)}$$
$$a_7*\text{Bairro}+a_8*\text{Centro}+a_9*\text{Comércio}+a_{10}*\text{Vista}+a_{11}*\text{Mês})^{4/3} + \varepsilon$$

Os coeficientes a_i (i=1,...,11) são os preços hedônicos correspondentes às variáveis X_i, enquanto que a_0 é o intercepto (ou constante) do modelo hedônico e ε é o erro ou desvio do modelo em relação ao valor original (da amostra). O formato adotado inicialmente para a variável dependente foi $\text{Valor}^{3/4}$, em vista da análise por Box-Cox desenvolvida anteriormente, e os coeficientes dos modelos foram estimados pelas várias técnicas exploradas, sendo que o modelo de estimação é o seguinte (Equação 34):

$$\text{Valor}^{3/4} = b_0+b_1*\text{Artocons}+b_2*\text{Razão}+b_3*\text{Med}+b_4*\text{Fin}+b_5*\text{Lux}+b_6*\text{Idade}+ \quad \text{(Equação 34)}$$
$$b_7*\text{Bairro}+b_8*\text{Centro}+b_9*\text{Comércio}+b_{10}*\text{Vista}+b_{11}*\text{Mês} + \varepsilon'$$

Os modelos utilizados (Equações 33 e 34) têm um aspecto aditivo, tais como os modelos encontrados na maioria dos trabalhos revisados na literatura relacionada com o mercado

imobiliário. Também poderiam ser utilizados modelos multiplicativos, explorando a interação entre variáveis independentes, os quais podem fornecer modelos mais precisos, mas que geralmente são mais complexos, em relação à estimação e interpretação.

7.2.1 Análise de regressão – modelos hedônicos tradicionais

A regressão foi desenvolvida inicialmente, sendo utilizada como referência para a análise das outras técnicas de modelagem, em função de ser a técnica mais utilizada atualmente em avaliação coletiva de imóveis, sendo bem conhecida e razoavelmente mais simples do que as demais, inclusive em termos de disponibilidade de *software* (é encontrada até no Microsoft Excel). No caso, novamente foi utilizado o pacote estatístico SPSS.

Foram utilizados os limites de significância recomendados pela norma brasileira de avaliações, NBR-5676 (ABNT, 1989), que consistem em 5% de significância para as variáveis (teste de hipótese usando a estatística *t*) e para o modelo (análise de variância com a distribuição *F*). Os resultados estão apresentados na Tabela 19, a seguir.

Há uma certa regularidade nos coeficientes, em termos de valores absolutos e sinais, embora também existam variações entre os grupos, indicando as diferenças no comportamento dos agentes do mercado em cada caso.

Em alguns casos, há diferenças nos modelos devido a peculiaridades estatísticas, como elevada correlação encontrada entre Centro e Comércio ou entre Med e Fin (variáveis binárias para o padrão construtivo) apenas no grupo 3, ou da própria amostra de dados, que não apresenta casos com Vista=1 neste mesmo grupo.

Tabela 19: Modelos hedônicos tradicionais – modelo de Mercado e modelos agrupados por área – variável dependente: Valor$^{3/4}$

Variável	Modelo							
	Mercado		Pequenos		Médios		Grandes	
	Coeficiente	t	Coeficiente	t	Coeficiente	t	Coeficiente	t
Intercepto	515,724	11,1	367,574	9,5	1.206,505	3,7	2.062,602	1,3
Artocons	35,260	404,7	38,137	248,6	32,771	80,7	27,228	25,8
Razão	883,674	18,7	809,235	21,8	579,913	2,6	6.251,419	3,8
Med	510,359	32,1	533,832	45,6	818,974	3,2		

Fin	1.734,842	69,5	1.342,890	61,8	2.346,930	9,1	3.376,341	7,3
Lux	4.251,930	53,8	2.221,619	25,2	5.108,435	9,4	7.448,704	12,8
Idade	-27,447	-56,5	-24,337	-63,7	-46,415	-22,8	-165,182	-10,3
Bairro	20,615	47,4	16,472	42,6	26,114	20,3	35,645	5,5
Centro	-43,290	-21,4	-40,784	-25,7	-121,903	-10,5	-668,033	-6,6
Comércio	-90,022	-20,5	-97,632	-29,0	-55,820	-2,2		
Vista	1.618,714	12,2	1.420,388	11,0	1.733,946	5,7		
Mês	-19,864	-39,0	-17,271	-43,5	-26,921	-12,2	-75,689	-5,6
R^2_a	0,950		0,884		0,846		0,888	
F_{calc}	41.969		14.581		1.416		299	

Os sinais dos coeficientes são os esperados (conforme conhecimento anterior), para a maioria das variáveis. Todos os atributos apresentados na Tabela 19 tiveram significância inferior a 5%. No caso do tempo (Mês), o coeficiente negativo indica a progressiva diminuição dos preços no período coberto pela amostra. O ajustamento dos modelos é similar, com coeficientes de determinação (ajustados para os graus de liberdade) entre 0,846 e 0,95, indicando que de 85% a 95% das variações da variável dependente podem ser explicadas pelos modelos apresentados. Considerando apenas o coeficiente de determinação, o ajustamento do modelo Mercado seria superior aos modelos parciais. O exame dos pressupostos, porém, revelou que os modelos segmentos atendem melhor aos requisitos de normalidade e de homocedasticidade. Os valores calculados para os testes dos modelos (F_{calc}) também foram expressivamente superiores aos mínimos exigidos pela Norma, mesmo ocorrendo alguma variação em função dos graus de liberdade (ABNT, 1989).

Foram examinados os resultados gerais dos modelos, através dos erros percentuais (EA) e dos erros padronizados (EP), conforme Tabela 20. Verifica-se que a subdivisão dos dados em três modelos permitiu um aperfeiçoamento dos resultados, gerando estimativas que levaram a um erro padrão 15% menor e um erro percentual 5% menor para os dados de teste, para o modelo de avaliação em massa, em relação ao de Mercado.

Tabela 20: Resultados obtidos com os modelos de regressão múltipla

Modelo	Treino			Teste		
	EA	EP	Casos	EA	EP	Casos
MERCADO	18,45	18.957	24.289	19,14	19.854	6.074

Pequenos	18,65	11.083	21.148	18,88	11.114	5.288
Médios	13,37	30.294	2.838	13,63	30.175	710
Grandes	11,38	73.908	303	10,12	75.868	76
MASSA	17,94	16.802	24.289	18,16	16.911	6.074

7.2.2 Superfícies de resposta

Os modelos com superfícies de resposta foram desenvolvidos com o acréscimo das potências das coordenadas (X, Y) às outras variáveis. As superfícies investigadas incluíram termos até o sexto grau, com resultados variáveis entre os grupos de dados, indicando diferenças entre os sub-mercados (Tabela 21). A inclusão dos termos foi testada pelo algoritmo *stepwise*, também utilizando o SPSS, com os mesmos limites do caso anterior.

Os coeficientes calculados também têm os sinais e valores esperados (contribuições para o valor estimado), para as variáveis iniciais. Para os termos das coordenadas, há expressivas variações entre os modelos, sendo que algumas combinações não aparecem em nenhum dos modelos, tais como XY^2 ou X^4. Os ajustamentos e testes são similares aos obtidos para os modelos de regressão "tradicionais", com ligeira vantagem para os modelos com superfícies nos coeficientes de determinação.

Tabela 21: Modelos hedônicos com superfícies – modelos de Mercado e modelos agrupados por área – variável dependente: $Valor^{3/4}$

	Modelo							
	Mercado		Pequenos		Médios		Grandes	
Variável	Coeficiente	t	Coeficiente	t	Coeficiente	t	Coeficiente	t
Intercepto	369,112	7,8	184,040	5,1	1228,235	3,9	2482,833	1,6
Artocons	35,239	404,1	37,755	272,9	32,630	81,6	27,135	25,8
Razão	957,321	19,9	885,638	26,5	672,659	3,0	6510,384	4,0
Med	492,503	27,9	542,180	46,8	787,387	3,1		
Fin	1733,279	66,4	1371,059	68,8	2309,529	8,9	3336,894	7,3
Lux	4320,717	54,6	2360,026	30,0	4980,645	9,2	7352,755	12,6
Idade	-26,214	-55,1	-23,077	-66,8	-46,811	-23,5	-169,354	-10,5
Bairro	19,365	40,3	14,704	37,6	25,057	18,5	34,955	5,5
Centro							-797,448	-6,9
Comércio	-88,508	-12,2	-84,248	-15,7	-108,604	-3,4		
Vista	1645,282	12,4	1412,421	12,3	1358,386	4,5		
Mês	-19,688	-38,3	-16,606	-46,4	-27,219	-12,5	-75,916	-5,7

Y	-50,394	-10,9	-73,303	-15,7	-149,865	-3,5		
X^2					-14,675	-7,6		
XY					35,096	3,4		
Y^2	-15,003	-8,4	-20,540	-12,4	-42,136	-4,6		
X^3			0,408	3,4				
X^2Y			0,827	7,5				
Y^3	2,244	6,1	2,354	6,6				
X^2Y^2			0,619	7,0	-12,548	-4,8		
XY^3	-0,495	-3,4	-0,778	-5,7				
Y^4	0,491	8,5	0,747	8,7				
X^5	$-2,22*10^{-3}$	-9,5	-0,0153	-5,4				
X^3Y^2	-0,0369	-6,4	-0,0724	-8,3	2,812	4,4		
X^2Y^3	0,0734	4,7	0,0963	6,7	-1,657	-4,7	1,117	2,2
XY^4	-0,0496	-3,3	-0,115	-5,2				
Y^5			0,0402	8,0				
X^6			$8,67*10^{-4}$	5,1				
X^5Y	$6,248*10^{-4}$	5,2						
X^4Y^2					-0,167	-4,0		
X^3Y^3					0,240	4,6		
X^2Y^4	$9,87*10^{-3}$	5,5	$9,03*10^{-3}$	7,2				
XY^5			$-3,65*10^{-3}$	-3,5				
Y^6	$-1,84*10^{-3}$	-7,8			$1,575*10^{-3}$	3,5		
R^2_a	0,951		0,886		0,851		0,889	
F_{calc}	21.232		7.907		812		270	

Os resultados apresentados na Tabela 22 também indicam equilíbrio com os modelos anteriores, com pequenas vantagens em alguns indicadores ou modelos, de parte a parte. Este equilíbrio pode ser tomado como uma indicação de que os modelos de regressão convencionais, apresentados na Tabela 19, são adequados, em termos da representação da localização. Em outras palavras, as superfícies são uma forma de investigação sobre a autocorrelação espacial: se houvesse uma parcela de erros correlacionados com o espaço, nos modelos anteriores, provavelmente as diferenças seriam significativas a favor dos modelos com superfícies, mais flexíveis neste aspecto.

Tabela 22: Resultados obtidos com as superfícies de resposta

Modelo	Treino			Teste		
	EA	EP	Casos	EA	EP	Casos
MERCADO	17,57	18.891	24.289	18,20	19.801	6.074
Pequenos	17,45	10.926	21.148	18,29	11.532	5.288
Médios	13,07	29.781	2.838	13,54	29.986	710
Grandes	11,24	73.476	303	10,19	76.720	76

MASSA	16,86	16.581	24.289	17,63	17.162	6.074

7.2.3 Redes neurais com explicação por regras difusas

As redes neurais foram treinadas com o *Stuttgart Neural Network Simulator* (SNNS). Foram utilizadas as mesmas variáveis dos modelos de regressão. Testes iniciais demonstraram que a variável dependente transformada ($Valor^{3/4}$) tinha melhor desempenho do que na forma linear (Valor), um resultado interessante, pois em princípio a rede neural deveria compensar internamente a não-linearidade das variáveis.

As redes neurais foram treinadas utilizando o módulo de retropropagação com termo de momento no SNNS (*BackPropMomentum*). Os parâmetros das redes (taxas de aprendizagem e de momento e o número de ciclos) foram ajustados por tentativas sucessivas, utilizando-se o critério de parada antecipada (encerramento do treinamento quando o erro do conjunto de teste começa a aumentar) para evitar *overfitting* (Haykin, 2001).

As redes foram compostas por três camadas, incluindo entrada (11 neurônios), saída (com um neurônio) e uma camada oculta (Figura 28). A camada oculta foi testada com J={I, 2I, 3I}, sendo J o número de neurônios ocultos e I o número de neurônios de entrada, com os melhores resultados sendo obtidos com igual número de neurônios de entrada e ocultos (I=J). Em função das peculiaridades dos dados do terceiro grupo (imóveis Grandes), foram testadas redes com 8, 10 e 11 variáveis, obtendo-se melhores resultados com as redes de 8 entradas. Também foram investigados modelos com 11 neurônios ocultos (rede 8-11-1), neste caso sem vantagens.

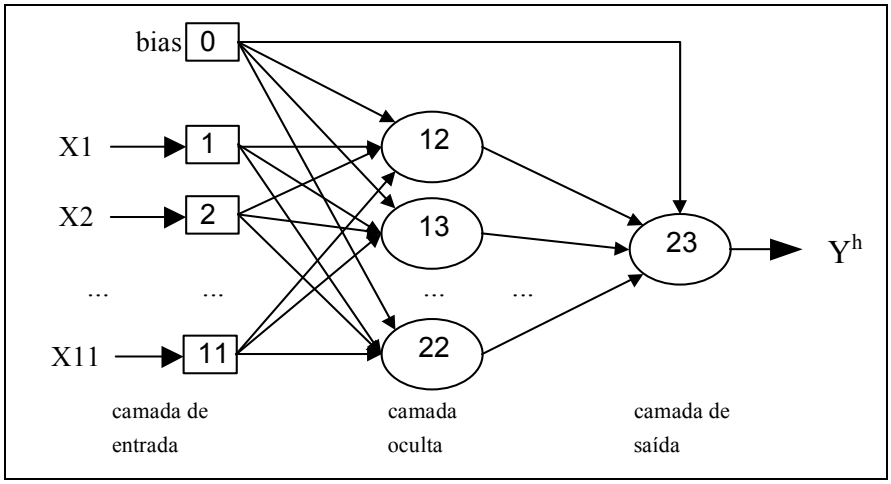


Figura 28: Representação esquemática das redes neurais utilizadas

Em função da posterior extração de regras, foram utilizadas funções de ativação tanh(X/2), as quais são numericamente equivalentes às funções logísticas$_{[-1,1]}$, pois estas funções têm algumas vantagens em termos de simplificação de derivadas, facilitando a extração (Cechin, 1998). As redes calculadas com funções logísticas$_{[0,1]}$ forneceram resultados semelhantes em termos de erros de treino e teste, embora com parâmetros de aprendizagem e momento muito distintos (em torno de 1,0 e 0,9, respectivamente) e menor quantidade de treinamento (número de ciclos)[51]. Os parâmetros das redes treinadas estão na Tabela 23, a seguir.

Tabela 23: Parâmetros das redes neurais treinadas – modelos de Mercado e modelos com dados agrupados por área

	Modelo			
Parâmetro	Mercado	Pequenos	Médios	Grandes
Função de ativação	Tanh (x/2)			
Algoritmo	*BackPropMomentum*			
Topologia (I-J-K)	11-11-1	11-11-1	11-11-1	8-8-1
Taxa de aprendizagem (η)	0,01	0,1	0,01	0,01
Taxa de momento (α)	0,01	0,01	0,01	0,1
Ciclos de treinamento	850	850	5010	7000

Os tempos de treinamento foram razoáveis, variando de 1 a 20 minutos, em geral, afastando um dos problemas apontados pela literatura (alguns autores relataram tempos de treinamento

[51] O formato destas três funções é similar, em forma de um "S" alongado. A função tanh(x/2) é a tangente hiperbólica, com x em radianos. A função logística$_{[-1,1]}$ é calculada por F(x)=[2/(1+exp(–x)]–1, com intervalo de resposta de F(x)=[–1,+1], enquanto que a logística$_{[0,1]}$ pode ser calculada por F(x)=1/(1+exp(-x)), com F(x)=[0,1].

de dezenas de horas).

Os resultados obtidos com as redes neurais estão na Tabela 24. O nível de erro dos modelos neurais é bom, sendo ligeiramente inferior aos modelos correspondentes de regressão (Tabela 20), exceto para o grupo dos imóveis Grandes, tanto para o erro padrão quanto para os erros percentuais. O equilíbrio entre erros de treinamento e de teste permite afastar a suspeita de overfitting. Considerando o erro padrão, o erro do modelo Mercado foi cerca de 9% menor do que com a regressão, enquanto que o erro do modelo de avaliação em massa foi 2,5% menor.

Tabela 24: Resultados obtidos com as redes neurais

Modelo	Treino			Teste		
	EA	EP	Casos	EA	EP	Casos
MERCADO	16,05	17.944	24.289	16,13	18.020	6.074
Pequenos	16,74	10.462	21.148	16,87	10.559	5.288
Médios	13,21	29.405	2.838	13,42	29.452	710
Grandes	12,21	75.065	303	10,79	76.715	76
MASSA	16,26	16.329	24.289	16,38	16.495	6.074

7.2.3.1 Representação simbólica das redes neurais

Embora haja o inconveniente da falta de um modelo hedônico, as redes neurais não são realmente uma "caixa preta", pois as redes podem ser representadas por um conjunto de equações. Trata-se de um formato complexo, quase incompreensível para o usuário final, mas há um modelo matemático. Por exemplo, para o modelo Mercado, a rede neural pode ser descrita pela Equação 35:

$$Y^h=F_{23}[-0,06440+1,31053*F_{12}(X_{12})-1,22626*F_{13}(X_{13})+ 0,78600*F_{14}(X_{14})+0,37089*F_{15}(X_{15})-1,68319*F_{16}(X_{16})+ 0,09843*F_{17}(X_{17})+0,45751*F_{18}(X_{18})-0,29391*F_{19}(X_{19})+$$ (Equação 35)

$1{,}32671*F_{20}(X_{20})-1{,}15142*F_{21}(X_{21})+0{,}09971*F_{22}(X_{22})]$

Onde Y^h é a saída da rede (valor estimado), X_j são os somatórios das entradas em cada neurônio oculto (j=12,...,22), F_j são as funções de ativação (no caso, iguais para todos os neurônios não-lineares) e os coeficientes são os pesos das conexões entre a camada oculta e a camada de saída. As saídas dos neurônios ocultos são dadas pela aplicação das funções de ativação aos somatórios ponderados das entradas em cada neurônio, posteriormente com a aplicação da função de ativação do neurônio de saída (F_{23}). Os somatórios X_j são calculados pelas relações apresentadas na Figura 29.

$X_{12}=-0{,}11692+0{,}24988*X_1+1{,}96036*X_2-0{,}52362*X_3+0{,}10678*X_4-0{,}54584*X_5-0{,}48432*X_6+0{,}64490*X_7-0{,}23931*X_8-0{,}24132*X_9-0{,}69206*X_{10}-0{,}57628*X_{11}$

$X_{13}=0{,}30686+0{,}29104*X_1-2{,}15555*X_2-0{,}23315*X_3-0{,}04429*X_4-0{,}23163*X_5-0{,}79767*X_6-0{,}43676*X_7-0{,}44846*X_8-0{,}61651*X_9-0{,}64560*X_{10}-0{,}81167*X_{11}$

$X_{14}=-0{,}30583+0{,}15531*X_1+1{,}15663*X_2+0{,}84697*X_3-0{,}46843*X_4+0{,}16577*X_5+0{,}49420*X_6-0{,}94483*X_7-0{,}46594*X_8+0{,}07521*X_9+0{,}67462*X_{10}-0{,}99405*X_{11}$

$X_{15}=0{,}27744-0{,}43872*X_1+0{,}19775*X_2-0{,}57421*X_3+1{,}02011*X_4+0{,}38012*X_5+1{,}04876*X_6-0{,}68285*X_7+0{,}37727*X_8-0{,}47133*X_9-0{,}23228*X_{10}+0{,}47400*X_{11}$

$X_{16}=0{,}93573-0{,}07050*X_1-2{,}70451*X_2-0{,}14799*X_3+0{,}82558*X_4+0{,}06034*X_5+0{,}08215*X_6-0{,}05571*X_7-0{,}64606*X_8+0{,}39128*X_9+0{,}63636*X_{10}+0{,}09500*X_{11}$

$X_{17}=0{,}11776+0{,}19480*X_1+0{,}20027*X_2+0{,}78390*X_3+0{,}61617*X_4-0{,}06145*X_5-0{,}59426*X_6+0{,}43289*X_7+0{,}27847*X_8-0{,}38570*X_9+0{,}20113*X_{10}+0{,}79948*X_{11}$

$X_{18}=0{,}46804+0{,}34544*X_1+0{,}79811*X_2-0{,}59574*X_3+1{,}10363*X_4-0{,}29349*X_5-0{,}82632*X_6+0{,}47843*X_7-0{,}71154*X_8+0{,}21238*X_9-0{,}49541*X_{10}-0{,}38349*X_{11}$

$X_{19}=-0{,}88106-0{,}60353*X_1-0{,}44679*X_2-0{,}46635*X_3-0{,}54670*X_4+0{,}70344*X_5-0{,}40715*X_6+0{,}34727*X_7+0{,}78608*X_8+0{,}43385*X_9-0{,}68569*X_{10}+0{,}73383*X_{11}$

$X_{20}=0{,}51339-0{,}32318*X_1+2{,}16562*X_2+0{,}33256*X_3-0{,}11329*X_4+0{,}54807*X_5-0{,}20542*X_6-0{,}46506*X_7-0{,}72492*X_8-0{,}86982*X_9+0{,}06222*X_{10}+0{,}41792*X_{11}$

$X_{21}=-0{,}99259-0{,}20164*X_1-1{,}32337*X_2+0{,}56630*X_3-0{,}37690*X_4+0{,}28157*X_5+0{,}32458*X_6+0{,}01931*X_7-0{,}61588*X_8-1{,}10797*X_9-0{,}14165*X_{10}-0{,}21962*X_{11}$

$X_{22}=0{,}78485+0{,}00463*X_1+0{,}78033*X_2+0{,}48156*X_3-0{,}85717*X_4-0{,}23684*X_5-0{,}50778*X_6+0{,}21482*X_7+0{,}00213*X_8-0{,}97040*X_9-0{,}60075*X_{10}+0{,}03420*X_{11}$

Figura 29: Somatórios dos produtos das entradas da rede pelos pesos das conexões com a camada oculta – rede neural de Mercado

Por sua vez, os X_i (i=1,...11) são as variáveis de entrada, em valores normalizados. No caso, X_1=Bairro/100, X_2=Artocons/10.000, X_3=Razão, X_4=Idade/100, X_5=Mês/100, X_6=Med, X_7=Fin, X_8=Lux, X_9=Centro/100, X_{10}=Comércio/100, X_{11}=Vista, tendo ainda $Y=Valor^{3/4}/100.000$. Substituindo os somatórios da Figura 29 na Equação 35, a rede pode finalmente ser expressa como na Figura 30.

$$Y^h=F_{23}[-0,06440+1,31053*F_{12}(-0,11692+0,24988*X_1+1,96036*X_2-0,52362*X_3+0,10678*X_4-0,54584*X_5-0,48432*X_6+0,64490*X_7-0,23931*X_8-0,24132*X_9-0,69206*X_{10}-0,57628*X_{11})-1,22626*F_{13}(0,30686+0,29104*X_1-2,15555*X_2-0,23315*X_3-0,04429*X_4-0,23163*X_5-0,79767*X_6-0,43676*X_7-0,44846*X_8-0,61651*X_9-0,64560*X_{10}-0,81167*X_{11})+0,78600*F_{14}(-0,30583+0,15531*X_1+1,15663*X_2+0,84697*X_3-0,46843*X_4+0,16577*X_5+0,49420*X_6-0,94483*X_7-0,46594*X_8+0,07521*X_9+0,67462*X_{10}-0,99405*X_{11})+0,37089*F_{15}(0,27744-0,43872*X_1+0,19775*X_2-0,57421*X_3+1,02011*X_4+0,38012*X_5+1,04876*X_6-0,68285*X_7+0,37727*X_8-0,47133*X_9-0,23228*X_{10}+0,47400*X_{11})-1,68319*F_{16}(0,93573-0,07050*X_1-2,70451*X_2-0,14799*X_3+0,82558*X_4+0,06034*X_5+0,08215*X_6-0,05571*X_7-0,64606*X_8+0,39128*X_9+0,63636*X_{10}+0,09500*X_{11})+0,09843*F_{17}(0,11776+0,19480*X_1+0,20027*X_2+0,78390*X_3+0,61617*X_4-0,06145*X_5-0,59426*X_6+0,43289*X_7+0,27847*X_8-0,38570*X_9+0,20113*X_{10}+0,79948*X_{11})+0,45751*F_{18}(0,46804+0,34544*X_1+0,79811*X_2-0,59574*X_3+1,10363*X_4-0,29349*X_5-0,82632*X_6+0,47843*X_7-0,71154*X_8+0,21238*X_9-0,49541*X_{10}-0,38349*X_{11})-0,29391*F_{19}(-0,88106-0,60353*X_1-0,44679*X_2-0,46635*X_3-0,54670*X_4+0,70344*X_5-0,40715*X_6+0,34727*X_7+0,78608*X_8+0,43385*X_9-0,68569*X_{10}+0,73383*X_{11})+1,32671*F_{20}(0,51339-0,32318*X_1+2,16562*X_2+0,33256*X_3-0,11329*X_4+0,54807*X_5-0,20542*X_6-0,46506*X_7-0,72492*X_8-0,86982*X_9+0,06222*X_{10}+0,41792*X_{11})-1,15142*F_{21}(-0,99259-0,20164*X_1-1,32337*X_2+0,56630*X_3-0,37690*X_4+0,28157*X_5+0,32458*X_6+0,01931*X_7-0,61588*X_8-1,10797*X_9-0,14165*X_{10}-0,21962*X_{11})+0,09971*F_{22}(0,78485+0,00463*X_1+0,78033*X_2+0,48156*X_3-0,85717*X_4-0,23684*X_5-0,50778*X_6+0,21482*X_7+0,00213*X_8-0,97040*X_9-0,60075*X_{10}+0,03420*X_{11})]$$

Figura 30: Representação matemática da rede neural de Mercado

O mesmo procedimento pode ser aplicado às demais redes compiladas. Percebe-se que esta representação ainda não é conveniente, pela dificuldade de interpretação, e também porque não pode ser reduzida ou comprimida pela composição das funções, pois as funções de ativação $F_j(x)$ são funções não-lineares. É necessário transformar a rede neural em um formato mais facilmente compreensível. Utilizando a técnica proposta por Cechin (1998), que consiste basicamente em substituir as funções de ativação por funções lineares, em alguns casos a composição torna-se possível, como se verá a seguir.

7.2.3.2 Extração de regras para o modelo Mercado[52]

O método de extração deve proporcionar regras com comportamento similar ao das redes neurais, em termos de valores estimados e de nível de erro. Para a definição das funções lineares que substituirão as funções de ativação, deve-se verificar a faixa de funcionamento de cada neurônio, investigando os níveis de ativação dos neurônios não lineares. Os valores, calculados com as equações apresentadas na Figura 29, são os apresentados na Tabela 25. Verifica-se que o sinal de ativação mínimo do oitavo neurônio oculto (j=19), ultrapassou os limites para a função de pertinência central (F(X)=0,5X), a qual deveria variar no intervalo [–2,+2], exigindo a computação de uma segunda regra (ver Figura 13).

Tabela 25: Níveis de ativação dos neurônios da rede neural de Mercado (para os casos de treino)

Neurônio	Níveis de ativação (X_j)			
	Média	Mínimo	Máximo	Centro do intervalo
Oculto – 12	-0,84224	-1,28485	0,57466	-0,35510
Oculto – 13	-0,55001	-0,88359	0,25445	-0,31457
Oculto – 14	0,62903	-1,12542	1,14663	0,01061
Oculto – 15	0,66548	-1,28753	1,68751	0,19999
Oculto – 16	1,05068	-0,13530	1,51408	0,68939
Oculto – 17	0,51019	-0,30461	1,60853	0,65196
Oculto – 18	-0,26875	-0,93147	1,39258	0,23055
Oculto – 19	-1,65539	-2,37002	-0,41265	-1,39134
Oculto – 20	0,53718	-0,33809	0,92237	0,29214
Oculto – 21	-0,46583	-1,60520	-0,13372	-0,86946
Oculto – 22	0,51963	-0,18885	1,38934	0,60025
Saída – 23	0,08180	0,00770	1,02933	0,51852

Este problema ocorreu apenas para o neurônio oculto 19, portanto o conjunto de regras que deve substituir a rede neural pode ser composto por apenas duas regras, uma para os casos em que $X_{19}<-2$ e outra para $X_{19}\geq-2$, sendo iguais quanto ao restante da rede. A função de ativação $F_{19}(X)=\tanh(X/2)$ pode ser substituída por $F(X)= -1*G_1+0,5X*G_2$, sem a necessidade do elemento $+1*G_3$, pois o intervalo do nível de ativação é todo negativo ($X_{19}=[-2,37002; -0,41265]$), sendo que G_i são as funções de pertinência correspondentes às funções

[52] Para demonstrar mais claramente o fluxo de análise, foram explicitados todos os passos desenvolvidos na extração de regras para o modelo Mercado, incluindo também os modelos que não atingiram resultados satisfatórios.

lineares que substituem as funções de ativação (a_i+b_iX).

É necessário calcular o valor da pertinência a cada uma das duas regras, o que incrementa a complexidade do sistema, mas, como as demais funções de pertinência têm valores iguais para os dois casos, não precisam ser calculadas e as funções de pertinência das duas regras são calculadas apenas em função da ativação do neurônio oculto 19, conforme a Equação 36:

$$G_1(X)= -F(X)+(1-F(X))*(1+F(X))/2*X \qquad \text{(Equação 36)}$$

$$G_2(X)=(1-F(X))*(1+F(X))$$

Onde $F(X)=\tanh(X/2)$ e $X=X_{19}$. O nível de ativação do neurônio oculto 19, utilizando os pesos correspondentes das conexões entre as camadas de entrada e oculta, é dado pelo somatório apresentado na Figura 29, que também pode ser expresso como na Equação 37, convertido para cálculo com os valores originais das variáveis:

$$X_{19}= -0,88106-0,0000447*\text{Artocons}-0,46635*\text{Razão}-0,40715*\text{Med}+ \qquad \text{(Equação 37)}$$
$$0,34727*\text{Fin}+0,78608*\text{Lux}-0,00547*\text{Idade}-0,00603*\text{Bairro}+$$
$$0,00434*\text{Centro}-0,00686*\text{Comércio}+0,07338*\text{Vista}+0,00703*\text{Mês}$$

As duas regras geradas são as apresentadas nas Equações 38 e 39. As funções de pertinência G_i envolvem a função de ativação original, não-linear, impedindo a composição das funções em uma única regra. Estas equações têm alguns coeficientes com sinais distintos dos esperados, tais como para as variáveis Razão (na regra R_1, Equação 38) e Vista (na R_2, Equação 39), mas como se tratam de regras difusas existe uma fusão dos efeitos finais das regras e a interpretação deve levar em conta as duas regras simultaneamente.

Os valores dos imóveis são calculados através da ponderação das estimativas de cada regra pela pertinência do caso à regra, ou seja, o valor final é dado por $\text{Valor}=(Y_1*G_1(X)+Y_2*G_2(X))/(G_1(X)+G_2(X))$. As regras geradas podem ser vistas como modelos hedônicos, similares em formato e coeficientes ao modelo correspondente calculado

com a regressão múltipla (Tabela 19 – modelo de Mercado), de fácil análise e compreensão pelo usuário final. Contudo, aplicando as duas regras obteve-se um nível elevado de erros, sendo que a raiz dos erros quadrados atingiu 24.423 e os erros percentuais foram de 31,12%, o primeiro sendo 35% superior ao erro obtido com as redes neurais e praticamente dobrando os erros no caso dos erros percentuais. Assim, esta conversão não atende a um dos requisitos necessários, o de manter aproximadamente o mesmo nível de erro da rede neural.

$$R_1\text{: SE X é } G_1 \text{ ENTÃO } Y_1=(2.024{,}944+39{,}017*\text{Artocons}-1.417{,}302*\text{Razão}+2.193{,}436*\text{Med}+2.457{,}398*\text{Fin}+3.742{,}691*\text{Lux}-105{,}356*\text{Idade}+7{,}423*\text{Bairro}-62{,}592*\text{Centro}-190{,}854*\text{Comércio}+471{,}619*\text{Vista}-5{,}650*\text{Mês})^{4/3} \quad \text{(Equação 38)}$$

$$R_2\text{: SE X é } G_2 \text{ ENTÃO } Y_2=(513{,}253+39{,}345*\text{Artocons}+2.009{,}321*\text{Razão}+489{,}799*\text{Med}+605{,}745*\text{Fin}+1.016{,}772*\text{Lux}-65{,}186*\text{Idade}+51{,}768*\text{Bairro}-94{,}470*\text{Centro}-140{,}472*\text{Comércio}-67{,}581*\text{Vista}-57{,}337*\text{Mês})^{4/3} \quad \text{(Equação 39)}$$

Segundo Cechin (1998), as diferenças entre a rede neural e as regras decorrem do erro de aproximação (diferença entre a função de ativação original e a função linear utilizada na extração das regras) e da soma de pertinências dos trechos lineares menor do que a unidade (falta de cobertura das regras para algumas partes da rede original). Para diminuir o erro, podem ser utilizados alguns mecanismos de refinamento, tais como a definição de funções de ativação e de pertinência diferentes para cada neurônio não-linear (mais ajustadas ao nível efetivo de ativação dos neurônios), as quais podem ser determinadas pela média dos sinais de ativação ou pelo centro do intervalo dos valores de ativação em cada neurônio não-linear, definição de outros formatos para as funções de ativação e de pertinência, ou através do

refinamento das próprias regras geradas, os quais foram investigados, conforme descrito a seguir.

7.2.3.2.1 Refinamento das regras utilizando parte da função não linear

Esta alternativa consiste da escolha da função linear mais adaptada ao trecho específico de funcionamento dos neurônios, utilizando a média dos sinais de ativação ou o ponto médio, considerando os níveis de ativação de cada neurônio individualmente (ver Tabela 25). Para a rede neural Mercado, os parâmetros calculados para as funções lineares foram os seguintes (Tabela 26).

Tabela 26: Funções lineares para a rede neural de Mercado (para os casos de treino)

	Funções lineares (F(x)=a+bX)			
	Cálculo pela média		Cálculo pelo centro do intervalo	
Neurônio	a_i	b_i	a_i	b_i
Oculto – 12	-0,04342	0,42085	-0,00364	0,48456
Oculto – 13	-0,01306	0,46401	-0,00254	0,48783
Oculto – 14	0,01919	0,45363	0,00001	0,49999
Oculto – 15	0,02252	0,44849	0,00066	0,49503
Oculto – 16	0,07842	0,38393	0,02488	0,44500
Oculto – 17	0,01051	0,46882	0,02125	0,45042
Oculto – 18	-0,00159	0,49108	0,00101	0,49341
Oculto – 19	-0,23341	0,26932	-0,15773	0,31903
Oculto – 20	0,01220	0,46560	0,00204	0,48948
Oculto – 21	-0,00807	0,47383	-0,04735	0,41625
Oculto – 22	0,01109	0,46771	0,01679	0,45754
Saída – 23	0,00005	0,49917	0,01102	0,46784

A partir destes trechos lineares, foram extraídas regras únicas nos dois casos (Equações 40 e 41). A Equação 40 foi definida com funções lineares calculadas pelas médias dos sinais de ativação, enquanto que a Equação 41 foi calculada com o ponto médio dos intervalos (ver Tabela 26). No primeiro caso, a regra extraída atingiu erros de teste de 25.750 e de 53,19%, enquanto que no segundo caso foram obtidos erros de 32.373 e de 67,54%, os quais são insuficientes, pois são muito superiores aos erros da rede neural original.

R_1: SE *Verdadeiro* ENTÃO Valor=(2.533,468+34,042* Artocons+ 198,827*Razão–1.209,174*Med+101,009*Fin+5813,205*Lux– 15,570*Idade+22,538*Bairro–34,398*Centro–91,655*Comércio+ 396,107*Vista–18,234*Mês)$^{4/3}$ (Equação 40)

R_1: SE *Verdadeiro* ENTÃO Valor=(1.598,004+34,424*Artocons+ 2.607,858*Razão+375,103*Med+422,053*Fin+4.159,024*Lux– 58,566*Idade+22,484*Bairro–104,069*Centro–124,066*Comércio+ 14,471*Vista–20,163*Mês)$^{4/3}$ (Equação 41)

O antecedente "Verdadeiro" indica que as duas regras aplicam-se a todos os casos. As funções de pertinência têm o formato geral G(X)=F(X)/(a+bX), onde F(X)=tanh(X/2) é a função de pertinência original e (a+bX) é a função linear que a substitui (ver Tabela 26). Não há necessidade de cálculo das funções de pertinência, pois há apenas um trecho linear para cada função de ativação. Porém, os somatórios das funções de pertinência são menores do que 1, o que é uma das fontes de erro da conversão da rede em regras.

Com este procedimento foram extraídas regras simples, mas o nível de erro obtido também não foi satisfatório, especialmente quanto ao erro percentual, tendo estas regras atingido desempenho inferior ao das regras anteriores (Equações 38 e 39).

7.2.3.2.2 Refinamento utilizando funções de pertinência triangulares

Aplicando funções de pertinência com formato triangular, pode-se obter o somatório unitário de pertinência. O núcleo da função de ativação sugerida por Cechin (1998) é uma função linear: 0,225X, e a função de ativação de cada neurônio não linear então é substituída por (Equação 42):

$$F(X)= -1*G_1+0{,}225X*G_2+1*G_3, \text{ onde:} \qquad \text{(Equação 42)}$$

$$G_1=1-G_2=X/4, \text{ para } X<0$$

$$G_2=1-|X|/4, \text{ para todo } X$$

$$G_3=1-G_2=X/4, \text{ para } X\geq 0$$

Este conjunto pode ser condensado em uma única função. Neste caso, as funções de ativação originais são substituídas por $F(X)=[X/4+0{,}225X*(1-|X|/4)]$. Aplicando esta formulação, os erros calculados foram de 29.354 e 23,42% para os dados de teste, os quais ainda estão longe do desempenho da rede neural.

Explorando outras alternativas, verificou-se que a alteração do componente angular desta função (0,225) permite sintonizar a regra extraída. Após algumas tentativas, o desempenho foi significativamente melhorado utilizando $F(X)= -1*G_1+0{,}255X*G_2+1*G_3$, obtendo-se então erros médios de 19.758 e de 20,68%. Neste caso, a saída da rede é calculada por $Y^h=X_{23}/4+0{,}255X_{23}*(1-|X_{23}|/4))$. Como o intervalo de saída é todo positivo ($X_{23}>0$), a operação de valor absoluto pode ser removida, e a regra fica como na Equação 43:

$$R_1: \text{SE } \textit{Verdadeiro} \text{ ENTÃO } Y^h=X_{23}/4+0{,}255X_{23}*(1-X_{23}/4) =0{,}505X_{23}-0{,}06375X_{23}^2 \qquad \text{(Equação 43)}$$

O sinal de ativação do neurônio de saída (X_{23}) é calculado pela Equação 44, sendo que os sinais de ativação dos neurônios ocultos (X_j, j=12,...,22) são calculados como antes. Para esta regra (representada pela composição do conjunto das Equações 43 e 44), o nível de erro é adequado, com uma perda razoavelmente pequena em relação à rede original, de 9,6% se considerado o erro padrão e 28% a mais no erro percentual. Porém, esta regra também tem o inconveniente de não possibilitar a composição dos elementos internos, gerando um modelo hedônico claro. Efetivamente, esta regra parece tão ou mais complexa do que a regra da Equação 35. A operação de módulo pode ser removida em alguns trechos, tal como nos neurônios ocultos 19, 21 e 23, que têm toda a variação do sinal de ativação em um único lado

do eixo (ver Tabela 25), mas mesmo assim não permitem uma simplificação razoável. Assim, esta alternativa também não é conveniente, mesmo com um nível de erro similar à rede neural, em função da complexidade.

$$X_{23}=-0,06440+1,31053*(X_{12}/4+(1-|X_{12}|/4)*0,255*X_{12})-1,22626*(X_{13}/4+(1-|X_{13}|/4)*0,255*X_{13})+0,78600*(X_{14}/4+(1-|X_{14}|/4)*0,255*X_{14})+0,37089*(X_{15}/4+(1-|X_{15}|/4)*0,255*X_{15})-1,68319*(X_{16}/4+(1-X_{16}|/4)*0,255*X_{16})+0,09843*(X_{17}/4+(1-|X_{17}|/4)*0,255*X_{17})+ 0,45751*(X_{18}/4+(1-|X_{18}|/4)*0,255*X_{18})-0,29391*(X_{19}/4+(1-|X_{19}|/4)* 0,255*X_{19})+1,32671*(X_{20}/4+(1-|X_{20}|/4)*0,255*X_{20})-1,15142*(X_{21}/4+(1-|X_{21}|/4)*0,255*X_{21})+0,09971*(X_{22}/4+(1-|X_{22}|/4)*0,255*X_{22}))$$ (Equação 44)

7.2.3.2.3 Refinamento da constante da equação das regras geradas

Durante a análise do desempenho das regras, verificou-se que os somatórios dos erros eram muito diferentes de zero, ou seja, havia tendências de superestimar ou subestimar valores nas regras geradas, independentemente dos parâmetros do caso. Utilizando uma analogia com a análise de regressão, na qual o intercepto é calculado de forma a tornar nulo o somatório dos erros, foi desenvolvido um ajuste nas constantes das Equações 40 e 41, sintonizando as mesmas para aproximar as estimativas das regras com as estimativas da rede neural. Esta alternativa de refinamento não está entre as apresentadas por Cechin (1998), mas atende ao espírito de sintonia das regras difusas (Cordón *et al.*, 2001).

Busca-se o valor da constante (a_0) que minimiza a diferença para as estimativas da rede neural. Em função de o sistema ser não linear (expoente ¾ na variável dependente), o valor da constante não é obtido diretamente. Assim, a constante foi definida através da aplicação de algoritmos genéticos. Os demais coeficientes não foram alterados para preservar a vinculação

com a rede neural. As novas constantes foram de $a_{0\text{-}40}$=2.295,407 e $a_{0\text{-}25}$=306,362 para as Equações 40 e 41, obtendo-se erros de 24.957 e 46,77% e de 21.735 e de 21,51%, respectivamente. A Equação 41, de melhor desempenho entre as duas, atingiu erros 21% e 33% superiores aos da rede neural original em relação ao EP e ao erro percentual, respectivamente, mesmo após o refinamento (Equação 45). O erro ainda é grande, mas como se busca também a facilidade de interpretação, em alguns casos pode ser conveniente adotar este procedimento. Tendo em vista o compromisso entre simplicidade e precisão do modelo, este foi o modelo adotado para estimar os valores para o modelo de Mercado.

R_1: SE *Verdadeiro* ENTÃO Valor=(306,362+34,042* Artocons+198,827* Razão–1.209,174*Med+101,009*Fin+5813,205*Lux–15,570*Idade+ 22,538*Bairro–34,398*Centro–91,655*Comércio+ 396,107*Vista– 18,234*Mês)$^{4/3}$ (Equação 45)

7.2.3.3 Regras para os grupos segmentados pela área

No caso dos modelos gerados para os dados agrupados por área, foi adotada a mesma seqüência de análise que no caso exposto acima (Mercado). Com as redes neurais geradas para os grupos, inicialmente foram analisados os níveis de ativação dos neurônios não-lineares. Na Tabela 27 estão apresentados os níveis médios e os centros dos intervalos de variação das entradas ponderadas de cada neurônio. Verifica-se que há diferenças entre estes dois valores, na maioria dos casos, resultando em regras bastante distintas em um e outro caso.

Em seguida, foram calculadas as funções lineares (a_i+b_iX) para os dois casos e testadas as regras correspondentes. O nível de erro não foi satisfatório e então foram investigadas todas as formas de refinamento. As regras finais foram obtidas pelo processo de sintonia do intercepto das equações. Os resultados calculados pelas regras indicam um nível de erro ligeiramente superior ao apurado com as redes neurais para os imóveis Médios e Grandes, e bastante superior para os imóveis Pequenos, ou seja, em todos os modelos há uma perda na

conversão, mesmo com o refinamento das regras (ver Tabela 28, a seguir).

Tabela 27: Níveis de ativação dos neurônios das redes neurais em cada grupo (para os dados de treino)

Neurônio	Modelo						
	Pequenos		Médios		Grandes		
	Média	Centro	Média	Centro	Neurônio	Média	Centro
Oculto – 12	-0,51758	-0,04109	-0,37703	-0,37936	Oculto – 9	0,29824	0,24948
Oculto – 13	0,05642	0,22711	-0,62060	-0,65471	Oculto – 10	-0,47811	-0,75578
Oculto – 14	0,03943	-0,43775	0,22786	-0,09772	Oculto – 11	-0,28652	-0,33880
Oculto – 15	0,28810	0,02949	-0,01288	0,05294	Oculto – 12	0,06437	0,39563
Oculto – 16	0,97074	0,71828	0,76369	0,49615	Oculto – 13	0,66935	0,44916
Oculto – 17	0,09328	0,28807	0,84435	1,01213	Oculto – 14	-0,06602	-0,15321
Oculto – 18	-0,04136	0,26693	0,07262	0,18175	Oculto – 15	0,36520	0,23072
Oculto – 19	-0,20505	-0,14520	-1,50264	-1,35639	Oculto – 16	0,88385	0,74004
Oculto – 20	0,80291	0,46010	0,44392	0,44647			
Oculto – 21	-0,13505	-0,53431	-0,80001	-0,92408			
Oculto – 22	0,07671	0,16470	0,81222	0,72543			
Saída – 23	0,06686	0,09128	0,18345	0,20838	Saída – 17	0,42914	0,55902

Na primeira regra, para os imóveis Pequenos, os sinais e valores dos coeficientes estão dentro do esperado, em valores razoáveis, a não ser para o coeficiente da variável Vista, que é muito diferente do coeficiente obtido com regressão. Porém, a diferença das estimativas para as estimativas da rede neural correspondente atingiu cerca de 8% no erro padrão e 35% no erro percentual (Equação 46).

$$R_1\text{: SE Artocons}<138{,}07\text{m}^2\text{ ENTÃO Valor}=(230{,}984+ 40{,}805*\text{Artocons}+ 1.366{,}038*\text{Razão}+ 205{,}339*\text{Med}+535{,}391*\text{Fin}+ 1.298{,}895*\text{Lux}- 35{,}249*\text{Idade}+17{,}732*\text{Bairro}-38{,}551*\text{Centro}-110{,}263*\text{Comércio}+ 74{,}173*\text{Vista}-16{,}103*\text{Mês})^{4/3} \quad \text{(Equação 46)}$$

Na regra para os imóveis Médios (Equação 47), houve o maior equilíbrio com a rede neural, e a diferença para a rede neural original foi de apenas 5% no erro padrão e 3% nos erros percentuais. Os coeficientes também estão dentro do esperado, exceto para a variável Razão, cujo coeficiente teve sinal negativo e para Vista, cujo coeficiente foi bem maior do que o da

regressão.

R_2: SE 138,07m^2≤Artocons≤355m^2 ENTÃO Valor=(2.657,403+32,585* Artocons–724,350*Razão+354,815*Med+1.819,161*Fin+5.600,965*Lux–38,302*Idade+32,133* Bairro–130,390*Centro–192,084*Comércio+4.626,546*Vista–32,107*Mês)$^{4/3}$ (Equação 47)

Para os imóveis terceiro grupo, a regra gerada também obteve um bom nível de ajustamento à rede neural, com cerca de apenas 2,6% de diferença no erro padrão e 8% no erro percentual, sem diferenças significativas nos coeficientes desta regra (Equação 48).

R_3: SE Artocons>355m^2 ENTÃO Valor=(977,981+27,628* Artocons+5.289,338*Razão+ 6.318,397*Fin+10.067,412*Lux–163,287*Idade+19,362*Bairro–691,527*Centro–81,525*Mês)$^{4/3}$ (Equação 48)

Ressalta-se que as regras geradas não compõem um sistema de regras difusas, mas sim um sistema de regras clássicas, pois cada regra é utilizada para calcular os valores dos imóveis de apenas um grupo, sem composição das estimativas para a obtenção do valor final.

7.2.3.4 Resultados dos modelos com a extração de regras utilizando lógica difusa

Como visto anteriormente, a perda na conversão decorre da simplificação das funções de pertinência, sendo o preço a pagar pelo aumento de interpretabilidade do sistema. Neste sentido, foram preferidos os modelos com regras únicas. Embora apresentem diferenças para os modelos neurais correspondentes, a vantagem destes modelos é que, havendo apenas uma regra, não há necessidade de calcular os valores de pertinência, tornando mais simples a compreensão e a utilização por parte do usuário final. Os resultados obtidos após as etapas de refinamento das regras foram os apresentados na Tabela 28.

A regra extraída para o modelo Mercado apresentou diferenças significativas em relação ao modelo neural, com erros maiores. Por outro lado, a união das estimativas das outras três regras, formando o modelo de avaliação em massa, também revelou diferenças importantes em relação às redes neurais, atingindo ao final erros de 17.340 e 21,60%, os quais são superiores aos erros originais, que eram de 16.495 para o erro padrão e de 16,38 para o erro percentual, ou seja, 5% de diferença no primeiro caso e 32% a mais de erro, considerando os erros percentuais (ver Tabelas 24 e 28). Embora os erros sejam superiores, a correlação das estimativas produzidas pelas redes neurais com as calculadas através das regras selecionadas é elevada, sendo superior a 0,99 para os modelos de Mercado e de avaliação em massa.

Tabela 28: Resultados obtidos com as regras extraídas das redes neurais

Modelo	Treino			Teste			
	EA	EP	Casos	EA	EP	Casos	Correlação com as redes neurais*
MERCADO (Eq.29)	21,64	21.116	24.289	21,51	21.735	6.074	0,9915
Pequenos (Eq.30)	22,81	11.339	21.148	22,78	11.404	5.288	0,9896
Médios (Eq.31)	13,78	31.523	2.838	13,85	30.953	710	0,9872
Grandes (Eq.32)	12,45	77.865	303	11,63	78.720	76	0,9967
MASSA	21,63	17.427	24.289	21,60	17.340	6.074	0,9976

*correlação das estimativas das regras com as estimativas dos modelos neurais correspondentes

É importante verificar os vários caminhos para a geração de regras, pois peculiaridades nos dados podem levar a melhores resultados por um dos meios, bem como os requisitos do usuário podem definir a preferência por um ou outro formato. Por outro lado, a extração de regras simples e com um nível de erro aceitável em alguns casos fornece indicações de que os dados não contêm fortes fugas à linearidade.

7.2.4 Modelos aditivos generalizados, com coeficientes e expoentes determinados por algoritmos genéticos

Um dos problemas apontados na revisão bibliográfica é o da escolha das formas funcionais, tendo em vista o relativo desconhecimento dos componentes (variáveis a serem incluídas) e

uma possível fuga à linearidade. A seleção das variáveis que devem participar do modelo foi realizada no Capítulo 6, e resta a definição dos formatos dos modelos, ou seja, a análise das alternativas ao modelo linear. Na metodologia tradicional, usando regressão múltipla, esta tarefa é desenvolvida iterativamente, por tentativas, ou por busca exaustiva em *softwares* estatísticos especiais para avaliação de imóveis, porém dentre formatos pré-determinados, ou através de Box-Cox.

Os modelos aditivos generalizados são uma forma não paramétrica de regressão, compostos por um somatório de funções, admitindo qualquer formato para estas funções. Podem ser vistos como uma alternativa ao procedimento de Box-Cox, e também ao uso de redes neurais na consideração de efeitos não-lineares. No caso, foram utilizadas funções com o formato $(c_i X_i)^{ei}$, sendo X_i as variáveis independentes, c_i uma constante e e_i um expoente. Explorações iniciais não indicaram a presença de múltiplos termos as variáveis independentes, nem de interações entre as variáveis (termos com $X_i X_j$), além das proporcionadas pela transformação da variável dependente (que assumiu o expoente ¾ nos modelos anteriores). Os coeficientes e expoentes foram determinados com algoritmos genéticos, com a estimação de quatro equações (para o conjunto Mercado e para os três grupos isoladamente). Neste caso, o sistema de equações resultante também pode ser considerado como um sistema de regras baseadas na lógica clássica.

Os modelos investigados incluíram as onze variáveis independentes selecionadas como relevantes, a variável dependente, a constante da equação, os coeficientes e os expoentes correspondentes, ou seja, é um modelo hedônico de preços generalizado, com a forma da Equação 49.

$$\text{Valor} = (c_0 + c_1 * \text{Artocons}^{e_1} + c_2 * \text{Razão}^{e_2} + \ldots + c_{11} * \text{Mês}^{e_{11}})^{e_0} \qquad \text{(Equação 49)}$$

A diferença em relação aos modelos de regressão (Equação 33) é a investigação dos expoentes em cada variável, com uma busca ampla no espaço de soluções. Foram utilizados algoritmos genéticos, através de uma rotina desenvolvida pelo autor. A população consistiu de indivíduos (cromossomos) com 24 genes (12 coeficientes e 12 expoentes), codificados como números reais e distribuídos conforme esquema apresentado na Figura 31. Os expoentes das

variáveis binárias, em função do formato destas variáveis, as quais assumem apenas valores 0 e 1, foram mantidos como unitários.

a_0	e_0	a_1	e_1	...	a_{10}	e_{10}	a_{11}	e_{11}

Figura 31: Aspecto dos indivíduos utilizados nos algoritmos genéticos para modelos aditivos generalizados

As populações iniciais foram geradas a partir do conhecimento anterior, no caso as equações de regressão apresentadas na Tabela 19, gerando os indivíduos através de variações aleatórias de ±50% sobre os coeficientes e sobre os expoentes. A avaliação das soluções empregou o erro padrão (EP), com a seguinte função de *fitness* (Equação 50).

$$F_i=1/(1+EP_i), \qquad \text{com } EP_i=[\Sigma_{i,j} (Y_j-Y^h_{i,j})^2/(n-2)]^{0,5} \qquad \text{(Equação 50)}$$

Onde F_i é a estimativa de *fitness* para a regra i, EP_i é o erro padrão calculado para o modelo i com os dados de treinamento, Y_j é o preço observado e $Y^h_{i,j}$ é o valor estimado pelo modelo i para o caso j. Foi utilizada uma seleção baseada no mecanismo de roleta, porém com seleção elitista inicial de 5%, ou seja, os 5% melhores indivíduos são copiados diretamente para a população provisória e depois os restantes 95% são selecionados proporcionalmente ao seu nível de ajustamento.

Os operadores genéticos utilizados na reprodução foram os operadores de cruzamento e mutação. Foram testadas populações de 50 a 100 indivíduos, com taxas de cruzamento variando na faixa de 0,6 a 0,95 e taxas de mutação na faixa de 0,001 a 0,2, com os melhores resultados ocorrendo com populações de 100 indivíduos e taxas de cruzamento de 0,8 e de mutação de 0,002. O algoritmo foi programado para terminar o processamento em 100 iterações ou quando a razão entre o erro padrão da melhor solução e da média da população fosse menor do que 0,01 por mais de 5 gerações seguidas (critério que geralmente encerrou o processamento).

A seleção para a mutação foi realizada em duas etapas, inicialmente selecionando aleatoriamente o indivíduo e em seguida selecionando um gene deste indivíduo. A taxa de mutação adotada corresponde a 5 mutações por iteração do algoritmo. O esquema de mutação empregado adotou um amortecimento progressivo dos efeitos da mutação (Equação 51):

$$Gene_{(i,\,g+1)}=Gene_{(i,\,g)}*[1+((\lambda-0{,}5)*(1-g/G))] \qquad \text{(Equação 51)}$$

Onde $Gene_i$ (i=1,...,24) é o gene selecionado para mutação, g é o número da geração atual (g=1,...,100), λ é um número aleatório no intervalo [0,1] e G é o total de gerações para o qual o algoritmo foi programado (no caso, 100). Este amortecimento tem a finalidade de diminuir a probabilidade de prejuízo para as soluções mais ajustadas, obtidas com o progresso do algoritmo.

O modelo Mercado estabilizou o nível de erro após 25 gerações, obtendo-se o modelo da Equação 52. Não há surpresas em quanto aos coeficientes das variáveis e seus sinais e os expoentes foram todos próximos ou iguais a 1, bem como o expoente geral foi próximo a 4/3 (ver Equação 17).

$$\begin{aligned} Valor=(&492{,}800+36{,}133*Artocons^{1,0073}+873{,}769*\ Razão^{1,0174}+ \\ &518{,}277*Med+1.586{,}775*Fin+4.190{,}182*Lux-32{,}541*\ Idade^{0,9886}+ \\ &23{,}669*Bairro^{0,9861}-43{,}120*Centro^{1,0495}-90{,}847*Comércio^{1,0332}+ \\ &1.640{,}422*Vista-19{,}462*Mês^{1,006})^{3,9746/3} \end{aligned} \qquad \text{(Equação 52)}$$

No caso dos imóveis Pequenos, o modelo foi obtido em 75 gerações, obtendo-se o modelo da Equação 53. Igualmente, os coeficientes, expoentes e sinais são similares aos da regressão correspondente (Tabela 19).

$$Valor=(425{,}240+37{,}712*Artocons^{0{,}9792}+795{,}873*\ Razão^{1{,}0193}+ \quad \text{(Equação 53)}$$
$$549{,}059*Med+1.347{,}676*Fin+2.183{,}953*Lux-25{,}119*\ Idade^{1{,}0136}+$$
$$15{,}889*Bairro^{0{,}992}-36{,}927*Centro^{1{,}0019}-96{,}602*Comércio^{0{,}9829}+$$
$$1.145{,}680*Vista-19{,}978*Mês^{0{,}988})^{4{,}028/3}$$

O grupo dos imóveis Médios teve seu modelo definido em 30 gerações, com coeficientes atingindo valores razoáveis e expoentes dos atributos também próximos da unidade (Equação 54).

$$Valor=(1.202{,}715+32{,}541*Artocons^{0{,}9982}+579{,}269*Razão^{0{,}9787}+ \quad \text{(Equação 54)}$$
$$819{,}063*Med+2.265{,}062*Fin+5.108{,}690*Lux-46{,}463*Idade^{0{,}9675}+$$
$$25{,}298*Bairro^{1{,}007}-122{,}315*Centro^{1{,}0001}-55{,}425*Comercio^{0{,}9909}+$$
$$1.778{,}914*Vista-27{,}176*Mes^{0{,}9975})^{4/3}$$

Para os imóveis maiores, o ajustamento se deu em 95 gerações, obtendo-se o modelo da Equação 55. As variáveis Med e Comércio foram testadas, mas foram eliminadas na evolução das soluções e não participam do modelo final. A variável Vista não participa por não existirem casos com esta característica no terceiro grupo de imóveis. Os expoentes são praticamente unitários, existindo pequenas diferenças dos coeficientes em relação ao modelo de regressão (Tabela 19).

$$Valor=(2.094{,}236+27{,}230*Artocons^{0{,}999}+6.411{,}882*Razão^{0{,}9995}+ \quad \text{(Equação 55)}$$
$$3.440{,}786*Fin+7.449{,}926*Lux-166{,}188*Idade^{0{,}9995}+35{,}030*Bairro^{1{,}0029}-$$
$$668{,}074*Centro^{1{,}0054}-79{,}793*Mês^{0{,}9996})^{4/3}$$

Em geral, os coeficientes foram coerentes com o esperado, em termos de sinais e valores

absolutos, inclusive no relacionamento entre as variáveis binárias (Med/Fin/Lux). A união das séries de estimativas para gerar o modelo de avaliação em massa indicou resultados similares aos da regressão. Os resultados obtidos com os modelos aditivos generalizados foram os seguintes (Tabela 29). Os erros obtidos são muito semelhantes aos da regressão, sem vantagens para nenhuma das duas técnicas, com respeito a este elemento.

Tabela 29: Resultados obtidos com os modelos aditivos generalizados

Modelo	Treino			Teste		
	EA	EP	N	EA	EP	N
MERCADO	17,51	19.323	24.289	18,01	20.038	6.074
Pequenos	17,76	11.302	21.148	17,92	11.333	5.288
Médios	13,22	30.651	2.838	13,31	29.991	710
Grandes	11,44	73.865	303	10,21	76.211	76
MASSA	17,18	16.932	24.289	17,29	17.019	6.074

Os formatos dos modelos também foram similares aos obtidos com a regressão, com a maioria dos expoentes das variáveis independentes muito próximos à unidade. A variável dependente teve expoentes próximos ou iguais ao expoente de 3/4, confirmando a análise de Box-Cox. Neste caso, pode-se concluir que os modelos hedônicos de regressão apresentados acima (Tabela 19) são adequados, ou seja, não há evidências de fugas sensíveis à linearidade.

7.2.5 Extração de regras difusas através de algoritmos genéticos

As regras difusas constituem uma forma interessante de consideração das imprecisões típicas do mercado imobiliário, tais como as transições entre regiões ou tipos de imóveis contíguos (interfaces entre sub-mercados). Nesta alternativa verificou-se a extração de regras difusas a partir da base de dados, com ajustamento das regras através de algoritmos genéticos, formando um Sistema Baseado em Regras Difusas Evolucionário (SBRDE). Foram utilizadas regras do tipo TSK, as quais são formadas por equações lineares e podem ser vistas como modelos hedônicos.

Foram estimados dois tipos de conjuntos difusos, utilizando diferentes abordagens. No primeiro, foi estimado um SBRDE com funções de pertinência baseadas na área total dos

imóveis (variável Artocons), desenvolvendo uma alternativa difusa comparável aos modelos apresentados acima. Foram investigados sistemas de regras com 3 a 7 regras, determinados através da abordagem Pittsburgh.

No segundo caso, foi construído um sistema de regras baseado em funções de pertinência que contemplam a localização. A fundamentação é a continuidade espacial do mercado, considerada através da união difusa das várias regras, cada uma delas "especializada" em determinada região, mas também contribuindo para a estimação nas demais, principalmente nas regiões contíguas. Este formato utilizou a abordagem Michigan, que permite maior flexibilidade na definição do número de regras, as quais são estimadas individualmente.

Nas duas abordagens, uma das vantagens observadas é a possibilidade de utilizar-se o conhecimento disponível, tais como modelos determinados por outras técnicas, como estimativa inicial (semente) para a geração das populações de teste. Outra característica é a flexibilidade do sistema.

Em função das características da estimação por regras difusas, não há modelos isolados para os grupos, mas um modelo único, sendo as estimativas compostas difusamente, geralmente com a participação de mais de uma regra no cálculo dos valores. Assim, o resultado é de apenas um modelo, ao contrário dos sistemas anteriores, nos quais foram desenvolvidos modelos com base no conjunto Mercado e nos sub-grupos (divididos por área).

7.2.5.1 Sistema difuso com base na área total

Este sistema tem regras no formato da Equação 56, sendo que a parte antecedente das regras tem os conjuntos difusos em função da área e a parte conseqüente tem equações do tipo TSK, sendo que as regras R_i são estimadas simultaneamente através de algoritmos genéticos.

$$R_i\text{: SE Área total é } A_i \text{ ENTÃO } Valor_i = (modelo_i) \qquad \text{(Equação 56)}$$

As funções de pertinência A_i foram definidas em formato triangular, apenas com as funções dos extremos em formato trapezoidal. No caso mais geral, com 7 parcelas, o conjunto de

funções de pertinência assume um formato como o da Figura 32, na qual os imóveis estão divididos em "muito pequenos" (mP), "pequenos" (P), "pequeno-médios" (PM), "médios" (M), "médio-grandes" (MG), "grandes" (G) e "muito grandes" (mG). Variando as áreas-limite pode-se obter diferentes configurações de sistemas, com diferentes coberturas para cada regra (faixas de atuação das regras).

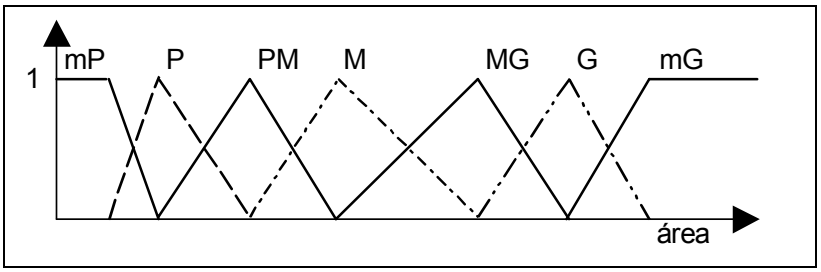


Figura 32: Exemplo de conjuntos difusos para a área total

Os limites de cada um dos conjuntos difusos foram determinados simultaneamente com os coeficientes dos modelos (codificados no mesmo cromossomo), porém baseados apenas em seleção (os valores originais, definidos aleatoriamente, não foram alterados). Esta estratégia foi preferida em função da maior estabilidade de evolução dos sistemas, detectada em testes iniciais. Os indivíduos foram definidos com a seguinte estrutura (Figura 33).

a_1	...	a_{12}	...	g_1	...	g_{12}	h_1	...	h_7

Figura 33: Aspecto dos indivíduos utilizados nos algoritmos genéticos para regras difusas

Onde os conjuntos $a_1,...,a_{12}$ a $g_1,...,g_{12}$ representam os coeficientes para a parte conseqüente de cada regra (cada um com 11 coeficientes para as variáveis e um intercepto), e $h_1,...,h_7$ são os limites das funções de pertinência (em termos de área total), conforme Figura 32. Cada indivíduo representa um sistema de regras completo. Por exemplo, os cromossomos para os conjuntos de 3 regras têm 39 genes (3x12+3), enquanto que os de 7 regras têm 91 genes (7x12+7).

A geração da população inicial utilizou o conhecimento disponível, na forma de equações

geradas pelos modelos anteriores. Para o conjunto de 3 regras, os indivíduos iniciais foram baseados nas equações determinadas anteriormente (ver Tabela 19). Nos modelos com 5 e 7 regras foram desenvolvidas regressões com base em 5 e 7 *clusters*, igualmente com base no critério "área total". Em todos os casos, foram aplicadas variações aleatórias de ±50% sobre os coeficientes para a geração da população inicial.

Em função das regras envolverem diferentes tamanhos de imóveis, com diferentes valores totais, a competição através do erro padrão (EP) privilegia os imóveis menores, os quais têm erros menores, em termos absolutos, e que são a maioria em número de casos. Assim, a evolução das regras tende a produzir um conjunto bem adaptado aos imóveis pequenos mas progressivamente inadequado para os médios e grandes. Para evitar esta situação, a função de ajustamento (*fitness*) baseou-se no erro absoluto percentual médio (EA), e teve o seguinte formato (Equação 57).

$$F_i=1/(1+EA_i), \qquad \text{com } EA_i=\Sigma_{i,j}\,[|Y_j-Y^h_{i,j}| / Y_j * 100\,] \qquad \text{(Equação 57)}$$

Onde F_i é a estimativa de *fitness* para a regra i, EA_i é o erro absoluto percentual médio calculado para a regra i com os dados de treinamento, Y_j é o preço observado e $Y^h_{i,j}$ é o valor estimado pela regra i para o caso j.

Da mesma forma que no modelo anterior, os operadores genéticos utilizados foram cruzamento e mutação. Os parâmetros foram explorados nos mesmos intervalos, ou seja, foram testadas populações de 50 a 100 indivíduos, taxas de cruzamento de 0,6 a 0,95 e taxas de mutação na faixa de 0,001 a 0,2. Os melhores resultados ocorreram com populações de 100 indivíduos, taxas de cruzamento de 0,8 e de mutação de 0,008 (0,002 em um dos casos) e seleção elitista de 5% dos indivíduos. O algoritmo foi programado para parar em 100 iterações ou quando a razão entre o erro absoluto da melhor solução e o erro médio da população fosse menor do que 0,01 por mais de 5 gerações seguidas.

Os resultados indicaram equilíbrio entre os sistemas desenvolvidos. Para o conjunto de três regras, o erro estabilizou-se em 50 gerações, sendo atingido um nível de erro padrão com os dados de teste de 16.833 e de 18,11 nos erros absolutos percentuais. Para o sistema com 5 regras, os erros foram de 18.822 e 18,62, respectivamente, sendo que o conjunto de regras foi

obtido em 95 gerações. Finalmente, o conjunto de 7 regras foi definido em 65 gerações e os erros foram de 17.718 e 18,35. Em face do equilíbrio entre os conjuntos testados, a preferência deve recair pelo sistema mais simples, ou seja, o conjunto com o menor número de regras, em função da interpretação mais fácil, além de possuir efetivamente o menor erro dentre os modelos estimados, no caso. O conjunto final, composto por 3 regras difusas, está apresentado nas Equações 58 a 60. Os conjuntos difusos A_1, A_2 e A_3 são definidos com base nas áreas de 11,265m^2, 236,326m^2 e 454,652m^2, respectivamente, sendo os dois extremos em formato trapezoidal e o central em formato triangular (ver Figura 34).

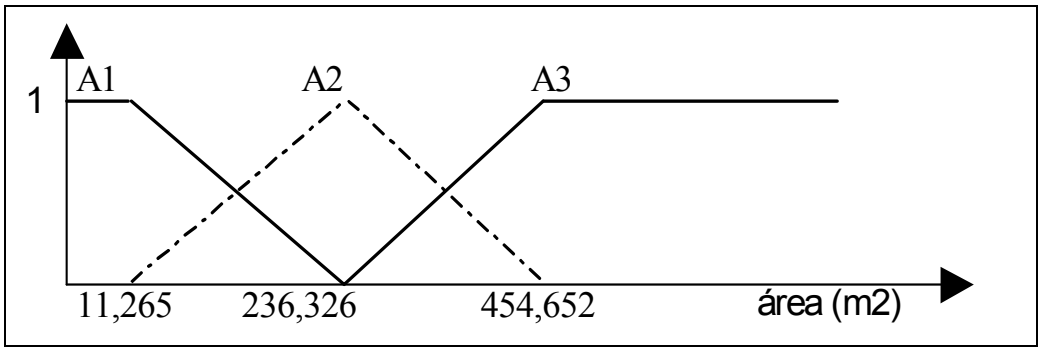


Figura 34: Conjuntos difusos para o sistema difuso baseado na área total com três regras

R_1: SE Área é A_1 ENTÃO $Valor_1$=(363,399+ 38,225*Artocons*+ 818,141*Razão+544,665*Med+1.016,166*Fin+ 2.107,615*Lux– 24,088*Idade+14,932*Bairro–33,617*Centro–94,569*Comércio+ 1.399,788*Vista–14,880*Mês)$^{4/3}$ (Equação 58)

R_2: SE Área é A_2 ENTÃO $Valor_2$=(1.147,260+33,432*Artocons+ 589,584*Razão+838,673*Med+2378,054*Fin+4.599,690*Lux– 50,621*Idade+26,549*Bairro–118,912*Centro–64,728*Comércio+ 1.917,465*Vista–28,918*Mês)$^{4/3}$ (Equação 59)

$$R_3\text{: SE Área é } A_3 \text{ ENTÃO Valor}_3=(1.414{,}908+27{,}101*\text{Artocons}+5.593{,}242*\text{Razão}+4.565{,}306*\text{Fin}+7.516{,}682*\text{Lux}-43{,}623*\text{Idade}+49{,}287*\text{Bairro}-726{,}720*\text{Centro}-103{,}582*\text{Mês})^{4/3} \quad \text{(Equação 60)}$$

Os coeficientes e sinais dos modelos apresentados são coerentes com o esperado, não revelando surpresas. Os resultados estão apresentados na Tabela 30. Há apenas um modelo difuso para representar os modelos de Mercado e de avaliação em massa, com equilíbrio com os resultados da regressão, sendo que o modelo difuso é superior ao modelo de Mercado com regressão, com erros praticamente iguais aos da regressão para o modelo de avaliação em massa. Os resultados calculados para os dados dos três grupos estão apresentados na Tabela 30, para comparação com os resultados anteriores.

Tabela 30: Resultados para as regras difusas extraídas da base de dados

Modelo	Treino			Teste		
	EA	EP	Casos	EA	EP	Casos
MERCADO (=MASSA)	17,30	16.862	24.289	17,56	16.980	6.074
Pequenos	17,87	11.117	21.148	18,16	11.235	5.288
Médios	13,70	30.076	2.838	13,93	30.526	710
Grandes	11,37	74.835	303	9,63	74.522	76

7.2.5.2 Sistema difuso com base na localização

A segunda alternativa envolve a localização dos imóveis. Cada regra é "especializada" em uma determinada região da cidade, cujo centro foi definido por análise de clusterização das coordenadas (X,Y) com os dados de treinamento. Foram inicialmente determinadas 10 regiões (correspondendo a 10 regras). O esquema adotado para o algoritmo genético foi semelhante ao do sistema baseado na área, exceto que as regras são individualmente apreciadas e a função de pertinência é espacial, contabilizando a distância de cada imóvel ao centro de referência da regra através das coordenadas dos imóveis (Equação 61):

$$M_{i,j} = 1 / (1 + [(X_i - X_j)^2 + (Y_i - Y_j)^2]^{0,5.k}) \qquad \text{(Equação 61)}$$

Onde $M_{i,j}$ é o valor da pertinência do caso j à regra i, (X_i,Y_i) é o centro da regra i, (X_j,Y_j) são as coordenadas de cada imóvel j e k é um expoente, que permite variar a abrangência da função. O expoente k foi investigado, obtendo-se os melhores resultados com expoente unitário para todas as regras (k=1). As funções de pertinência empregadas definem na verdade regiões de pertinência no espaço urbano, obtidas pela giração (em 360^o) da curva exponencial $M_{i,j}$. Em função do formato desta função, o valor de pertinência não é normalizado, ou seja, não está restrito ao intervalo [0,1], e depende da proximidade dos centros das regras, entre si.

As regras foram examinadas utilizando a pertinência $M_{i,j}$ como uma penalidade na função de *fitness*, para forçar o ajustamento localizado. Com esta finalidade, a função de ajuste tomou a seguinte forma (Equação 62):

$$F_i = 1/(1+EA_i'), \quad \text{com } EA_i' = \Sigma_{i,j} [|Y_j - Y^h_{i,j}| / Y_j * 100 / M_{i,j}] \qquad \text{(Equação 62)}$$

Onde F_i é a medida do ajustamento da regra i, EA_i' é a medida de erro percentual modificada e os outros parâmetros são os descritos acima. Com a aplicação desta penalidade, as regras mais ajustadas aos casos próximos ao centro da região são beneficiadas, pois os erros dos casos distantes (com pertinência pequena) são inflados.

Na primeira fase (geração da base inicial), cada uma das 10 regras foi ajustada individualmente, com populações de 100 regras similares, inicialmente baseadas nas regras obtidas por regressão tradicional (ver Tabela 19), optando-se pelos modelos mais adequados aos imóveis da região correspondente (conforme a área total média no local) e aplicando-se variações aleatórias de ±50% para gerar estas populações. As regras foram ajustadas na ordem do número de casos contidos em cada *cluster* (ou seja, a regra R_1 corresponde ao grupo com maior quantidade de casos). Para facilitar a compreensão das regras, elas foram identificadas com as regiões da cidade correspondentes às regiões de pertinência maior (ver Tabela 31, regras R_1 a R_{10}).

Na segunda fase (refinamento da base), verificou-se o desempenho da base de regras inicial, identificando as regiões da cidade com os maiores erros e incluindo progressivamente novas regras para estas regiões. A base final contou com 15 regras, acrescentando-se regras para compensar erros localizados (no espaço) ou para determinados tipos de imóveis (regras R_{11} a R_{15}, Tabela 31). Com este conjunto, os erros obtidos com o conjunto de dados de teste foram de 20.249 e 21,27%, para EP e erros percentuais, ou seja, mesmo com o refinamento os resultados finais são inferiores aos obtidos no modelo difuso anterior, baseado na área total.

Tabela 31: Centros das funções de pertinência para o SBRDE baseado na localização

Regra	Região	Centro	
		X	Y
R_1	Centro/Cidade Baixa	0,659	-0,673
R_2	Petrópolis	3,728	-1,570
R_3	São João/Higienópolis	3,597	1,708
R_4	Azenha	1,287	-3,136
R_5	Rubem Berta	11,566	2,063
R_6	Cristo Redentor	7,176	1,818
R_7	Cavalhada/Cristal	-0,678	-7,427
R_8	Bom Jesus	7,654	-1,903
R_9	Vila Nova	1,427	-10,996
R_{10}	Restinga	8,082	-13,574
R_{11}	Floresta/Moinhos de Vento (imóveis médios)	2,200	0,780
R_{12}	Bela Vista	4,000	-0,400
R_{13}	Humaitá	3,850	5,440
R_{14}	Moinhos de Vento/Mont'Serrat (imóveis grandes)	3,200	0,286
R_{15}	Centro	–0,528	-0,394

O procedimento de acréscimo de regras poderia ser estendido, possivelmente com o aperfeiçoamento das estimativas, porém com incremento na complexidade e na dificuldade de compreensão da base. Outra forma de aperfeiçoamento é a subdivisão de cada uma das regras em conjuntos de regras especializados em faixas de áreas, adotando um sistema misto das duas alternativas apresentadas, com área e distância influindo simultaneamente no cálculo da pertinência às regras.

7.2.6 Análise dos modelos de avaliação coletiva

Foi desenvolvido um exame geral das técnicas. Além dos parâmetros de erro já apresentados (EP e EA), os quais foram utilizados no ajustamento e na seleção inicial dos modelos, foram calculados o coeficiente de dispersão (COD), o coeficiente de correlação entre os valores originais da variável dependente e as estimativas dos modelos ($r_{y.y}^{h}$) e foram somados os casos com erros absolutos menores do que 5% e 10% e maiores do que 50%. Os resultados estão apresentados na Tabela 32, sendo todos baseados no conjunto de dados de teste (com 6.074 casos). As redes neurais foram incluídas nas tabelas para permitir a comparação com as demais técnicas, e principalmente com as regras extraídas destas redes. Contudo, como as redes neurais não são uma alternativa completa, pela falta de explicação dos resultados, não foram incluídas nos cálculos das médias.

Tabela 32: Comparação entre os modelos de Mercado e de avaliação em massa e entre as técnicas empregadas

Técnica	Modelo									
	MERCADO					MASSA				
	COD	$r_{y.y}^{h}$	<5%	<10%	>50%	COD	$r_{y.y}^{h}$	<5%	<10%	>50%
Regressão tradicional	18,7	0,9762	1141	2285	311	17,8	0,9821	1262	2403	277
Regressão – superfícies	17,8	0,9763	1218	2428	284	18,6	0,9775	1218	2312	263
Redes neurais (isoladas)	16,4	0,9807	1339	2535	158	15,7	0,9832	1406	2612	203
Regras extraídas das redes	22,1	0,9726	1062	2050	506	21,5	0,9810	1248	2371	561
Modelos aditivos	18,4	0,9762	1234	2367	267	17,6	0,9821	1299	2451	219
Regras difusas	17,5	0,9820	1216	2453	237	17,5	0,9820	1278	2453	237
Médias[53]	*18,9*	*0,9767*	*1174*	*2317*	*321*	*18,6*	*0,9809*	*1261*	*2398*	*311*

Um dos pontos importantes a ser destacado é o elevado nível médio do coeficiente de dispersão. Organismos internacionais, como a IAAO, recomendam que o coeficiente de dispersão assuma valores menores do que 15% nos modelos de avaliação em massa. Embora esta recomendação seja destinada a imóveis residenciais unifamiliares, pode-se aplicar para apartamentos, por analogia (IAAO, 1990). No caso, o COD é efetivamente elevado, pois a média dos resultados para os modelos gerados é de 18%. Os resultados, entretanto, dependem também da qualidade dos dados empregados. Os dados empregados neste trabalho são dados

[53] Sem os resultados das redes neurais (consideradas isoladamente).

fiscais, os quais podem conter tendências ou outros problemas que os afastem do comportamento real do mercado imobiliário, e o próprio tamanho da amostra induz a uma maior heterogeneidade e dificuldade de geração de modelos. A divisão da base em diversas parcelas ou a alteração dos níveis dos testes para *outliers* poderiam diminuir os erros. Há indícios neste sentido, pois os modelos com os grupos de imóveis Médios e Grandes em geral atingiram erros menores, enquanto que os modelos gerais e para os imóveis Pequenos, com um conjunto maior de dados, atingiram erros maiores. Por exemplo, sem alterar os modelos, apenas excluindo os casos com erros absolutos acima de 50%, foram obtidos coeficientes de dispersão na ordem de 15%, melhorando também os demais indicadores. Da mesma foram, se os modelos fossem re-estimados sem estes dados, provavelmente seriam obtidos modelos com desempenho melhor.

Todavia, para a comparação entre as técnicas utilizadas, é mais importante o nível de erro relativo. Neste sentido, a correlação entre o valor original e as estimativas é alta ($r_{y.y}^{h}$>0,97 em todos os modelos), mas há diferenças nos outros indicadores de erro. Os modelos tiveram em média 5% de casos com erros absolutos maiores do que 50%. Detalhando esta medida, foram detectados pouco mais de 100 casos em que os imóveis atingiram mais de 50% de erro absoluto em todas as técnicas (cerca de 2% dos casos de teste), enquanto que mais de 5.200 casos não atingiram este nível de erro em nenhum dos modelos, ou seja, mais de 85% dos casos. Por outro lado, quanto aos erros pequenos, todos os modelos tiveram cerca de 20% de erros absolutos menores do que 5% e mais de 1/3 dos casos com erros absolutos menores do que 10%.

Considerando todos os indicadores de erro, os piores modelos foram os das regras extraídas das redes neurais e o de superfícies baseadas na amostra segmentada por área, enquanto que os de menor erro foram os baseados em regras difusas. Por exemplo, o sistema de regras difusas atingiu um incremento de cerca de 6% no coeficiente de dispersão, em relação à regressão (Tabela 32). Nos modelos de Mercado, as superfícies e os modelos difusos são os melhores. A divisão da base revelou-se interessante, pois os modelos de avaliação em massa são superiores aos modelos de Mercado, exceto no caso das superfícies, inclusive com mais casos dentro das faixas dos 5% e 10% de erro, também com exceção das superfícies.

7.3 MODELOS PARA AVALIAÇÃO INDIVIDUAL

Para examinar a aplicação das técnicas selecionadas em tarefas de avaliação individual foram escolhidos seis casos, de diferentes tipos e localizações, para os quais foram ajustados os modelos segundo as várias técnicas. Os resultados em geral foram similares, em termos de desempenho relativo das técnicas. Para evitar repetições foi escolhido apenas um destes imóveis para discussão dos modelos. O imóvel avaliando escolhido é um apartamento de 138m^2 de área privativa e 10 anos de idade, de padrão Fino (Fin=1), situado na Av. Getúlio Vargas, esquina com a Rua Botafogo, no bairro Menino Deus (Bairro=42), com coordenadas (0,650; –2,900), e ainda com Razão=0,75, Vista=0, distância ao centro de 2,972km e distância ao comércio de 0,912km. O valor estimado deve ajustar-se ao mês de agosto de 2001, época da coleta dos dados (Mês=37).

7.3.1 Identificação de casos similares

A avaliação de um caso isolado pode ser aprimorada com a separação de uma amostra mais homogênea da base contendo os casos de maior similaridade principalmente quanto à localização, variável de mais difícil medição.

A forma de seleção para a avaliação individual foi proposta no capítulo 5 (Equação 24) e emprega um mecanismo de vizinhança próxima associando as características relevantes com uma medida da distância entre os imóveis. Considerando como base os coeficientes da equação de regressão para o modelo de Mercado (Tabela 19) e calibrando os coeficientes k_0 e k_1, a medida de similaridade dos casos da base para o avaliando foi então calculada pela Equação 63.

$$SIM_{a,b} = 1/[(1+0{,}01*(X-0{,}650)^2+(Y+2{,}900)^2)*(1+0{,}000001*(35{,}260*(Artocons-138)^2+883{,}674*|Razão-0{,}75|+510{,}359*Med+1734{,}842*|Fin-1|+4251{,}930*|Lux-1|-27{,}447*|Idade-10|+20{,}615*(Bairro-37)^2-43{,}290*|Centro-2{,}720|-90{,}022*|Comércio-0{,}912|+1618{,}714*|Vista|-19{,}864*|Mês-37|)^{4/3})]*100\% \quad \text{(Equação 63)}$$

Os componentes seguem as definições apresentadas no capítulo 5. A Equação 63 foi testada, verificando-se o grau de acerto das amostras recuperadas. Para aprimorar os resultados, foram adotados expoentes iguais a dois para as variáveis Artocons e Bairro, em função da maior importância do tamanho dos imóveis e do bairro, em relação às outras variáveis. Com esta configuração, os testes realizados indicaram que a seleção realizada era adequada, identificando os casos relevantes.

Após o ordenamento dos casos da base pela sua semelhança com o avaliando, foram selecionados os casos com similaridade acima de 90%, resultando em 347 casos, formando uma amostra de tamanho razoável para utilização com técnicas de inteligência artificial, de um lado, e um conjunto de grande homogeneidade, de outro. A descrição dos dados está na Tabela 33. Verifica-se que algumas situações não estão representadas na amostra, tais como imóveis com padrão Popular ou Luxo e imóveis com vista panorâmica, reduzindo o número de variáveis a serem analisadas.

Tabela 33: Descrição do conjunto de imóveis selecionados para avaliação individual (com base em 347 casos)

Variável	Unidade	Mínimo	Máximo	Média	Desvio padrão
Artocons	m^2	69,76	180,84	125,49	24,36
Razão	-	0,40	1,00	0,73	0,09
Tipo2	-	3	4	3,41	0,50
Pop	-	0	0	0	-
Med	-	0	1	0,59	0,49
Fin	-	0	1	0,41	0,49
Lux	-	0	0	0	-
Idade	anos	0	50	14,07	11,25
X	km	-0,17	1,12	0,52	0,32
Y	km	-4,00	-1,75	-2,95	0,53
Bairro	-	29	37	36,91	0,86
Centro	km	1,78	4	3,02	0,51
Comércio	km	0,24	1,76	1,02	0,34
Vista	-	0	0	0	-
Mês	mês	5	36	25,58	8,57
Valor	R$	46.603,74	237.448,80	126.925,01	38.548,39

O posicionamento dos casos pode ser identificado através da Figura 35, na qual o losango identifica o imóvel avaliando e os pontos são os 347 casos da amostra, que estão situados em 137 prédios, ou seja, há em média 2,5 casos da amostra em cada localização.

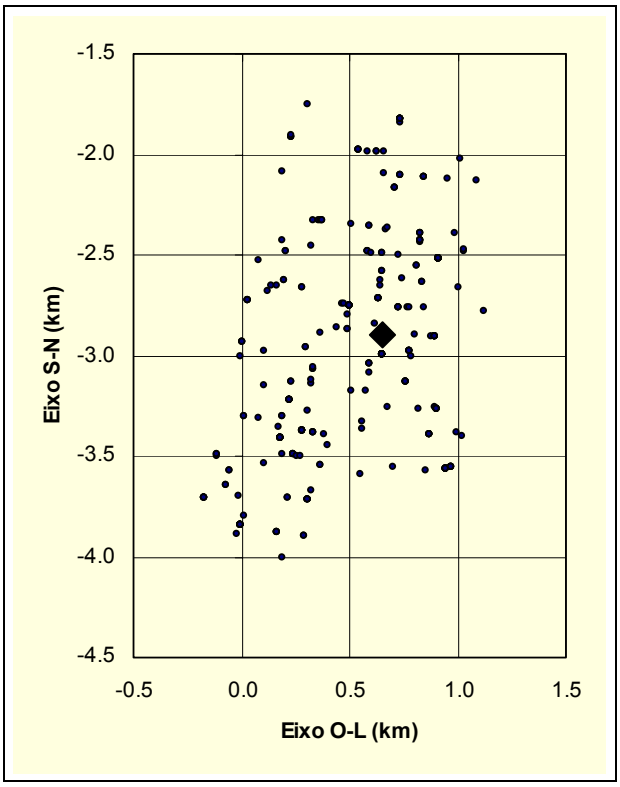


Figura 35: Posição dos casos selecionados para desenvolver os modelos de avaliação individual

7.3.2 Complementação dos dados com vistorias

Para verificar os efeitos do detalhamento sobre os dados, foram desenvolvidas vistorias no local, aplicadas sobre uma amostra-piloto de 115 casos (aproximadamente 1/3 da amostra), selecionada aleatoriamente. Nesta análise de campo foram produzidas fotografias dos imóveis, com análise da micro-localização e do padrão de construção dos prédios, bem como levantamento de informações sobre o uso predominante (residencial/comercial), posição do imóvel no quarteirão e a altura do prédio, medida em pavimentos. As variáveis coletadas foram as seguintes:

a) Vizinhança: variável qualitativa que indica a classificação da vizinhança (de 1 a 5);

b) Padrão: variável qualitativa para o padrão construtivo do prédio (de 1 a 5);

c) Uso_Com: variável binária que indica o uso prioritário nas proximidades do imóvel (100m) – 0=residencial, 1=comercial;

d) Esquina: indica se o imóvel é de esquina (1) ou não (0);

e) Andares: número de pavimentos do prédio;

Os novos dados foram preparados, seguindo-se a mesma seqüência utilizada no capítulo anterior, com os resultados apresentados na Tabela 34.

Tabela 34: Variáveis coletadas nas vistorias de campo (com base em 115 casos)

Variável	Tipo	Mínimo	Máximo	Média	Desvio-padrão
Vizinhança	Discreta	1	5	2,52	1,20
Padrão	Discreta	1	10	5,58	1,54
Uso_Com	Binária	0	1	0,17	0,38
Esquina	Binária	0	1	0,084	0,28
Andares	Discreta	2	14	7,96	3,10

A análise do desempenho destas variáveis através de visualização indicou, inicialmente, o relacionamento com algumas das variáveis anteriores. A matriz de correlações apresentada na Tabela 35 indica bom relacionamento das variáveis de Vizinhança, Padrão e número de pavimentos do prédio com o valor dos imóveis, sem serem notadas colinearidades com as variáveis independentes anteriores.

Tabela 35: Matriz de correlações das novas variáveis independentes com a dependente e algumas independentes (com base em 115 casos)

Variáveis	Vizinhança	Padrão	Uso_Com	Esquina	Andares
Artocons	0,589	0,596	-0,233	-0,217	0,726
Razão	-0,451	0,323	0,267	0,370	-0,569
Fin	0,386	0,605	-0,397	-0.063	0,564
Idade	-0,647	0,620	0,329	0,167	-0,596
Comércio	0,404	0,393	0,061	-0,021	0,427
Mês	0,184	-0,106	0,164	0,025	0,108
Valor	0,646	0,795	-0,233	-0,182	0,769

Estas variáveis foram testadas através da abordagem de envoltório, com os mesmos

procedimentos e técnicas empregados no capítulo 6 (ver item 6.5), empregando os 115 casos para os quais as variáveis foram coletadas. Em alguns modelos estas variáveis não tiveram desempenho superior às variáveis anteriores. Em outros, contribuíram com pequena redução do nível de erros. Assim, tendo em vista que estas variáveis exigem considerável custo e tempo de coleta, o procedimento de campo não foi estendido aos demais casos da amostra para avaliação individual. Os imóveis foram então divididos em dois conjuntos, seguindo a proporção de 80% para treinamento e 20% para teste.

7.3.3 Desenvolvimento dos modelos de avaliação individual

Inicialmente foram desenvolvidos os modelos de regressão, no formato tradicional e utilizando superfícies matemáticas. Os modelos gerados para estas duas configurações tiveram razoável grau de ajustamento, com cerca de 89% de explicação, bem como as variáveis participantes do modelo tiveram coeficientes convencionais e significância abaixo de 5%, de acordo com as estatísticas *t* calculadas. As significâncias dos modelos, utilizando a distribuição *F*, também ficaram abaixo do limite de 5% (Tabela 36). Os resultados gerais em termos de erros cometidos estão apresentados adiante, na Tabela 37.

Tabela 36: Modelos hedônicos tradicional e com superfícies – modelos individuais – variável dependente: $Valor^{3/4}$

Variável	Técnica			
	Regressão		Superfícies	
	Coeficiente	t	Coeficiente	t
Intercepto	620,833	1,5	871,828	2,2
Artocons	45,505	33,3	45,276	33,7
Razão	1.397,825	3,6	1.187,644	3,2
Fin	833,974	11,0	859,500	11,8
Idade	-37,539	-10,4	-33,214	-9,4
Comércio	-189,103	-2,0	-399,236	-3,7
Mês	-13,386	-3,6	-15,637	-4,4
XY^5	-		-3,324	-5,3
X^2Y^4	-		-9,826	-4,2
R^2_a	0,889		0,899	
F_{calc}	369,6		308,3	

O modelo seguinte foi o de redes neurais, complementado pela extração de regras difusas. A

rede neural construída com os dados selecionados atingiu bons resultados, bem como a regra difusa extraída a partir da rede (ver Tabela 37), que assumiu a seguinte configuração (Equação 64):

$$R_1\text{: SE } \textit{Verdadeiro} \text{ ENTÃO Valor}=(796{,}054+45{,}941*\text{Artocons}+ 1.348{,}791*\text{Razão}+781{,}506*\text{Fin}-40{,}896*\text{Idade}-250{,}842*\text{Comércio}-15{,}646*\text{Mês})^{4/3} \quad \text{(Equação 64)}$$

Em seguida, foi ajustado o modelo aditivo generalizado, com os mesmos dados. O modelo foi desenvolvido utilizando algoritmos genéticos, com funcionamento similar ao utilizado para os modelos coletivos. O algoritmo convergiu em 107 iterações, adotando-se taxa de cruzamento de 0,8, de mutação de 0,002, com uma população de 100 indivíduos, gerada com variações aleatórias aplicadas sobre o modelo de regressão da Tabela 36. A equação desenvolvida apresenta diversos termos não lineares, mas ainda próximos à unidade, bem como o expoente geral está próximo ao expoente de Box-Cox (Equação 65).

$$\text{Valor}=(759{,}768+44{,}681*\text{Artocons}^{0{,}9966}+ 1294{,}125*\text{Razão}^{0{,}9101}+ 786{,}657*\text{Fin}-44{,}171*\text{Idade}^{0{,}9577}-250{,}445*\text{Comércio}^{0{,}8659}-15{,}815*\text{Mês}^{0{,}9676})^{4{,}0162/3} \quad \text{(Equação 65)}$$

Por fim, foi desenvolvido o modelo baseado em regras difusas. Foram testados dois sistemas de regras, com duas e três regras, geradas pelo sistema Pittsburgh, com melhor desempenho do sistema de três regras. Foram utilizados algoritmos genéticos para o ajuste e a área total foi novamente utilizada como elemento de fuzzificação, com três conjuntos difusos, sendo os dois extremos trapezoidais e o central, triangular, com áreas-limite em 73,29m^2, 135,61m^2 e 194,64m^2, com formato similar ao da Figura 34. O sistema de regras gerado foi o seguinte (Equações 66 a 68).

R_1: SE Área total é A_1 ENTÃO Valor=(727,878+43,405*Artocons+ 1.393,087*Razão +790,694*Fin–35,076*Idade–233,988*Comércio– 14,153*Mês)$^{4/3}$ (Equação 66)

R_2: SE Área total é A_2 ENTÃO Valor=(762,759+ 46,089*Artocons+ 1.392,282*Razão+787,725*Fin–45,394*Idade–231,022*Comércio – 10,433*Mês)$^{4/3}$ (Equação 67)

R_3: SE Área total é A_3 ENTÃO Valor=(712,475+44,997*Artocons+ 1.377,973*Razão+792,861*Fin–52,667*Idade–255,711*Comércio– 19,043*Mês)$^{4/3}$ (Equação 68)

7.3.4 Análise dos modelos de avaliação individual

Após o desenvolvimento dos modelos para o caso de avaliação individual, foram procedidos diversos testes, especialmente quanto aos erros cometidos pelos modelos, utilizando os mesmos parâmetros utilizados para os modelos coletivos. A Tabela 37 apresenta os resultados para os modelos, bem como os valores estimados para o imóvel avaliando, descrito acima. Novamente as redes neurais são inseridas para comparação.

Tabela 37: Resultados para os modelos de avaliação individual

Técnica	Treino – 278 casos		Teste – 69 casos						Estimativa para o
	EA	EP	EA	EP	COD	$r_{y.y}^h$	e<5%	e<10%	avaliando
Regressão tradicional	8,94	12.612	7,52	13.013	7,6	0,9454	28	45	153.097,42
Regressão – superfícies	8,62	11.923	7,12	12.560	7,0	0,9492	30	48	153.233,06
Redes neurais (isoladas)	8,69	12.163	7,52	12.769	7,5	0,9439	35	48	151.122,93
Regras extraídas das redes	9,02	12.610	7,52	12.763	7,6	0,9470	28	47	152.376,50
Modelos aditivos	8,95	12.549	7,56	12.793	7,6	0,9466	27	45	152.776,38

Regras difusas	8,84	12.375	7,27	12.527	7,2	0,9508	27	47	156.795,49
Médias[54]	*8,87*	*12,414*	*7,40*	*12,731*	*7,4*	*0,9478*	*28*	*46*	*153.655,77*

Verifica-se que novamente há equilíbrio entre as técnicas exploradas, em todos os parâmetros estudados. Os modelos com superfícies e com regras difusas têm um desempenho levemente superior. Entretanto, na comparação destes resultados com os obtidos dos modelos coletivos há uma diferença importante. Não há casos com erros acima de 50%, e a parcela de erros menores do que 5% é de 42% da amostra, em média, enquanto que os casos com erros menores do que 10% representam mais de 2/3 da amostra de teste. Estes resultados são melhores do que os anteriores, obtidos com os modelos coletivos. Os valores estimados também são bastante similares, variando próximo a R$ 153 mil (Tabela 37).

Outra verificação foi desenvolvida utilizando os modelos de avaliação em massa para avaliar os mesmos casos de teste (69 casos). Os valores estimados apresentam evidente diferença quanto aos modelos Individuais. As estimativas baseadas nos modelos de avaliação em massa tiveram erros maiores, em todas as técnicas (Tabela 38). Os erros padronizados (EP) calculados pelos modelos individuais são cerca de 35% menores do que os calculados pelos modelos de avaliação em massa, bem como os erros absolutos percentuais (EA) são 20% menores. Verifica-se que a construção de modelos específicos propicia um bom incremento de precisão, e que todas as técnicas evoluem de forma similar.

Tabela 38: Resultados com os dados de teste da amostra para os modelos de avaliação individual aplicando os modelos de avaliação em massa

Técnica	"Teste" – 69 casos						
	EA	EP	COD	$R_{y,y}{}^{h}$	e<5%	e<10%	e>50%
Regressão tradicional	11,51	19.951	8,7	0,9311	19	31	2
Regressão - superfícies	12,03	20.468	9,1	0,9255	11	28	2
Redes neurais (isoladas)	10,72	18.494	8,1	0,9390	16	34	2
Regras extraídas das redes	11,45	20.314	8,5	0,9298	18	27	2
Modelos aditivos	11,57	19.890	8,7	0,9306	20	31	2
Regras difusas	10,52	17.780	8,3	0,9418	19	33	2
Médias[55]	*11,30*	*19.483*	*8,6*	*0,9330*	*17*	*31*	*2*

[54] Calculadas sem as redes neurais.

[55] Idem à nota anterior.

7.4 DISCUSSÃO DOS RESULTADOS

Os requisitos básicos exigidos para os modelos de avaliação, precisão e explicabilidade, foram atendidos de forma similar por todas as técnicas exploradas. Há equilíbrio nos níveis de erro, bem como os modelos obtidos são semelhantes em formato. Os modelos aditivos generalizados são levemente mais complexos e as regras extraídas das redes neurais são menos precisas, mas a diferença de ambos em relação aos demais não é grande.

A sensibilidade aos dados também é similar. As redes neurais, por exemplo, já foram apontadas como imunes aos efeitos de *outliers*, em um dos primeiros trabalhos com dados de mercado imobiliário (Tay e Ho, 1994), o que foi contestado posteriormente (Lenk *et al.*, 1997; Worzala *et al.*, 1995). As etapas de preparação e modelagem dos dados desenvolvidas indicaram que a presença de dados identificados como discrepantes diminuiu o desempenho das redes neurais de forma semelhante ao que ocorreu com a regressão. Com as outras técnicas o comportamento foi similar.

O tempo de processamento e análise é importante, embora seja influenciado por diversos fatores, tais como o domínio do analista sobre as técnicas e a capacidade dos *softwares* e da máquina utilizada. Porém, como uma indicação do esforço relativo, pode-se relatar que o ajustamento dos modelos baseados em regressão foi claramente mais rápido, enquanto que os modelos que dependem de um processo de otimização iterativa, tais como os algoritmos genéticos, exigiram de 20 a 50 vezes mais tempo de desenvolvimento, dependendo da complexidade dos modelos e da quantidade de casos dos conjuntos de treinamento e teste (fator que não afeta sensivelmente o ajustamento dos modelos de regressão). Deve ser levado em conta que os modelos de regressão têm seus coeficientes calculados com base no Método dos Mínimos Quadrados (MMQ), na maioria dos pacotes estatísticos, usando formulações previamente determinadas através de cálculo diferencial. Embora exigindo o exame dos pressupostos, geralmente através do exame de gráficos, é um processo potencialmente mais rápido. Além disto, existem *softwares* amigáveis e poderosos, o que geralmente não ocorre para os sistemas híbridos, que exigem mais conhecimentos do usuário.

7.4.1 Aperfeiçoamento dos modelos

Os resultados finais indicaram que o nível de erro de alguns modelos é razoavelmente elevado. Por exemplo, o erro absoluto dos modelos de avaliação em massa e de Mercado em geral foi de 17 a 18%. Porém, o erro absoluto dos modelos para os imóveis Grandes ficou abaixo dos 14%, para os apartamentos Médios foi de aproximadamente 10%, e para os modelos individuais foi de cerca de 7,5%. Estes resultados podem ser considerados bons, considerando as indicações da literatura e a finalidade exploratória do trabalho. Como há equilíbrio entre as técnicas para cada tipo de modelo estimado, pode-se concluir que o nível de erro decorre mais das características dos dados do que das técnicas empregadas.

De qualquer forma, há algumas alternativas de aperfeiçoamento. O nível de precisão absoluto poderia ser incrementado através de um maior rigor na seleção dos casos, por exemplo alterando o limite para seleção de *outliers* de 4 para 3 ou 2 desvios-padrão. Também podem ser acrescentados outros atributos, potencialmente importantes, tais como número de dormitórios e número de garagens, os quais não estavam disponíveis, no caso. Além de participarem nos modelos diretamente, aumentando a explicação dos preços, estes atributos podem identificar segmentos de mercado distintos, com diferentes perfis de compradores e possivelmente com diferentes preços hedônicos, pois os modelos podem variar para cada segmento.

Em termos relativos, a precisão das ferramentas poderia ser incrementada pela seleção de subconjuntos de casos e de atributos relevantes distintos para cada uma delas. Entretanto, a exploração realizada quando da seleção dos casos e atributos indicou que esta otimização conduzia a incrementos marginais, com pequeno impacto nos resultados finais.

A relevância dos casos e atributos poderia igualmente ser re-estudada para cada um dos sub-modelos, para identificar possíveis particularidades dos dados em função de comportamento diferenciado nos sub-mercados. Entretanto, apenas a identificação de grupos mais homogêneos já produziu modelos relativamente melhores (com menor nível de erro), com os mesmos atributos. No caso dos modelos de avaliação em massa, a subdivisão em maior quantidade de sub-grupos, especialmente no caso dos apartamentos Pequenos, poderia incrementar a precisão geral, mas com o aumento do número de modelos e do trabalho de modelagem. Um processamento semi-automatizado poderá auxiliar a solução deste dilema.

Ademais, os resultados obtidos podem ser explicados, em parte, pelas limitações dos dados utilizados. A variável dependente coletada consiste de estimativas fiscais, e não de preços de mercado, os quais estão indiretamente representados. Portanto o fenômeno que está sendo mapeado efetivamente é o procedimento adotado na avaliação dos imóveis, e não o comportamento do mercado. Parte das características originais dos dados pode ter sido perdida, bem como algumas tendências podem ter sido introduzidas. Por exemplo, o comportamento claramente linear dos dados provavelmente é um reflexo da adoção freqüente de tabelas prévias ou modelos lineares na estimação dos valores fiscais.

7.4.2 Relevância dos atributos

O exame do desempenho dos atributos deve levar em conta a contribuição marginal (impacto da variação do atributo na variação dos preços) e também o custo de coleta. Algumas variáveis disponíveis no cadastro municipal ou *proxies* baseadas no conhecimento anterior foram substitutas interessantes para atributos subjetivos, como no caso dos atributos coletados em vistorias realizadas em amostra-piloto, para o modelo individual.

Além do custo e tempo de obtenção, a própria variabilidade decorrente da subjetividade pode ser prejudicial. No caso de grandes bases de dados, os prazos de desenvolvimento do trabalho exigiriam a participação de vários analistas, e a diversidade de critérios é um empecilho a ser enfrentado para garantir a consistência das apreciações individuais. A inclusão de fotografias na base de dados pode auxiliar, permitindo a revisão ou apreciação conjunta de alguns atributos medidos de forma qualitativa, tais como padrão construtivo e estado de conservação, mas outros, tal como a micro-localização, são de mais difícil consideração[56].

Os efeitos dos atributos podem ser verificados através dos seus preços implícitos (preços hedônicos), calculados através da derivada parcial da função hedônica. Esta tarefa é mais simples quando as funções são logarítmicas ou lineares. No caso, a maioria das funções computadas conta com o expoente determinado para a variável dependente por Box-Cox, exceto para o caso dos modelos aditivos generalizados, que têm expoentes também para as variáveis independentes. Por exemplo, empregando o modelo de Mercado obtido com os

[56] Neste caso, a alternativa seria extrair medidas de localização a partir dos dados.

modelos aditivos generalizados (Equação 52), o preço hedônico para a área total dos apartamentos (Artocons) é o apresentado na Equação 69[57].

$$PH_{Artocons}=\partial Valor/\partial Artocons=48{,}221*Artocons^{0{,}0073}*(492{,}800+36{,}133*Artocons^{1{,}0073}+873{,}769*Razão^{1{,}0174}+518{,}277*Med+1.586{,}775*Fin+4.190{,}182*Lux+23{,}669*Bairro^{0{,}9861}-32{,}541*Idade^{0{,}9886}-43{,}120*Centro^{1{,}0495}-90{,}847*Comércio^{1{,}0332}+1.640{,}422*Vista-19{,}462*Mês^{1{,}006})^{0{,}9746/3} \quad \text{(Equação 69)}$$

Esta relação pode ser simplificada, utilizando novamente a função original (Equação 52), obtendo-se então a Equação 70.

$$PH_{Artocons}=\partial Valor/\partial Artocons=48{,}221*Artocons^{0{,}0073}*(Valor^{3/3{,}9746})^{0{,}9746/3} \quad \text{(Equação 70)}$$
$$=48{,}221*Artocons^{0{,}0073}*Valor^{0{,}2452}$$

Assim, o efeito da variação da área do imóvel no valor estimado depende da magnitude do próprio valor, o qual depende das características do imóvel. Por exemplo, considerando os valores médios destas duas variáveis (ver Tabela 17), o efeito calculado é de: $PH_{Artocons}=48{,}221*90{,}74^{0{,}0073}*74.310{,}64^{0{,}2452}=779{,}67$. Ou seja, mantendo constantes os demais atributos, um imóvel terá uma variação marginal do valor de mercado de aproximadamente 780 Reais para uma variação de um metro quadrado na área total. Levando em conta o Valor médio, a variação de uma unidade de área provoca uma variação de 1,05% no valor estimado, neste trecho da função.

Este tipo de informação pode ser utilizado na tomada de decisões durante o projeto de uma nova edificação, por exemplo. Para as outras funções e atributos, os preços hedônicos podem ser obtidos e analisados de forma semelhante. No caso de atributos binários, porém, há algumas diferenças na derivação, conforme alertam Halvorsen e Palmquist (1980). Para os

[57] Ver Rosen (1974) e Sheppard (1999). Na derivada parcial da função hedônica escolhida em relação ao atributo Artocons, os demais atributos comportam-se como constantes, ou seja, é como se a função fosse $Valor=(k+36{,}133*Artocons^{1{,}0073})^{3{,}9746/3}$, onde k é uma constante.

modelos baseados em regressão, existem outros elementos de comparação, tal como o exame dos coeficientes *t* calculados para as variáveis (Hair *et al.*, 1998; Neter *et al.*, 1990).

7.4.3 As técnicas alternativas como incremento de segurança das estimativas

As técnicas examinadas podem ser utilizadas como um veículo para o teste das condições da regressão. Geralmente os pressupostos são verificados com base em testes padronizados ou pelo exame de gráficos, tarefa que contém um componente subjetivo. A utilização de ferramentas alternativas pode fornecer testes mais robustos empiricamente de cada um dos problemas que os modelos de regressão enfrentam no mercado imobiliário.

Especificamente, as superfícies matemáticas podem ser utilizadas para verificar as medidas de localização incluídas, cuja imprecisão pode levar à ocorrência de autocorrelação espacial. As redes neurais e os modelos aditivos generalizados podem identificar relacionamentos não lineares e os modelos baseados em regras difusas são úteis para verificar a influência das transições entre segmentos de mercado, com vinculação ao exame da homocedasticidade dos resíduos. Adicionalmente, os modelos aditivos consistem em uma alternativa ao procedimento de Box-Cox, o qual pode introduzir tendências nos parâmetros dos modelos (Anglin e Gençay, 1996; Blackley *et al.*, 1984). Estas técnicas alternativas podem ser aplicadas em amostras-piloto e, havendo resultados similares com estas técnicas e com a regressão, a hipótese de violações sérias nos dados seria rejeitada.

Outra forma de aumento da segurança das estimativas é através da utilização de composições das estimativas de cada técnica, em um sistema multi-agente. Os modelos gerados por técnicas distintas podem apresentar deficiências locais, para um certo tipo de imóvel ou região da cidade, e a utilização de médias, ponderadas pelo nível geral de erro, pode tornar as estimativas mais robustas (Wood, 1976).

Por fim, os modelos desenvolvidos podem ser vistos como funções de preços hedônicos e, alternativamente, também podem ser interpretados como equações de regressão, linear ou não-linear, obtidos por diferentes técnicas de cálculo. Assim, as técnicas aplicadas são alternativas ao MMQ, no caso da regressão linear, e as redes neurais e os modelos aditivos generalizados são alternativas às técnicas de ajustamento dos modelos de regressão não linear.

É necessário desenvolver pesquisas adicionais mas, em princípio, se os dados e resultados dos modelos atenderem às condições da análise de regressão, os testes de hipóteses sobre o modelo e sobre os coeficientes poderão ser utilizados também nestas outras técnicas.

8 CONCLUSÃO

8.1 CONSIDERAÇÕES INICIAIS

Como exposto inicialmente, o mercado imobiliário tem características especiais que afetam as estimativas obtidas através do formato tradicional, que emprega estatística inferencial. Erros de até 15% do valor estimado têm sido indicados como aceitáveis pela literatura. Face à importância das avaliações e do mercado imobiliário, em termos econômicos e sociais, é importante buscar alternativas de aperfeiçoamento das avaliações.

Em parte, os problemas são provocados por erros nos dados observados do mercado. A utilização direta destes dados, sem uma etapa de preparação, pode invalidar os modelos estatísticos, em função da ruptura de pressupostos da análise de regressão. Neste aspecto, as técnicas da área de descobrimento de conhecimento em bases de dados (DCBD) podem colaborar, fornecendo um processo mais consistente de análise dos dados.

Em relação aos modelos, existem algumas tentativas de aperfeiçoamento, tais como a utilização de regressão não-linear e estatística espacial. Por outro lado, diversas técnicas da área de inteligência artificial apresentam potencial para utilização em avaliações. Tendo em vista o desenvolvimento recente, algumas ainda não foram completamente exploradas, e outras técnicas apresentaram resultados divergentes na literatura da área. Estas alternativas representam uma mudança expressiva de paradigma de análise. O processo atual de avaliação de imóveis pode ser classificado como um ramo da econometria, a qual é fortemente embasada na estatística inferencial, que, por sua vez, fundamenta a análise no teste de hipóteses ou de modelos previamente sugeridos pelo analista, através de amostras geralmente pequenas e de alta qualidade. As técnicas e métodos das áreas de descobrimento de conhecimento e inteligência artificial utilizam uma visão bastante diferente, pois o

descobrimento de conhecimento parte da análise dos dados para então obter *insights* sobre os relacionamentos existentes, e as técnicas de inteligência artificial são não-paramétricas, geralmente apenas com condições formais sobre os dados. Este contexto recomendava a exploração empírica, de forma ampla, verificando o desempenho de cada uma das técnicas. Neste trabalho, algumas alternativas foram investigadas, e os resultados apontam para um equilíbrio de desempenho, com modelos e níveis de erro semelhantes.

Por outro lado, existe uma parcela expressiva de subjetividade no processo atual de avaliação de imóveis, com influência significativa nos processos de seleção de casos e de atributos para as avaliações, dificultando a comparação dos trabalhos produzidos por diferentes avaliadores e também com possível impacto sobre a precisão. Neste caso, o desenvolvimento de um formato de análise sistemática dos dados, mais objetivo, também contribui para aumentar a precisão.

Ademais, podem ser identificados três tipos distintos de modelos do mercado imobiliário, com utilizações especiais para análise geral, avaliação em massa e avaliação individual, e com algumas diferenças na seleção dos dados e no desenvolvimento dos modelos, mas com dois requisitos fundamentais comuns, que são a exigência de explicabilidade dos modelos e de um nível razoável de precisão.

8.2 PROPOSTA INVESTIGADA E RESULTADOS ATINGIDOS

Procurando contribuir para o desenvolvimento do setor, foi apresentada a proposta de uma nova abordagem para as avaliações, considerando as características do mercado imobiliário e os requisitos dos tipos de avaliação, e explorando as técnicas da área de descobrimento de conhecimento e de Inteligência Artificial, através de um sistema de avaliações, consistindo de uma base de dados geral, preparada antecipadamente, e de técnicas alternativas para geração dos modelos preditivos.

A utilização de uma base única aumenta a robustez das estimativas, pois a preparação dos dados conduz à formação de uma base de dados mais consistente para a geração de modelos, fornecendo um contexto geral de mercado para a seleção objetiva de casos e atributos. A utilização de mais de uma técnica de modelagem e a separação de uma parte dos dados para

testar os modelos obtidos ampliam a confiança sobre os resultados.

Foi desenvolvida uma aplicação do sistema proposto, utilizando dados das declarações do Imposto sobre a Transmissão de Bens Imóveis (ITBI) do Município de Porto Alegre, com mais de 30 mil casos de transações de apartamentos residenciais. Estes dados foram preparados de acordo com a sistemática proposta, sendo também coletadas diversas novas variáveis. A evolução dos resultados obtidos com a análise de regressão, durante a preparação dos dados, é uma indicação da importância desta etapa.

A seleção de atributos foi realizada utilizando as abordagens de filtro e envoltório, com um conjunto de técnicas contribuindo para a seleção final, evitando as tendências que poderiam ser geradas pelo uso de uma única técnica. A seleção de casos foi realizada em duas etapas. Na primeira parte foram identificados casos irrelevantes, considerados como *outliers*, os quais prejudicam a análise[58]. Esta seleção foi realizada através das indicações de três ferramentas de modelagem, utilizando a regressão tradicional, regressão com superfícies matemáticas e redes neurais, também com o objetivo de eliminar tendências provocadas pelas técnicas. A seleção de casos foi realizada em paralelo com a seleção de atributos, construindo progressivamente modelos mais aprimorados, com maior potencial para identificar os *outliers*.

Considerando os tipos de avaliações, foi proposto um esquema de seleção dos casos mais relevantes em cada situação. Os modelos gerais (modelos de mercado) foram desenvolvidos com toda a base de dados, servindo também como guia para os demais formatos. Os modelos destinados à avaliação de massa foram desenvolvidos com base em parcelas da base, representando sub-mercados, identificados com a técnica de análise de agrupamento (clusterização). Por fim, a seleção de casos para os modelos de avaliação individual foi realizada através de um mecanismo de vizinhança própria baseado em medidas de posicionamento espacial e de diferenciação entre os imóveis da base e o caso em análise (avaliando), utilizando uma formulação especialmente desenvolvida. Os dados foram divididos em três grupos para a avaliação de massa, com apartamentos Pequenos, Médios e Grandes, e foi identificada uma amostra para avaliação individual. Em todos os modelos os dados foram divididos na proporção de 1/5, reservando-se 20% para testar os resultados.

[58] Para o tipo de análise desenvolvido. Para outras tarefas de descobrimento de conhecimento, estes casos podem ser de grande interesse.

Em seguida foi realizada a investigação das técnicas de modelagem, comparando o desempenho para os três formatos de avaliação. Foram desenvolvidos modelos utilizando as técnicas disponíveis em cinco propostas distintas, incluindo a regressão tradicional, regressão com superfícies matemáticas, redes neurais com explicação por lógica difusa, modelos aditivos generalizados e sistemas de regras difusas, ambos ajustados com algoritmos genéticos.

As principais inovações introduzidas neste trabalho são a seqüência de tratamento dos dados e os sistemas híbridos para desenvolvimento dos modelos de avaliação. A seleção de casos e atributos utilizou um processo baseado em técnicas de descobrimento de conhecimento em bases de dados e modelos gerais (modelos de mercado), propostos como alternativa ao processo tradicional de análise com base em pequenas amostras. Em relação às técnicas, os sistemas baseados em regras difusas evolucionárias e a extração de regras a partir de redes neurais foram introduzidos para o domínio do mercado imobiliário, sendo que os primeiros obtiveram bons resultados, atingindo desempenho superior à regressão em alguns modelos.

Em geral, os resultados foram geralmente similares, indicando que todas as técnicas apresentadas podem ser utilizadas em modelos de avaliação individual ou coletiva. Verificou-se que a sensibilidade aos casos espúrios (potenciais *outliers*) é semelhante, pois as técnicas têm aproximadamente o mesmo número de elementos com erros grandes, em todos os formatos. A utilização de mais de uma técnica permite uma forma empírica segura de verificação dos efeitos dos dados sobre a regressão. Em particular, as superfícies de resposta são interessantes para examinar os efeitos de localização e as redes neurais para identificar relacionamentos não-lineares. Com os resultados semelhantes, pode-se concluir que não há rupturas significativas dos pressupostos.

Quanto ao tamanho do conjunto de dados utilizado, verificou-se que os conjuntos menores conduziram a resultados melhores, o que está vinculado à variabilidade do mercado imobiliário. Uma indicação importante é que as técnicas de inteligência artificial não tiveram o desempenho degradado com a diminuição da amostra, nem a regressão foi prejudicada com o aumento do tamanho da amostra.

Os erros encontrados foram elevados para os modelos coletivos, atingindo coeficientes de dispersão (COD) de 18%, em média, embora com menos de 5% de casos com erros absolutos

superiores a 50% e com correlação de 0,98 entre os valores observados e os estimados pelos modelos. Estes resultados são parcialmente explicados pelo tipo de informação utilizada (estimativas fiscais), mas revelam que as técnicas exploradas têm sensibilidade semelhante aos dados, mesmo aquelas classificadas como não-paramétricas.

Por fim, os resultados obtidos com os modelos de avaliação individual também foram similares, para todas as técnicas. Os erros foram muito inferiores aos obtidos com os modelos coletivos, atingindo cerca de 7,4% no coeficiente de dispersão e com nenhum erro absoluto acima de 50%.

8.3 SUGESTÕES PARA PESQUISAS FUTURAS

Os resultados obtidos no âmbito deste trabalho devem ser interpretados restritivamente, sendo considerados em função do tipo de dados utilizados. Há evidências, na literatura, de que as técnicas mais adequadas podem ser diferentes para diferentes tarefas ou para utilização com diferentes tipos de dados. Assim, para estender a utilização das técnicas e da abordagem propostas, devem ser testadas outras fontes de dados e outros tipos de imóveis, tais como terrenos, residências, imóveis comerciais e imóveis rurais, verificando as eventuais influências destes tipos de imóveis na preparação dos dados, nos resultados e na escolha das técnicas de modelagem. Também existem outras técnicas disponíveis, tais como os sistemas de informações geográficas, que podem ser explorados.

Há algumas questões relevantes nos domínios do mercado imobiliário, que podem ser sugeridas para pesquisa futura, tais como as seguintes.

Em muitas situações, o avaliador enfrenta o problema da falta de dados, em função do desaquecimento ou da especificidade do segmento analisado, o que pode dificultar a aplicação de algumas das técnicas apresentadas, inclusive a análise de regressão. A literatura indica que existe uma relativa vinculação entre os sub-mercados, por efeitos macroeconômicos ou em função de um desenvolvimento econômico urbano ou regional similares. Assim, pode ser possível desenvolver modelos com dados de vários segmentos ou mesmo de outros locais. Uma das áreas de pesquisa é descobrir os estes vínculos, identificando quais os locais e as informações relevantes para este tipo de análise e que tipo de conexão existe entre os

segmentos, em termos espaciais e temporais. Uma abordagem possível é o desenvolvimento de modelos simultâneos para os diversos segmentos, possivelmente através de redes neurais com múltiplas saídas ou árvores de regras difusas.

Mesmo com o esforço para tornar o processo de avaliação mais objetivo, sabe-se que há uma parcela subjetiva nas avaliações. A busca pelo conhecimento de especialistas, através da compilação das regras ou procedimentos adotados na seleção de casos, coleta de atributos ou identificação de modelos, pode contribuir para aprimorar o processo de avaliação.

Outra questão de pesquisa é a investigação dos efeitos de intervenções públicas ou privadas sobre o mercado imobiliário. Ações como a implantação de *shopping centers* ou a abertura de uma nova via expressa podem modificar significativamente os valores dos imóveis. O desenvolvimento de modelos gerais destes processos pode auxiliar os profissionais envolvidos no planejamento urbano ou no desenvolvimento de empreendimentos, aprimorando as previsões dos cenários de mercado futuros.

Os preços imobiliários contêm uma parcela aparentemente aleatória. A identificação de padrões nesta parcela pode ser investigada com a seqüência de análise proposta neste trabalho. Para tanto, podem ser coletados outros dados referentes às transações, tais como velocidade de venda e características dos compradores e vendedores (renda, nível de escolaridade, perfil familiar e preferências de consumo, entre outras), desenvolvendo uma aplicação de descobrimento de conhecimento na busca por padrões nas transações, tais como as características dos imóveis que são vendidos mais rapidamente, ou quais os atributos preferidos conforme o perfil do comprador. Outras aplicações incluem a identificação de regiões ou tipos de imóveis com maior rentabilidade, comparando a distribuição espacial dos valores das construções, para venda e aluguel, com a distribuição dos valores dos terrenos.

As técnicas de avaliação não consideram explicitamente o risco de erro. Uma abordagem com diversas ferramentas, como foi sugerido, pode tornar as estimativas mais robustas. Entretanto, nas estimativas para hipotecas, em função do longo prazo de pagamento, é importante inferir o comportamento futuro. Esta é uma tarefa significativamente diferente, e devem ser investigadas as técnicas e os dados adequados. As pesquisas sugeridas acima podem contribuir para compreender melhor o funcionamento do mercado, aperfeiçoando as predições.

Outra questão importante é a coleta e disponibilização de dados em escala comercial, tal como ocorre em outros países, viabilizando as análises de descobrimento de conhecimento, as quais exigem geralmente uma grande quantidade de dados. Pode ser desenvolvido um *data warehouse* para o mercado imobiliário, reunindo diversos tipos de informação. Uma das alternativas seria a união das bases de dados das prefeituras (cadastro e dados do ITBI, por exemplo), com as informações da Caixa Econômica Federal (que é o maior agente financiador). Dados oriundos do Censo ampliariam o escopo da análise, que provavelmente seria baseado em um sistema de informações geográficas *on-line*. A conexão com a Receita Federal poderá diminuir a evasão de receitas, tanto para o Imposto de Renda quanto para os tributos imobiliários, financiando o sistema. Estando disponível um sistema deste tipo, surge também a possibilidade de pesquisas explorando o potencial dos dados para a tomada de decisão de investimentos, para tributação e também sobre o impacto do incremento do nível de informação geral sobre o funcionamento do mercado imobiliário.

REFERÊNCIAS BIBLIOGRÁFICAS

AAMODT, A.; PLAZA, E. Case-based reasoning: Foundational issues, methodological variations, and system approaches. **AI Communications**, v.7, n.1, p.39-59, Mar. 1994.

AAMODT, A.; SANDTORV, H. A.; WINNEM, O. M. Combining case based reasoning and data mining - A way of revealing and reusing RAMS experience. In: European Safety and Reliability Conference (ESREL'98), 1998, Trondheim. **Proceedings...** Rotterdam: Balkena, 1998, p.1345-1351.

ABNT (ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS). **Avaliações de imóveis urbanos**: NBR-5676. Rio de Janeiro: ABNT, 1989.

ABNT (ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS). **Avaliação de bens** - Parte 1 - Procedimentos gerais: NBR-14653-1. Rio de Janeiro: ABNT, 2000.

ABNT (ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS). **Avaliação de custos unitários e preparo de orçamento de construção para incorporação de edifício em condomínio**: NBR-12721. Rio de Janeiro: ABNT, 1992.

ABNT (ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS). **Avaliação de custos unitários e preparo de orçamento de construção para incorporação de edifício em condomínio**: NB-140. Rio de Janeiro: ABNT, 1966

ABRAMO, P. **A dinâmica imobiliária**: elementos para o entendimento da espacialidade urbana. 1988. Dissertação (Mestrado em Planejamento Urbano) - Instituto de Pesquisa e Planejamento Urbano, Universidade Federal do Rio de Janeiro, Rio de Janeiro.

ABRAMO, P. A ordem urbana Walraso-Thüneniana e suas fissuras: O papel da interdependência nas escolhas de localização. **Cadernos IPPUR**, a.13, n.2, p.69-91, 1999.

ABRAMO, P. **Mercado e ordem urbana**: Do caos à teoria da localização residencial. Rio de Janeiro: Bertrand Brasil/FAPERJ, 2001a.

ABRAMO, P. (Org.). **Cidades em transformação**: Entre o plano e o mercado. Rio de Janeiro: Ed. Autor, 2001b.

ABUNAHMAN, S. A. **Curso básico de engenharia legal e de avaliações**. São Paulo: PINI, 1999

ACCETTA, G. J. Presenting convincing residential appraisals. **Appraisal Journal**, p.168-173, April, 1999.

ADRIAANS, P.; ZANTINGE, D. **Data mining**. Harlow: Addison-Wesley, 1996

AGRAWAL, R.; IMIELINSKI, T; SWAMI, A.. Mining association rules between sets of items in large databases. In: ACM SIGMOD Conference on Management of Data, 1993, Washington. **Proceedings...** Washington: ACM, 1993, p.207-216.

ALBRITTON, H. D. **Controversies in real estate property valuation**: A commentary. Chicago: American Institute of Real Estate Appraisers, 1982.

ALCALÁ, R.; CASILLAS, J.; CORDÓN, O.; HERRERA, F.; ZWIR, S. J. I. Techniques for learning and tuning fuzzy rule-based systems for linguistic modeling and their applications. In: LEONDES, C. T. (Ed.). **Knowledge-based systems**. Techniques and applications, v.3. Academic Press, 2000, p.889-941.

ALTHOFF, K.-D.; BARTSCH-SPÖRL, B. Decision support for case-based applications. **Wirtschaftsinformatik**, n.38, v.1, p.6-14, 1996.

AMARAL, F. C. **Data mining**: Técnicas e aplicações para o marketing direito. São Paulo: Berkeley, 2001.

ANGLIN, P. M.; GENÇAY, R. Semiparametric estimation of a hedonic price function. **Journal of Applied Econometrics**, v.11, n. 6, p.633-648, Nov./Dec. 1996.

ANSELIN, L. Gis research infraestructure for spatial analysis of real estate markets. **Journal of Housing Research**, v.9, n.1, p.113-134, 1998.

ARBATLI, A. D.; AKIN, H. L. Rule extraction from trained neural networks using genetic algorithms. **Nonlinear Analysis, Theory, Methods and Applications**, v.30, n.3, p.1639-1648, 1997.

AUBIN, J.-P. **Neural networks and qualitative physics**. A viability approach. Cambridge: Cambridge University, 1996.

BÄCK , T.; KURSAWE, F. Evolutionary algorithms for fuzzy logic: A brief overview. In: BOUCHON-MEUNIER, B.; YAGER, R. R.; ZADEH, L. A. (Eds.). **Fuzzy logic and soft computing**, v.4. Singapore: World Scientific, 1995.

BÄCK, T.; HOFFMEISTER, F.; SCHWEFEL, H-P. Applications of evolutionary algorithms. **Technical report SYS-2/92**, Department of Computer Science. Dortmund: University of Dortmund, 1992.

BAGNOLI, C.; SMITH, H. C. The theory of fuzzy logic and its application to real estate valuation. **Journal of Real Estate Research**, v.16, n.2, p.169-199, 1998.

BALARINE, O. F. O. **Determinação do impacto de fatores sócio-econômicos na formação do estoque habitacional em Porto Alegre**. Porto Alegre: EDIPUCRS, 1996.

BALCHIN, P. N.; KIEVE, J. L. **Urban land economics**. 3ed. London: MacMillan, 1986.

BALCHIN, P.; BULL, G. H.; KIEVE, J. L. **Urban land economics and public policy**. 5ed. Basingstoke: MacMillan, 1995.

BALL, M. J. Recent empirical work on the determinants of relative house prices. **Urban Studies**, v.10, p.213-233, 1973.

BARBOSA Filho, D. S. Avaliação de terras conflagradas pelas fraldas urbanas. In: I Congresso Brasileiro de Engenharia de Avaliações (COBREAP), 1974, São Paulo. **Anais...** São Paulo: IBAPE, 1974, p.173-195.

BARBOSA, E. P.; BIDURIN, C. P. Seleção de modelos de regressão para predição via validação cruzada: Uma aplicação em avaliação de imóveis. **Revista Brasileira de Estatística**, v.52, n.197/198, p.105-120, Jan./Dez. 1991.

BARNETT, V.; LEWIS, T. **Outliers in statistical data**. 2ed. New York: John Wiley, 1984.

BAROSSI Filho, M.; BRAGA, M. B. Metodologia da econometria. In: VASCONCELLOS, M. A. S.; ALVES, D. (Eds.) **Manual de econometria**: Nível intermediário. São Paulo: Atlas, 2000, cap.1.

BARTIK, T. J.; SMITH, V. K. Urban amenities and public policy. In: MILLS, E. S. (Ed). **Handbook of regional and urban economics**, v.2 (urban economics). Amsterdam, Elsevier, 1987, chap.31, p.1207-1254.

BELSLEY, D. A.; KUH, E.; WELSCH, R. E. **Regression diagnostics:** Identifying influential data and sources of collinearity. New York: John Wiley, 1980.

BELL, G.; GRAY, J. N. The revolution yet to happen. In: DENNING, P. J.; METCALFE, R. M. (Eds.). **Beyond calculation**. Berlin: Springer-Verlag, 1997, p.5-32.

BELL, R. **Real estate damages**: An analysis of detrimental conditions. Chicago: Appraisal Institute, 1999.

BENÍTEZ, J. M.; CASTRO, J. L.; REQUEÑA, I. Are artificial neural networks black boxes? **IEEE Transactions on Neural Networks**, v.8, n.5, p.1156-1164, Sept. 1997.

BENNETT, R. J.; HORDIJK, L. Regional econometric and dynamic models. In: NIJKAMP, P. (Ed). **Handbook of regional and urban economics**, v.1 (Regional economics). Amsterdam: Elsevier, 1986, chap.10, p.407-441.

BERRINI, L. C. **Avaliações de imóveis**. 3ed. São Paulo: Freitas Bastos, 1957.

BERRY, M. J. LINOFF, G. **Data mining techniques**: For marketing, sales, and customer support. New York: John Wiley, 1997.

BERRY, M. J.; LINOFF, G. **Mastering data mining**. New York: Wiley Computer Publishing, 2000.

BIAVA, A. H. R. **Contribuição de melhoria**: Necessidades de inovação fiscal. São Paulo: IPE, 1986.

BIBLE, D. S.; HSIEH, C-H. Applications of geographic information systems for the analysis of apartment rents. **Journal of Real Estate Research**, v.12, n.1, p.79-88, 1996.

BLACKLEY, P.; FOLLAIN Jr., J. R.; ONDRICH, J. Estimation of hedonic models: How serious is the iterative OLS variance bias? **Review of Economics and Statistics**, v.66, n.2, p.348-353, May, 1984.

BLEKAS, K.; STAFYLOPATIS, A. Real-coded genetic optimization of fuzzy clustering. In: 4th European Congress on Intelligent Techniques and Soft Computing (EUFIT'96), 1996, Aachen. **Proceedings...**,v.1. Aachen: Elite, 1996, p. 461-465.

BLUM, A. L.; LANGLEY, P. Selection of relevant features and examples in machine learning. **Artificial Intelligence**, n.97, p.245-271, 1997.

BOND, M. T.; SEILER, V. L.; SEILER, M. J. Residential real estate prices: A room with a view. **Journal of Real Estate Research**, v.23, n.1/2, p.129-137, 2002.

BONISSONE, P. P. Soft computing: The convergence of emerging reasoning technologies. **Soft Computing**, v.1, n.1, p.6-18, 1997.

BONISSONE, P. P.; CHEETAM, W.; GOLIBERSUCH, D. C.; KHEDKAR, P. Automated residential property valuation: An accurate and reliable approach based on soft computing. In: RIBEIRO, R.; ZIMMERMANN, H.; YAGER, R. R.; KACPRZYK, J. (Eds.). **Soft computing in financial engineering**. Heidelberg: Physica-Verlag, 1998.

BONISSONE, P. P.; CHEN, Y-T.; GOEBEL, K.; KHEDKAR, P. S. Hybrid soft computing systems: Industrial and commercial applications. **Proceedings of the IEEE**, v.87, n.9, p.1641-1667, Sept. 1999.

BONRIGHT, J. C. **The valuation of property**. v.1. New York: McGraw-Hill, 1937.

BORST, R. A. Artificial neural networks: The next modeling/calibration technology for the assessment community? **Property Tax Journal**, v.10, n.1, p.69-94, Mar. 1991.

BOURASSA, S. C.; HAMELIN, F.; HOESLI, M.; MACGREGOR, B. D. Defining housing submarkets. **Journal of Housing Economics**, v.8, n.2, p. 160-183, June, 1999.

BOURASSA, S. C.; HOESLI, M. The structure of housing submarkets in a metropolitan region. **Paper 99.15** - Ecole des Hautes Etudes Commerciales. Genève: Université de Genève, 1999.

BOYLE, M. A.; KIEL, K. A. A survey of house price hedonic studies of the impact of environmental externalities. **Journal of Real Estate Literature**, v.9, n.2, p.117-144, 2001.

BRACHMAN, R. J.; ANAND, T. The process of knowledge discovery in databases: A human-centered approach. In: FAYYAD *et al.*, 1996b. chap.2, p.37-57.

BRADLEY, P. S.; FAYYAD, U.; REINA, C. Scaling clustering algorithms to large databases. In: 4th International Conference on Knowledge Discovery and Data Mining (KDD-98), 1998, New York. **Proceedings...** Menlo Park (CA): AAAI Press, 1998.

BRAGA, A. de P.; LUDERMIR, T. B.; CARVALHO, A. C. P. L. F. **Redes neurais artificiais:** Teoria e aplicações. Rio de Janeiro: LTC, 2000.

BREMAEKER, F. E. J. de. Mitos sobre as finanças dos municípios brasileiros. **Revista de Administração Municipal**, v.41, n.212, p.6-21, Jul./Set., 1994.

BRUHA, I. Pre- and post-processing in machine learning and data mining. In: PALIOURAS, G.; KARKALETSIS, V.; SPYROPOULOS, C. D. (Eds.). **Machine learning and its applications**. Berlin: Springer-Verlag, 2001, p.258-266.

BUENO, T. C. D.; WANGENHEIM, C. Von; HOESCHL, H. C.; MATTOS, E.; BARCIA, R. M. Uso da teoria jurídica para recuperação em amplas bases de textos jurídicos. In: 19º Congresso Nacional da Sociedade Brasileira de Computação, 1999, Rio de Janeiro. **Anais...** Rio de Janeiro: EntreLugar, 1999**,** v.4, p.107-120.

BYRNE, P. Fuzzy analysis: A vague way of dealing with uncertainty in real estate analysis? **Journal of Property Valuation and Investment**, v.13, n.3, p.22-41, 1995.

CABENA, P.; HADJINIAN, P.; STADLER, R.; VERHEES, J.; ZANASI, A. **Discovering data mining**: From concept to implementation. Prentice-Hall PTR: Upper Saddle River (NJ), 1997.

CAIRES, H. R. R. de. **Novos tratamentos matemáticos em engenharia de avaliações**. São Paulo: Pini, 1978.

CAIRES, H.; CAIRES, H. R. R. de. **Avaliação de glebas urbanizáveis**. São Paulo: Pini, 1984.

CAMARGO, E. C. G.; MONTEIRO, A. M. V.; FELGUEIRAS, C. A.; FUKS, S. Integração de geoestatística e sistemas de informação geográfica: Uma necessidade. In: 5º Congresso e Feira para Usuários de Geoprocessamento da América Latina (GIS Brasil'99), 1999, Salvador. **Anais...** (Em Cd-Rom). Curitiba: FatorGIS, 1999.

CAN, A. Gis and spatial analysis of housing and mortgage markets. **Journal of Housing Research**, v.9, n.1, p.61-86, 1998.

CAN, A. The measurement of neighborhood dynamics in urban house prices. **Economic Geography**, v.66, n.3, p.254-272, July, 1990.

CAPLES, S. C.; HANNA, M. E.; PREMEAUX, S. R. Least squares versus least absolute value in real estate appraisals. **Appraisal Journal**, p.18-24, Jan. 1997.

CARNEGHI, C. Appraisal arbitration: The role of the real estate appraiser in resolving value disputes. **Appraisal Journal**, p.119-125, Apr. 1999.

CARVALHO, L. A. V. de. **Data mining**: A mineração de dados no marketing, medicina, economia, engenharia e administração. São Paulo: Érica, 2001.

CASE, K. E. Real estate and the macroeconomy. **Brookings Papers on Economic Activity**, n.2, p.119-162, 2000.

CASE, K. E.; SHILLER, R. J. A decade of boom and bust in the prices of single-family homes: Boston and Los Angeles, 1983 to 1993. **New England Economic Review**, p.40-52, Mar./Apr. 1994.

CASE, K. E.; SHILLER, R. J. Forecasting prices and excess returns in the housing marketing. **AREUEA Journal**, v.18, n. 3, p.153-273, 1990.

CASE, K. E.; SHILLER, R. J. Prices of single family homes since 1970: New indexes for four cities. **New England Economic Review**, p.45-56, Sept./Oct., 1987.

CASE, K. E.; SHILLER, R. J. The efficiency of the market for single-family homes. **American Economic Review**, v.79, n.1, p.125-137, Mar. 1989.

CASTRO, J. L.; MANTAS, C. J.; BENÍTEZ, J. M. Interpretation of artificial neural networks by means of fuzzy rules. **IEEE Transactions on Neural Networks**, v.13, n.1, p.101-116, Jan. 2002.

CECHIN, A. L. **The extraction of fuzzy rules from neural networks**. Aachen: Shaker Verlag, 1998.

CECHIN, A. L.; SOUTO, A.; GONZÁLEZ, M. A. S. Análise de imóveis através de redes neurais artificiais na cidade de Porto Alegre. **Scientia**, v.10, n.2, p.5-32, Jul./Dez. 1999.

CECHIN, A. L.; SOUTO, A.; GONZÁLEZ, M. A. S. Real estate value at Porto Alegre city using artificial neural networks. In: 6th Brazilian Symposium on Neural Networks (SBRN'2000), Nov. 2000, Rio de Janeiro. **Proceedings...** Rio de Janeiro: SBRN, 2000.

CENECORTA, A. X.; SMOLKA, M. O. (Orgs.). **Los pobres de la ciudad y la tierra**. Zinacantepec (México): El Colegio Mexiquense/Lincoln Institute of Land Policy, 2000.

CHANDIAS, M. E. **Tasación de inmuebles urbanos**. Buenos Aires: Alsina, 1954.

CHAO, L.-C.; SKIBNIEWISKI, M. J. Fuzzy logic for evaluating alternative construction technology. **Journal of Construction Engineering and Management**, v.124, n.4, p.297-304, July, 1998.

CHEMERIS, I. **A função social da propriedade**. O papel do Judiciário diante das invasões de terras. São Leopoldo: Ed.Unisinos, 2002.

CIOS, K. J.; PEDRYCZ , W.; SWINIARSKI, R. W. **Data mining**: Methods for knowledge discovery. Boston: Kluwer, 1998.

CLAYTON, J. Further evidence on real estate market efficiency. **Journal of Real Estate Research**, v.15, n.1/2, p.41-57, 1998.

CLIFF, A. D.; ORD, J. K. **Spatial autocorrelation**. London: Pion, 1973.

CONNELLAN, O.; JAMES, H. Estimate realisation price (ERP) by neural networks: forecasting commercial property values. **Journal of Property Valuation and Investment**, v.16, n.1, p.71-86, 1998.

CORDÓN, O.; HERRERA, F. A two-stage evolucionary process of designing TSK fuzzy rule-based systems**. IEEE Transactions on Systems, Man, and Cybernetics** - Part B: Cybernetics, v.29, n.6, 1999.

CORDÓN, O.; HERRERA, F.; HOFFMANN, F.; MAGDALENA, L. **Genetic fuzzy systems:** Evolutionary tuning and learning of fuzzy knowledge bases. World Scientific: Singapore, 2001.

CROSBY, N.; LAVERS, A.; MURDOCH, J. Property valuation variation and the 'margin of error' in the UK. **Journal of Property Research**, v.15, n. 4, 305-330, 1998a.

CROSBY, N.; LAVERS, A.; MURDOCH, J. Property valuations: The role of the margin of error test in establishing negligence. **University of Western Australia Law Review**, v.27, n.2, p.156-197, July, 1998b.

CUTHBERTSON, K.; HALL, S. G.; TAYLOR, M. P. **Applied econometric techniques**. New York: Harvester Wheatsheaf, 1992.

CZERNOWSKY, R. M. J. Expert systems in real estate valuation. **Journal of Valuation**, v.8, n.4, p.376-393, 1989.

DANIEL, C.; WOOD, F. S. **Fitting equations to data**. 2ed. New York: John Wiley, 1980.

DANTAS, R. A. **Engenharia de avaliações**. São Paulo: Pini, 1998.

DANTAS, R. A.; CORDEIRO, G. M. Evaluation of the Brazilian city of Recife's condominium market using generalized linear models. **Appraisal Journal**, p.247-257, July, 2001.

DE CESARE, C. M. de. **An empirical analysis of equity in property taxation:** A case study from Brazil. 1998. PhD Thesis - University of Salford, Salford (UK).

DEAN, T.; ALLEN, J. ALOIMONOS, Y. **Artificial intelligence**: Theory and practice. Menlo Park: Addison-Wesley, 1995.

DEDDIS, W. G.; McCLUSKEY, W. J.; MANNIS, A.; McBURNEY, D. **The price is right?** Using computer-based mass appraisal techniques to value residential property. London: Royal Institute of Chartered Surveyors, 1998.

DEOGUN, J. S.; RAGHAVAN, V. V.; SARKAR, A.; SEVER, H. Data mining: Research trends, challenges, and applications. In: LIN, T. Y.; CERCONE, N. (Eds.). **Roughs sets and data mining**: Analysis of imprecise data. Boston: Kluwer Academic Publishers, 1997, p. 9-45.

DERYCKE, P.-H. **La economia urbana**. Madrid: IEAL, 1971.

DETWEILER, J. H.; RADIGAN, R. E. Computer-assisted real estate appraisal: A tool for the practicing appraiser. **Appraisal Journal**, p.91-101, Jan. 1996

DEWEESE, G. S. The role of the professional appraiser in REIT valuations. **Appraisal Journal**, p.236-241, Jan. 1998.

DIAZ III, J. An investigation into the impact of previous expert value estimates on appraisal judgement. **Journal of Real Estate Research**, v.13, n.1, p.57-66, 1997.

DIAZ III, J. How appraisers do their work: A test of the appraisal process and the development of a descriptive model. **Journal of Real Estate Research**, v.5, n.1, p.1-16, 1990.

DIAZ III, J.; HANSZ, J. A. How valuers use the value opinions of others. **Journal of Property Valuation and Investment**, v.15, n.3, p.256-260, 1997.

DIAZ, M. D. M. Multicolinearidade. In: VASCONCELLOS, M. A. S.; ALVES, D. (Eds.) **Manual de econometria**: Nível intermediário. São Paulo: Atlas, 2000, cap.6.

DIN, A.; HOESLI, M.; BENDER, A. Environmental variables and real estate prices. **Urban Studies**, v.38, n.11, p.1989-2000, 2001.

DING, C.; SIMONS, R.; BAKU, E. The effect of residential investment on nearby property values: Evidence from Cleveland, Ohio. **Journal of Real Estate Research**, v.19, n.1/2, p.23-48, 2000.

DODGSON, J. S.; TOPHAM, N. Valuing residential properties with the hedonic method: A comparison with the results of professional valuations. **Housing Studies**, v.5, n.3, p.209-213, 1990.

DOWSE, G. Valuations at issue: Market value, what is it? In: 6th Annual Pacific-Rim Real Estate Society Conference, Jan. 2000, Sydney. **Proceedings...** Sydney: PRRES, 2000.

DUBIN, R. A. Estimation of regression coefficients in the presence of spatially autocorrelated error terms. **Review of Economics and Statistics**, v. 70, p.466-474, 1988.

DUBIN, R. A. Predicting house prices using multiple listings data. **Journal of Real Estate Finance and Economics**, v.17, n.1, p.35-59, July, 1998.

DUBIN, R. A. Spatial autocorrelation and neighbourhood quality. **Regional Science and Urban Economics**, v.22, n.3, p.433-452, Sept. 1992.

DUBIN, R. A.; SUNG, C.-H. Spatial variation in the price of housing: Rent gradients in non-monocentric cities. **Urban Studies**, v.24, p.193-204, June, 1987.

DUBOIS, D.; ESTEVA, F.; GARCIA, P.; GODO, L.; MÁNTARAS, R. L. de; PRADE, H. Fuzzy modeling of case-based reasoning and decision. In: 2nd International Conference on Case-Based Reasoning (ICCBR-97), 1997, Providence (USA). **Proceedings...** Berlin: Springer, 1997, p.599-610.

ECKERT, J. K.; O'CONNOR, P. M.; CHAMBERLAIN, C. Computer-assisted real estate appraisal: A California savings and loan case study. **Appraisal Journal**, p.524-532, Oct. 1993.

EICHENBAUM, J. Incorporating location into computer-assisted valuation. **Property Tax Journal**, v.8, n.2, p.151-169, 1989.

ENGLE, R. F.; GRANGER, C. W. J. Cointegration and error correction: representation, estimation and testing. **Econometrica**, v.55, n.2, p.251-276, Mar. 1987.

ESTER, M.; FROMMELT, A.; KRIEGEL, H-P. SANDER, J. Spatial data mining: Database primitives, algorithms and efficient DBMS support. **Data Mining and Knowledge Discovery**, v.4, n.2/3, p.193-216, July, 2000.

ESTER, M.; KRIEGEL, H-P. SANDER, J.; XU, X. Clustering for mining in large spatial databases. **Künstliche Intelligenz**, n.12, v.1, p.18-24, 1998.

EVANS, A. W. The property market: Ninety per cent efficient? **Urban Studies**, v.32, n.1, p.5-29, 1995.

EVANS, A.; JAMES, H.; COLLINS, A. Artificial neural networks: An application to residential valuation in UK. **Journal of Property Tax Assessment & Administration**, v.1, n.3, p.78-92, May, 1995.

FANNING, S. F.; GRISSON, T. V.; PEARSON, T. D. **Market analysis for valuation appraisals**. Chicago: Appraisal Institute, 1994.

FAYYAD, U. M.; PIATETKSY-SHAPIRO, G.; SMITH, P. From data mining to knowledge discovery: An overview. In: FAYYAD *et al*., 1996b, chap.1, p.1-34, 1996a.

FAYYAD, U. M.; PIATETSKY-SHAPIRO, G.; SMYTH, P.; UTHURUSAMY, R. (Eds.). **Advances in knowledge discovery and data mining**. Menlo Park (CA)/Cambridge (MA): AAAI Press/MIT, 1996b.

FERNANDES, J. F. Conceitos gerais. Métodos avaliatórios. In: IBAPE, 1983, p.19-21.

FERNANDEZ, G. R. Novos instrumentos urbanos: Mecanismos de reparcelamento e compensação fundiário-imobiliária. In: ABRAMO, 2001b, p.191-217.

FIKER, J. **Avaliação de terrenos e imóveis urbanos**. São Paulo: Pini, 1993.

FRANCHI, C. C. **Avaliação das características que contribuem para a formação do valor de apartamentos na cidade de Porto Alegre**. 1991. Dissertação (Mestrado em Engenharia) – Curso de Pós-Graduação em Engenharia Civil, Universidade Federal do Rio Grande do Sul, Porto Alegre.

FRANK Jr., C. R. **Statistics and econometrics**. New York: Holt, Rinehart and Winston, 1971.

FURTADO, F. Instrumentos para a recurperação de mais-valias na América Latina: Debilidade na implementação, ambigüidade na interpretação. **Cadernos IPPUR**, a.11, n.1/2, p.163-205, 1997.

FURTADO, F. **Urbanização de terras e ocupação do solo urbano: Elementos para a análise do processo de crescimento das cidades brasileiras**. 1993. Dissertação (Mestrado em Planejamento Urbano) - Instituto de Pesquisa e Planejamento Urbano, Universidade Federal do Rio de Janeiro, Rio de Janeiro.

GALLIMORE, P.; FLETCHER, M.; CARTER, M. Modeling the influence of location on value. **Journal of Property Valuation and Investment**, v.14, n.1, p.6-19, 1996.

GANTI, V.; GEHRKE, J.; RAMAKRISHNAN, R. Mining very large databases. **IEEE Computer**, v.32, n.8, p.38-45, Aug. 1999.

GAU, G. W. Efficient real estate markets: Paradox or paradigm? **AREUEA Journal**, v.15, n.2, p.1-12, 1987.

GEHRKE, J.; RAMAKRISHNAN, R.; GANTI, V. RainForest – A framework for fast decision tree construction of large datasets. In: 24th Very Large Data Base Conference (VLDB'98), 1998, San Francisco. **Proceedings...**. San Francisco: Morgan Kaufmann, 1998, p.416-427.

GELBTUCH, H. C.; MACKMIN, D.; MILGRIM, M. **Real estate valuation in global markets**. Chicago: Appraisal Institute, 1997.

GEORGE, H. **Progreso y miseria.** New York: Robert Schalkenbach Foundation, 1972.

GILBERTSON, B. G. Appraisal or valuation: An art or a science? **Real Estate Issues**, v.26, n.3, p.86-89, Fall, 2001.

GODDARD, B. L. The role of graphic analysis in appraisals. **Appraisal Journal**, p.429-435, Oct. 1999.

GOLDBERG, D. E. **Genetic algorithms in search, optimization, and machine learning**. Reading: Addison-Wesley, 1989.

GOLDBERG, D. E. Real-coded genetic algorithms, virtual alpha- bets, and blocking. **Complex Systems**, v.5, p.139-167, 1991.

GÓMEZ-SKARMETA, A. F.; JIMÉNEZ, F. Generating and tuning fuzzy rules using hybrid systems. In: 6th IEEE International Conference on Fuzzy Systems (FUZZ/IEEE'97), 1997, Barcelona. **Proceedings...** Piscataway (NJ): IEEE, 1997, p. 247-252.

GONÇALVES, M. F. dos R. O imposto predial e territorial urbano e a progressividade. **Revista de Administração Municipal**, v.35, n.189, p.6-14, Out./Dez. 1988.

GONDIM, L. M. P. O plano diretor como instrumento de um pacto social urbano: Quem põe o guizo no gato? **Ensaios FEE**, v.16, n.2, p.472-490, 1995.

GONZÁLEZ, A. J.; LAUREANO-ORTIZ, R. A case-based reasoning approach to real estate property appraisal. **Expert System With Applications**, v.4, p.229-246, 1992.

GONZÁLEZ, M. A. S.; FORMOSO, C. T. Análise da utilização de dados do imposto de transmissão para atualização das plantas de valores. In: 1º Congresso Brasileiro de Avaliações para Fins Tributários (CONBRAFT), 1995, Cachoeira do Sul (RS). **Anais...** Porto Alegre: IGEL, 1995a, p.88-99.

GONZÁLEZ, M. A. S.; FORMOSO, C. T. Integração dos tributos imobiliários urbanos: Uma planta de valores inferencial única, determinada com dados de ITBI. In: 8º Congresso Brasileiro de Engenharia de Avaliações e Perícias (COBREAP), 1995, Florianópolis. **Anais...** Florianópolis: IBAPE, 1995b, p.326-335.

GONZÁLEZ, M. A. S. **A Engenharia de avaliações na visão inferencial**. São Leopoldo: Ed. Unisinos, 1997.

GONZÁLEZ, M. A. S. **A formação do valor de alugueis de apartamentos residenciais na cidade de Porto Alegre**. 1993. Dissertação (Mestrado em Engenharia) – Curso de Pós-Graduação em Engenharia Civil, Universidade Federal do Rio Grande do Sul, Porto Alegre.

GONZÁLEZ, M. A. S. Avaliação de aluguéis residenciais com ponderação pela distância ao imóvel-objeto. In: 8º Congresso Brasileiro de Engenharia de Avaliações e Perícias (COBREAP), 1995, Florianópolis. **Anais...** Florianópolis: IBAPE, Nov.1995a, p.17-22.

GONZÁLEZ, M. A. S. Montagem de base de dados para cálculo automático de coordenadas. **Caderno Brasileiro de Avaliações e Perícias**, v.7, n.80, p.292-297, Fev. 1996.

GONZÁLEZ, M. A. S. Plantas de valores inferenciais: A espacialidade considerada através de trend surfaces. In: 8º Congresso Brasileiro de Engenharia de Avaliações e Perícias (COBREAP), 1995, Florianópolis. **Anais...** Florianópolis: IBAPE, Nov.1995b,p.390-397.

GONZÁLEZ, M. A. S.; ERBA, D. A. Cálculo em massa de valores imobiliários para tributação sobre a propriedade utilizando inferência e modelos digitais de valores (MDV). In: 9º Congresso Brasileiro de Engenharia de Avaliações e Perícias (COBREAP), 1997, São Paulo. **Anais...** São Paulo: IBAPE, 1997, v.1, p.315-322.

GONZÁLEZ, M. A. S.; SOIBELMAN, L.; FORMOSO, C. T. A new approach to spatial analysis in CAMA. In: 9th European Real Estate Society Conference (ERES 2002), June, 2002, Glasgow. **Proceedings...** Glasgow: ERES, 2002a.

GONZÁLEZ, M. A. S.; SOIBELMAN, L.; FORMOSO, C. T. Explaining results in a mass-appraisal model. In: 9th European Real Estate Society Conference (ERES 2002), June, 2002, Glasgow. **Proceedings...** Glasgow: ERES, 2002b.

GOODALL, B. **The economics of urban areas**. Oxford: Pergamon, 1972.

GRAFF, R. A.; YOUNG, M. S. The magnitude of random appraisal error in commercial real estate valuation. **Journal of Real Estate Research**, v.17, n.1, p.33-54, 1999.

GREFENSTETTE, J. Optimization of control parameters for genetic algorithms. **IEEE Transactions on Systems, Man, and Cybernetics**, v.16, n.1, p.122-128, 1986.

GRILICHES, Z. **Price indexes and quality change**. Cambridge: Harvard University Press, 1971.

GUEDES, J. C. O emprego de inteligência artificial na avaliação de bens. In: 8º Congresso Brasileiro de Avaliações e Perícias (COBREAP), 1995, Florianópolis. **Anais...** Florianópolis: IBAPE, 1995, p.368-374.

GUJARATI, D. N. **Econometria básica**. São Paulo: Makron, 2000.

HAIR Jr., J. F.; ANDERSON, R. E.; TATHAM, R. L.; BLACK, W. C. **Multivariate data analysis.** 5ed. Upper Saddle River (NJ): Prentice-Hall, 1998.

HALVORSEN, R.; PALMQUIST, R. The interpretation of dummy variables in semilogarithmic equations. **American Economic Review**, v.70, n.3, p.474-5, June, 1980.

HALL, M. A. Correlation-based feature selection for discrete and numeric class machine learning. In: 17th International Conference on Machine Learning (ICML 2000), 2000, Stanford. **Proceedings...** San Francisco: Morgan Kaufmann, 2000**,** p.359-366.

HALLINAN, J.; JACKWAY, P. Simultaneous evolution of feature subset and neural classifier on high-dimensional data. In: Conference on Digital Image Computing: Techniques and Applications (DICTA'99), 1999, Perth (Austrália). **Proceedings...** Piscataway (NJ): IEEE, 1999.

HAN, J.; KAMBER, M. **Data mining**: Concepts and techniques. San Francisco: Morgan Kaufmann, 2001.

HAND, D.; MANNILA, H.; SMYTH, P. **Principles of data mining**. Cambridge: MIT, 2001.

HANSZ, J. A.; DIAZ III, J. Valuation bias in commercial appraisal: A transaction price feedback experiment. **Real Estate Economics**, v.29, n.4, p.553-565, Winter, 2001.

HARMANN, H. H. **Modern factor analysis**. 3ed. Chicago: University of Chicago, 1976.

HARVEY, J. **Urban land economics**. 4ed. London: MacMillan, 1996.

HASTIE, T.; TIBSHIRANI, R. Generalized additive models. In: **Encyclopedia of Statistical Sciences**, v.2. New York: Wiley Computer Publishing, 1997.

HASTIE, T.; TIBSHIRANI, R. Generalized additive models. **Statistical Science**, v.1, p.297-318, 1986.

HÄTÖNEN, K.; KLEMETTINEN, M.; MANNILA, H.; RONKAINEN, P.; TOIVONEN, H. Knowledge discovery from telecommunication network alarm databases. In: 12th International Conference on Data Engineering (ICDE 1996), 1996, New Orleans. **Proceedings...** Piscataway (NJ): IEEE, 1996, p.115-122.

HAYKIN, S. **Redes neurais** - Fundamentos. Porto Alegre: Artmed, 2001.

HAYTER, A. J. **Probability and statistics**: For engineers and scientists. Boston: PWS, 1996.

HENDERSON, J. M.; QUANDT, R. E. **Teoria microeconômica**: Uma abordagem matemática. São Paulo: Pioneira, 1976.

HENDRY, D. F. Encompassing. **National Institute Economic Review**, p.88-92, Aug. 1988.

HENDRY, D. F.; MIZON, G. E. Serial correlation as a convenient simplification, not a nuisance: A comment on a study of the demand for money by the Bank of England. **Economic Journal**, v.88, p.549-563, Sept. 1978.

HERRERA, F.; LOZANO, M.; VERDEGAY, J. L. Generating fuzzy rules from examples using genetic algorithms. In: BOUCHON-MEUNIER, B.; YAGER, R. R.; ZADEH, L. A. (Eds.), **Fuzzy logic and soft computing**. Singapore: World Scientific, 1995, p.11-20.

HOWES, P.; CROOK, N. Using input parameter influences to support the decisions of feedforward neural networks. **Neurocomputing**, n.24, p.191-206, 1999.

HSIA, M.; BYRNE, P. Computer assisted valuation: The development of an automated prototype. **Journal of Valuation**, v.7, n.4, p.395-407, 1989.

HUANG, S. H.; XING, H. Extract intelligible and concise fuzzy rules from neural networks. **Fuzzy Sets and Systems**, v.132, n.2, p.233-243, Dec. 2002.

IAAO (INTERNATIONAL ASSOCIATION OF ASSESSING OFFICERS). **Standards on ratio studies**. Chicago: IAAO, 1990.

IBAPE (INSTITUTO BRASILEIRO DE AVALIAÇÕES E PERÍCIAS DE ENGENHARIA). **Avaliações para garantias**. São Paulo: Pini, 1983.

IBAPE (INSTITUTO BRASILEIRO DE AVALIAÇÕES E PERÍCIAS DE ENGENHARIA). **Engenharia de avaliações**. São Paulo: Pini, 1974.

IHERING, R.Von. **Teoria simplificada da posse**. Salvador: Progresso, 1957.

ILLICH, N.; SIMONOVIC, S. Evolutionary algorithm for minimization of pumping cost. **Journal of Computing in Civil Engineering**, v.12, n.4, p.232-240, Oct. 1998.

ISAKSON, H. R. The nearest neighbors appraisal technique: An alternative to the adjustment grid methods. **AREUEA Journal**, v.14, n.2, p.274-286, 1986.

ISAKSON, H. R. The review of real estate appraisals using multiple regression analysis. **Journal of Real Estate Research**, v.15, n.1/2, p.177-190, 1998.

ISAKSON, H. R. Using multiple regression analysis in real estate appraisal. **Appraisal Journal**, v.69, n.4, p.424-430, Oct. 2001.

ISHIKAWA, M. Rule extraction by sucessive regularization. **Neural Networks**, v.13, p.1171-1183, 2000.

JENSEN, D. L. Artificial intelligence in computer-assisted mass appraisal. **Property Tax Journal**, v.9, n.1, p.5-24, Mar. 1990.

JENSEN, D. L. Alternative modeling techniques in computer-assisted mass appraisal. **Property Tax Journal**, v.6, n.3, p.193-237, Sept. 1987.

JOHN, G. H.; KOHAVI, R.; PFLEGER, K. Irrelevant features and the subset selection problem. In: 11th International Conference on Machine Learning, 1994, San Francisco. **Proceedings...** San Francisco: Morgan Kaufmann, 1994, p.121-129.

JUD, G. D.; WATTS, J. M. Schools and housing values. **Land Economics**, v.57, n.3, p. 459-470, Aug. 1981.

JUDGE, G. G.; HILL, R. C.; GRIFFITHS, W.; LÜTKEPOHL, H.; LEE, T.-C. **Introduction to the theory and practice of econometrics**. 2ed. New York: John Wiley, 1985.

KACPRZYK, J. **Multistage fuzzy control**. A model-based approach to fuzzy control and decision making. Chichester: John Wiley, 1997.

KAIN, J. F.; QUIGLEY, J. M. Measuring the value of housing quality. **Journal of the American Statistical Association**, v.65, n.330, p.532-548, June, 1970.

KANDAL, A.; FRIEDMAN, M. Defuzzification using most typical values. **IEEE Transactions on Systemas, Man, and Cybernetics - B**, v.28, n.6, p.901-906, Dec. 1998.

KANG, H. B.; REICHERT, A. K. An evaluation of alternative estimation techniques and functional forms in developing statistical appraisal models. **Journal of Real Estate Research**, v.2, n.1, p.1-29, 1987.

KANUNGO, T.; MOUNT, D. M.; NETANYAHU, N. S.; PIATKO, C.; SILVERMAN, R.; WU, A. Y. Computing nearest neighbors for moving points and applications to clustering. In: 10th ACM/SIAM Symposium on Discrete Algorithms (SODA 1999), 1999, Baltimore. **Proceedings...** Washington: ACM/SIAM, 1999, p.931-932.

KARRAY, F.; ZANELDIN, E.; HEGAZY, Y.; SHABEEB, A. H. M.; ELBELTAGY, E. Tools of soft computing as applied to the problem of facilities layout planning. **IEEE Transactions on Fuzzy Systems**, v.8, n.4, p.367-379, Aug. 2000.

KASKI, S. Data exploration using self-organizing maps. **Acta Polytechnica Scandinavica**, n.82. Espoo: Finish Academy of Technology, 1997.

KASKI, S.; KANGAS, J.; KOHONEN, T. Bibliography of self-organizing map (SOM) papers: 1981-1997. **Neural Computing Surveys**, v.1, p.102-350, 1998.

KATHMAN, R. M. Neural networks for the mass appraisal of real estate. **Computer Environment and Urban Systems**, v.17, p.373-384, 1993.

KAUKO, T. Can housing market segmentation be captured with a neural network modeling approach? In: Conference of the European Network for Housing Research (ENHR 2000), 2000, Gävle (Sweden). **Proceedings...** Uppsala (Sweden): Uppsala Universitet, 2000.

KAUKO, T. Exploring the prices of residential apartments and locality features within an artificial neural network approach – Evidence from Finland. In: AREUEA International Conference, 1997, Berkeley. **Proceedings...** Chicago: AREUEA, 1997.

KELEJIAN, H. H.; ROBINSON, D. P. Spatial autocorrelation: A new computationally simple test with an application to per capita county police expenditures. **Regional Science and Urban Economics**, v.22, p.317-331, 1992.

KINNARD Jr., W. N. **Income property valuation**. Lexington (MA): Heath Lexington, 1971.

KINNARD Jr., W. N. New thinking in appraisal theory. **Real Estate Appraiser**, Aug. 1966.

KLEMETTINEN, M.; MANNILA, H.; RONKALMEN, P.; TOIVONEN, H.; VERKAMO, A. I. Finding interesting rules from large sets of discovered association rules. In: 3rd International Conference on Information and Knowledge Management (CIKM'94), 1994, Gaithersburg (EUA). **Proceedings...**. Washington: ACM, 1994, p.401-407.

KLEMETTINEN, M.; MANNILA, H.; TOIVONEN, H. A data-mining methodology and its application to semi-automatic knowledge acquisition. In: 8th International Conference and Workshop on Database and Expert Systems Applications (DEXA-97), 1997, Toulouse (France). **Proceedings...** Piscataway (NJ): IEEE, 1997, p.670-676.

KMENTA, J. **Elementos de econometria**. São Paulo: Atlas, 1978.

KODRATOFF, Y. Technical and scientific issues of KDD (Or: Is KDD a science?). In: JANTKE, K.; SHINOHARA, T.; ZEUGMANN, T. (Eds.) **Algorithms learning theory**. Berlin: Springer-Verlag, 1995, p.261-265.

KOHAVI, R. Data mining and visualization. In: 6th Annual Symposium on Frontiers of Engineering, 2000, Irvine (CA). **Proceedings...** Washington: National Academy of Engineering, 2000.

KOHAVI, R.; JOHN, G. H. Wrappers for feature subset selection. **Artificial Intelligence**, v.97, n.1-2, p.273-324. 1997.

KOLODNER, J. **Case-based reasoning**. San Mateo (USA): Morgan Kaufmann, 1993.

KOVÁCS, Z. L. **Redes neurais artificiais:** Fundamentos e aplicações. 2ed. São Paulo: Collegium Cognitio/Acadêmica, 1996.

KUMMEROW, M. Thinking statistically about valuations. **Appraisal Journal**, p.318-325, July, 2000.

LANG, J. R.; JONES, W. H. Hedonic property valuation models: Are subjective measures of neighborhood amenities needed? **AREUEA Journal**, v.7, n.4, p.451-465, Winter, 1975.

LANGE, O. O objeto e método da economia. **Literatura Econômica**, v.7, n.2, p.207-230, 1985.

LAPOLLI, A. R. S.; DE CESARE, C. M.; LUNARDI, M. L. F.; OLIVEIRA, O. S. de; GRANDO, P. A**.** Metodologia para a determinação de regiões homogêneas de valorização imobiliária, tendo em vista a geração de informações cadastrais: o caso do município de Porto Alegre. In: 1º Congresso Brasileiro de Cadastro Técnico Multifinalitário (COBRAC), Ago. 1994, Florianópolis. **Anais...** Florianópolis: UFSC, 1994, tomo III, p.216-223.

LAVENDER, S. D. **Economics for builders and surveyors**. Essex (UK): Longman, 1990.

LEAKE, D. B. (Ed.). **Case-based reasoning**: Experiences, lessons and future directions. Menlo Park (CA)/Cambridge: AAAI/MIT, 1996.

LEAL, J. A. A. Financiamento habitacional e os requisitos para desenvolver o mercado de títulos hipotecários no Brasil: Uma análise a partir da experiência americana e chilena. In: 9º Encontro Nacional da ANPUR, 2001, Rio de Janeiro. **Anais...** Rio de Janeiro: ANPUR, 2001, v.3, p.1436-1445.

LEAL, J. A. A. **Políticas de integração da tributação sobre a renda e sobre a propriedade imobiliária urbana**. 1990. Dissertação (Mestrado em Planejamento Urbano) – Instituto de Pesquisa e Planejamento Urbano, Universidade Federal do Rio de Janeiro, Rio de Janeiro.

LENK, M. M.; WORZALA, E. M.; SILVA, A. High-tech valuation: Should artificial neural networks bypass the human valuer? **Journal of Property Valuation and Investment**, v.15, n.1, p.8-26, 1997.

LENTZ, G. H.; WANG, K. Residential appraisal and the lending process: A survey of issues. **Journal of Real Estate Research**, v.15, n.1/2, p.11-39, 1998.

LEU, S-S.; CHEN, A-T.; YANG, C-H. Fuzzy optimal model for resource-constrained construction scheduling. **Journal of Computing in Civil Engineering**, v.13, n.3, p.207-216, July, 1999.

LEWIS, O. M.; WARE, J. A.; JENKINS, D. Identification of residential property sub-markets using evolutionary and neural computing techniques. **Neural Computing and Applications**, v.10, p.108-119, 2001.

LEWIS, O. M.; WARE, J. A.; JENKINS, D. A novel neural network technique for the valuation of residential property. **Neural Computing and Applications**, v.5, n.4, p.224-229, 1997.

LI, C.; BISWAS, G. Knowledge-based scientific discovery in geological databases. In: 1st International Conference on Knowledge Discovery and Data Mining (KDD-95), Aug. 1995, Montreal (Canada). **Proceedings...** Menlo Park (CA): AAAI Press, 1995, p.204-210.

LI, M. M.; BROWN, H. J. Micro-neighbourhood externalities and hedonic housing prices. **Land Economics**, v.56, n.2, p.125-141, May, 1980.

LIMA, G. P. de A. Planta genérica de valores de terreno urbano – Organização e atualização. **Caderno Brasileiro de Avaliações e Perícias**, a.2, n.18, p.150-152, Dez. 1990.

LIPORONI, A. S. Cadastro imobiliário e planta de valores genéricos. **Caderno Brasileiro de Avaliações e Perícias**, n.54, p.167-175, Dez. 1993.

LIPPAI, I.; HEANEY, J.P.; LAGUNA, M. Robust water system design with commercial intelligent search optimizers. **Journal of Computing in Civil Engineering**, v.12, n.3, p.135-143, July, 1998.

LIU, H.; SETIONO, R. A probabilistic approach to feature selection - A filter solution. In: 13th International Conference on Machine Learning (ICML 1996), 1996, Bari. **Proceedings...** San Francisco: Morgan Kaufmann, 1996.

LUCENA, J. M. P. de. **O mercado habitacional no Brasil**. Rio de Janeiro: Fundação Getúlio Vargas, 1985. (Série Teses, n.9).

MACEDO, P. B. R. Hedonic price models with spatial effects: An application to the housing markets of belo Horizonte, Brazil. **Revista Brasileira de Economia**, v.52, n.1, p.63-81, Jan./Mar. 1998.

MACLENNAN, D. Some thoughts on the nature and purpose of house price studies. **Urban Studies**, v.14, p.59-71, 1977.

MADDALA, G. S. **Econometrics**. New York: McGraw-Hill, 1988.

MAIRE, F. Rule-extraction by backpropagation of polyedra. **Neural Networks**, v.12, p.717-725, 1999.

MALINVAUD, E. **Métodos estadísticos de la econometria**. Barcelona: Ariel, 1967.

MAN, K. F.; TANG, K. S.; KWONG, S. **Genetic algorithms**: Concepts and designs. London: Springer, 1999.

MANNILA, H. Methods and problems in data mining (a tutorial). In: International Conference on Database Theory (ICDT-97), 1997, Delphi (Greece). **Proceedings...** Berlin: Springer-Verlag, 1997, p. 41-55.

MARASCHIN, C. **Alterações provocadas pelo shopping center em aspectos da estrutura urbana - Iguatemi, Porto Alegre, RS**. 1993. Dissertação (Mestrado em Planejamento Urbano). Programa de Pós-Graduação em Planejamento Urbano e Regional, Universidade Federal do Rio Grande do Sul, Porto Alegre.

MARICATO, H. **Habitação e cidade**. 5ed. São Paulo: Atual, 1997.

MARICATO, H. **Política habitacional no regime militar**. Petrópolis: Vozes, 1987.

MARIR, F.; WATSON, I. Case-based reasoning: A categorised bibliography. **Knowledge Engineering Review**. v.9, n.3, 1994.

MARK, J.; GOLDBERG, M. A. Multiple regression analysis and mass assessment: A review of the issues. **Appraisal Journal**, v.56, n.1, p.89-109, Jan. 1988.

MARKS, D. The effect of rent control on the price of rental housing: An hedonic approach. **Land Economics**, v.60, n.1, p.81-94, Feb. 1984.

MARSTON, A.; AGG, T. R. **Engineering valuation**. New York: McGraw-Hill, 1936.

MASON, C.; QUIGLEY, J. M. Non-parametric hedonic housing prices. **Housing Studies**, v.11, n.3, p.373-385, 1996.

McCLUSKEY, W. J. Predictive accuracy of machine learning models for mass appraisal of residential property. **New Zealand Valuer's Journal**, p.41-47, July, 1996.

McCLUSKEY, W. J.; ANAND, S. The application of intelligent hybrid techniques for the mass appraisal of residential properties. **Journal of Property Investment and Finance,** v.17, n.3, p.218-239, 1999.

McCLUSKEY, W. J.; BORST, R. A. An evaluation of MRA, comparable sale analysis, and ANNs for the mass appraisal of residential properties in North Ireland. **Assessment Journal**, v.4, n.1, p.47-55, Jan./Feb. 1997

McCLUSKEY, W. J.; DEDDIS, W. G.; LAMONT, I. G.; BORST, R. A. The application of surface generated interpolation models for the prediction of residential property values. **Journal of Property Investment and Finance**, v.18, n.2, p.162-176, 2000.

McCLUSKEY, W. J.; DEDDIS, W.G.; MANNIS, A.; McBURNEY, D.; BORST, R.A. Interactive application of computer assisted mass appraisal and geographic information systems. **Journal of Property Valuation and Investment**, v.15, n.5, p.448-465, 1997.

McGREAL, S.; ADAIR, A.; MCBURNEY, D.; PATTERSON, D. Neural networks: The prediction of residential values. **Journal of Property Valuation and Investment**, v.16, n.1, p.57-70, 1998.

McMICHAEL, S. L. **McMichael's appraising manual**. 4ed. Englewood Cliffs (NJ): Prentice-Hall, 1962.

McSHERRY, D. An adaptation heuristic for case-based estimation. In: 4th European Workshop on Case-Based Reasoning (EWCBR-98), 1998, Dublin (Ireland). **Proceedings...** Berlin: Springer, 1998, p.184-195.

MEIRELLES, H. L. **Direito de construir**. 7ed. São Paulo: Malheiros, 1996.

MELAZZO, E. S. **Mercado imobiliário, expansão territorial e transformações intra-urbanas**. 1993. Dissertação (Mestrado em Planejamento Urbano) - Instituto de Pesquisa e Planejamento Urbano, Universidade Federal do Rio de Janeiro, Rio de Janeiro.

MENDONÇA, M. C. **Engenharia legal**: Teoria e prática profissional. São Paulo: Pini, 1999.

MILON, J. W.; GRESSEL, J.; MULKEY, D. Hedonic amenity valuation and functional form specification. **Land Economics**, v.60, n.4, p.378-387, Nov. 1984.

MILLS, A. C.; REYNOLDS, A. **The valuation of apartment properties**. Chicago: Appraisal Institute, 1999.

MITCHELL, M. **An introduction to genetic algorithms**. Cambrige (MA): MIT Press, 1999.

MITCHELL, P. S. The evolving appraisal paradigm. **Appraisal Journal**, p.189-197, Jan. 1993.

MITRA, S.; PAL, S. K.; MITRA, P. Data mining in soft computing framework: A survey. **IEEE Transactions on Neural Networks**, v.13, n.1, p.3-14, Jan. 2002.

MÖLLER, L. F. C. **Planta de valores genéricos**. Porto Alegre: Sagra-DC Luzzatto, 1995.

MOORE, D. S.; McCABE, G. P. **Introduction to the practice of statistics**. 3ed. New York: W. H. Freeman, 1998.

MOREIRA, A. L. **Princípios de engenharia de avaliações**. 4ed São Paulo: Pini, 1997

MOREIRA, T. A. A questão fundiária e as políticas sociais: Limitações às desapropriações de terras. In: 9º Encontro Nacional da ANPUR, 2001, Rio de Janeiro. **Anais...** Rio de Janeiro: ANPUR, 2001, v.3, p.1620-1643.

MORTON, T. G. Factor analysis, multicollinearity, and regression appraisal models. **Appraisal Journal,** v.45, p.578-587, Oct.1977.

MOSCOVITCH, S. K. Qualidade de vida urbana e valores de imóveis: Um estudo de caso para Belo Horizonte. **Nova Economia**, número especial, p.247-278, 1997.

MUTH, R. F. **Urban economic problems**. New York: Harper and Row, 1975.

MYRDAL, G. **Aspectos políticos da teoria econômica**. Rio de Janeiro: Zahar, 1962.

NATH, R.; RAJAGOPALAN, B.; RYKER, R. Determining the saliency of input variables in neural network classifiers. **Computers and Operations Research**, v.24, n.8, p.767-773, Aug. 1997.

NETER, J.; WASSERMANN, W.; KUTNER, M. H. **Applied linear statistical models**. Regression, analysis of variance, and experimental designs. 3ed. Burr Ridge (IL): Richard D. Irwin, 1990.

NEWELL, G.; KISHORE, R. The accuracy of commercial property valuations. In: 4th Annual Pacific-Rim Real Estate Society Conference, Jan. 1998, Perth (Austrália). **Proceedings...** Perth: PRRES, 1998.

NEWSOME, B. A.; ZIETZ, J. Adjusting comparable sales using multiple regression analysis – The need of segmentation. **Appraisal Journal**, p.129-135, Jan. 1992.

NG, R. T.; HAN, J. Efficient and effective clustering methods for spatial data mining. In: 20th International Conference on Very Large Data Bases (VLDB'94), 1994, Santiago. **Proceedings...** San Francisco: Morgan Kaufmann, 1994, p.144-155.

NGUYEN, N. T.; CRIPPS, A. Predicting housing value: A comparison of multiple regression analysis and artificial neural networks. **Journal of Real Estate Research**, v.22, n.3, 2001.

NGUYEN, H. T.; WALKER, E. A. **A first course in fuzzy logic**. 2ed. Boca Ratón: Chapman and Hall, 2000.

NIE, N. H. *et al*. **SPSS** - Statistical Package for the Social Sciences. 2ed. New York: McGraw-Hill, 1975.

NIKOLOPOULOS, C. **Expert systems**. Introduction to first and second generation and hybrid knowledge based systems. New York: Marcel Dekker, 1997.

NILSSON, N. J. **Artificial inteligence**: A new synthesis. San Francisco: Morgan Kaufmann, 1998.

NORDVIK, V. Prices and price expectation in the market for owner occupied housing. **Housing Studies**, v.10, n.3, p.365-380, 1995.

O'ROARTY, B.; McGREAL, S.; ADAIR, A. ; PATTERSON, D. Case-based reasoning and retail rent determination. **Journal of Property Research**, v.14, n.4, p.309-328, 1997a.

O'ROARTY, B.; PATTERSON, D.; McGREAL, S.; ADAIR, A. A case-based reasoning approach to the selection of comparable evidence for retail rent determination. **Expert Systems with Applications**, v.12, n.4, p.417-428, 1997b.

OLIVEIRA, L. R. de; RETIK, A.; WATSON, I. The integration of VR and CBR to visually represent past experiences. In: 4th Conference of Applications of Artificial Intelligence in Structural Engineering, 1997, Lahti (Finland). **Proceedings…** Tampere (Finland): Tampere University of Technology, 1997, p.105-116.

OLSEN, E. O. An econometric analysis of rent control. **Journal of Political Economy**, v.80, n.6, p.1081-1100, Nov./Dec., 1972.

PACE, R. K. Appraisal using generalized additive models. **Journal of Real Estate Research**, v.15, n.1/2, p.77-99, 1998.

PACE, R. K.; BARRY, R.; SIRMANS, C. F. Spatial statistics and real estate. **Journal of Real Estate Finance and Economics**, v.17, n.1, p.5-15, July, 1998.

PACHARAVANICH, P.; ROSSINI, P. Examining the potential for the development of computerised mass appraisal in Thailand. In: 7th Annual Pacific-Rim Real Estate Society Conference, Jan. 2001, Adelaide. **Proceedings...** Adelaide: PRRES, 2001.

PACHARAVANICH, P.; WONGPINUNWATANA, N.; ROSSINI, P. The development of a case-based reasoning system as a tool for residential valuation in Bangkok. In: 6th Annual Pacific-Rim Real Estate Society Conference, Jan. 2000, Sydney. **Proceedings...** Sydney, PRRES, 2000.

PANAYIOTOU, P. A.; PATTICHIS, C.; JENKINS, D.; PLIMMER, F.; ECONOMIDES, C; PATTICHIS, C. S; MALIOTIS, G; PAPACONSTANTINOU, M; MICHAELIDES, N; SCHIZAS, C. N. A modular artificial neural network valuation system. In: 10th Mediterranean Electrotechnical Conference (MeleCon2000), 2000, Lemesos (Cyprus). **Proceedings...** Piscataway (NJ): IEEE Press, 2000, v.2, p.457-460.

PARKER, D. R. R. Valuation accuracy - An australian perspective. In: 4th Annual Pacific-Rim Real Estate Society Conference, Jan. 1998, Perth. **Proceedings...** Perth: PRRES, 1998.

PELLEGRINO, J. C. A propósito do valor potencial - raízes, problemas e implicações. In: IBAPE, 1983, p.9-17.

PEREIRA, P. L. V. Co-integração e suas representações: uma resenha. **Revista Brasileira de Econometria**, v.11, n.2, p.185-216, Nov. 1991.

PINDYCK, R. S.; RUBINFELD, D. L. **Microeconomia**. 5ed. São Paulo: Prentice-Hall, 2002.

PINHO, D. B.; VASCONCELLOS, M. A. S. de. **Manual de economia**. 3ed. São Paulo: Saraiva, 1997.

PLATE, T. A.; BERT, J.; GRACE, J.; BAND, P. Visualizing the function computed by a feedforward neural network. **Neural Computation**, v.12, p.1337-1353, 2000.

POGODZINSKI, J. M.; SASS, T. R. Zoning and hedonic housing price models. **Journal of Housing Economics**, v.1, p.271-292, 1991.

PORTO ALEGRE. **Estatísticas**. Porto Alegre: Prefeitura Municipal, 1970 a 2000. (anual).

PORTUGAL, M. S.; FERNANDES, L. G. L. Redes neurais artificiais e previsão de séries econômicas: Uma introdução. **Nova Economia**, v.6, n.1, p.51-74, Jul. 1996.

PYLE, D. **Data preparation for data mining**. San Francisco: Morgan Kaufmann, 1999.

QUINLAN, J. R. **C4.5 programs for machine learning**. San Mateo (CA): Morgan Kaufmann, 1993.

RAMANATHAN, R. **Introductory econometrics** - With applications. 4ed. Forth Worth (USA): Dryden, 1998.

RAMSLAND, M. O.; MARKHAM, D. E. Market-supported adjustments using multiple regression analysis. **Appraisal Journal**, p.181-191, Apr. 1998.

RATCLIFFE, J. S.; SIMON, T. Y.; THOMAS, P. N.; ZHAOLING, Y.; YABIAO, Y. **An examination of land and property appraisal techniques suitable for application in the people's republic of China**. Hong Kong: Department of Building and Real Estate/H.K. Polytechnic, 1993.

RAYBURN, W. B.; TOSH, D. S. Artificial intelligence: The future of appraising. **Appraisal Journal**, p.429-435, Oct. 1995.

REATEGUI, E.; CAMPBELL, J. A.; BORGHETTI, S. Using neural netwok to learn general knowledge in a case-based system. **Case-Based Reasoning Research and Development**, 1995.

REICH, Y. Machine learning techniques for civil engineering problems. **Microcomputers in Civil Engineering**, v.12, n.4, p.307-322, 1997.

RENSTIRENA, E. T. Risk assessment. **Appraisal Journal**, p.288-292, July, 1996.

RIBEIRO, F. L. Using multiagent systems technology and case-based reasonig in the valuation of residential and office properties. In: Annual Construction Research

Conference (COBRA-1999), 1999, London. **Proceedings...** London: Royal Institution on Chartered Surveyors (RICS), 1999, p.33-44.

RING, A. A. **The valuation of real estate**. Englewood Cliffs: Prentice-Hall, 1965.

ROBINSON, R. **Housing economics and public policy**. London: MacMillan, 1979.

ROBNIK-SIKONJA, M.; KONONENKO, I. An adaptation of Relief for attribute estimation in regression. In: 14th International Conference on Machine Learning (ICML 1997), 1997, Nashville (TE). **Proceedings...** San Francisco: Morgan Kaufmann, 1997, p. 296-304.

ROLNIK, R.; CYMBALISTA, R. (Orgs.). **Instrumentos urbanísticos contra a exclusão social**. São Paulo: Pólis, 1997. (Publicações Pólis, n.29).

ROSEN, S. Hedonic prices and implicit markets: product differenciation in pure competition. **Jornal of Political Economy**, n.82, p.34-55, 1974.

ROSSINI, P. Application of artificial neural networks to the valuation of residential property. In: 3rd Annual Pacific-Rim Real Estate Society Conference, Jan. 1997, Palmerston North (New Zealand). **Proceedings...** Palmerston North: PRRES, 1997.

ROVATTI, J. F. **A fertilidade da terra urbana em Porto Alegre**: Uma leitura da intervenção do estado na cidade. 1990. Dissertação (Mestrado em Planejamento Urbano) - Instituto de Pesquisa e Planejamento Urbano, Universidade Federal do Rio de Janeiro, Rio de Janeiro.

ROVATTI, J. F. **Produção capitalista de moradias em Porto Alegre (anos oitenta).** In: Seminário de Incorporação Imobiliária. Rio de Janeiro: IPPUR/UFRJ, 1991.

RUMELHART, D. E.; MACCLELLAND, J. L. **Parallel distributed processing**: Explorations in the microstructure of cognition. Cambridge: MIT, 1995.

SALENGUE, L. G. de P.; MARQUES, M. M. Reavaliação de planos diretores: O caso de Porto Alegre. In: PANIZZI, W. M.; ROVATTI, J. F. **Estudos Urbanos** – Porto Alegre e seu planejamento. Porto Alegre: UFRGS, 1993, p.155-164.

SAUTER, B. Computers and comparable sales. In: WOOLERY, A.; SHEA, S. (Eds.) **Introduction to computer assisted valuation**. Boston: Oelgeschlager, Gunn and Hain/Lincoln Institute of Land Policy, 1985, chap. 9, p.141-147.

SCIASCIA, G. **Regras de Ulpiano**. São Paulo: Ed.Autor, 1952.

SEELEY, I. H. **Building economics**. 2ed. London: MacMillan, 1976.

SETIONO, R. Extracting M-of-N rules from trained neural networks. **IEEE Transactions on Neural Networks**, v.11, n.2, p.512-519, Mar. 2000.

SETIONO, R. Extracting rules from neural networks by pruning and hidden-unit splitting. **Neural Computation**, n.9, p.205-225, 1997.

SETIONO, R.; THONG, J.Y.L.; YAP, C.-S. Symbolic rule extraction from neural networks. An application to identifying organizations adopting IT. **Information and Management**, v.14, p.91-101, 1998.

SHAFER, J.; AGRAWAL, R.; MEHTA, M. SPRINT: A scalable parallel classifier for data mining. In: 22th International Conference on Very Large Data Bases (VLDB'96), 1996, Bombay. **Proceedings...** San Francisco: Morgan Kaufmann, 1996.

SHARIF, M.; WARDLAW, R. Multireservoir systems optimization using genetic algorithms: Case study. **Journal of Computing in Civil Engineering**, v.14, n.4, p.255-263, Oct. 2000.

SHEPPARD, S. Hedonic analysis of housing markets. In: CHESHIRE, P. C.; MILLS, E. S. (Eds.) **Handbook of applied urban economics**, v.3. New York: Elsevier, 1999, chap.8.

SHILLER, R. J.; WEISS, A.N. Evaluating real estate valuation systems. **Journal of Real Estate Finance and Economics**, v.18, n.2, p.147-161, Mar. 1999.

SIU, K. K.; YU, S. M. Using response surface analysis in mass appraisal to examine the influence of location on property values in Hong Kong. In: 7th Annual Pacific-Rim Real Estate Society Conference, Jan. 2001, Adelaide. **Proceedings...** Adelaide: PRRES, 2001.

SMALLEY, S. P. Appraisal: Science or art? **Appraisal Journal**, p.165-171, 1995.

SMITH, H. C. Inconsistencies in appraisal theory and practice. **Journal of Real Estate Research**, v.1, n.1, p.1-17, 1986.

SMITH, L. B.; ROSEN, K. T.; FALLIS, G. Recent development in economic models of housing markets. **Journal of Economic Literature**, v.26, p.29-64, Mar. 1988.

SMITH, T. R. Statistical implications of the most probable price. **Appraisal Journal**, p.81-86, Jan. 1995.

SMOLKA, M. O. Argumentos para a rehabilitação do IPTU e do ITBI como instrumentos de intervenção urbana (progressista). In: 1º Congresso Brasileiro de Cadastro Técnico Multifinalitário (COBRAC), Ago. 1994, Florianópolis. **Anais...** Florianópolis: UFSC, 1994a, Tomo III, p.170-187.

SMOLKA, M. O. Problematizando a intervenção urbana: Falácias, desafios e constrangimentos. **Cadernos IPPUR**, a.8, n.1, p.29-42, Abr. 1994b.

SMOLKA, M. O. et al. **Dinâmica imobiliária e estruturação urbana**: O caso do Rio de Janeiro. 1989. Relatório de pesquisa. Instituto de Pesquisa e Planejamento Urbano, Universidade Federal do Rio de Janeiro, Rio de Janeiro.

SMOLKA, M. O.; FURTADO, F. (Eds.). **Recuperación de plusvalías en América Latina**. Santigo: Eurelibros, 2001.

SOH, C. K.; YANG, J. Fuzzy controlled genetic algorithm search for shape optimization. **Journal of Computing in Civil Engineering**, v.10, n.2, p.143-214, April, 1996.

SOH, C. K.; YANG, Y. Genetic programming-based approach for structural optimization. **Journal of Computing in Civil Engineering**, v.14, n.1, p.31-37, Jan. 2000.

SOIBELMAN, L.; KIM, H. Data preparation process for construction knowledge generation through knowledge discovery in databases. **Journal of Computing in Civil Engineering**, v.16, n.1, p.39-48, June, 2002.

SOVAT, R. B.; CARVALHO, A. C. P. L. F. de. Um ambiente para desenvolvimento de sistemas de raciocínio baseado em casos. In: 19º Congresso Nacional da Sociedade Brasileira de Computação, 1999, Rio de Janeiro. **Anais...** Rio de Janeiro: EntreLugar, 1999, v.4, p.133-148.

SPANOS, A. **Statistical foundations of econometric modeling**. Cambridge: Cambridge University Press, 1989.

SPEARS, W. M. Crossover or mutation? In: WHITLEY, L. D. (Ed.). **Foundation of genetic algorithms**, v.2, p.221-237. San Mateo: Morgan Kaufmann, 1993.

SPEARS, W. M. Adapting crossover in evolutionary algorithms. In: 4th Annual Conference on Evolutionary Programming, 1995, San Diego (CA). **Proceedings...** Cambridge (MA): MIT Press, 1995, p.367-384.

SPENCE, M. T.; THORSON, J. A. The effect of expertise on the quality of appraisal services. **Journal of Real Estate Research**, v.15, n.1/2, p.205-215, 1998.

SPSS. **User's guide base 9.0**. Chicago: SPSS Inc., 1999.

STONE, C.; HANSEN, M.; KOOPERBERG, C.; TRUONG, Y. K. Polynomial splines and their tensor products in extended linear modeling. **Ann Statistics**, v.25, p.1371-1470, 1997.

STRASZHEIM, M. The theory of urban residential location. In: E.S. MILLS (ed). **Handbook of regional and urban economics**, v.2 (urban economics). Amsterdam: Elsevier, 1987, chap.18, p.717-757.

SYCARA, K. P. Using case-based reasoning for plan adaptation e repair. In: DARPA Case-Based Reasoning Workshop, 1988. **Proceedings...** San Francisco: Morgan Kaufmann, 1988, p.425-434.

TAY, D. P. H.; HO, D. K. H. Intelligent mass appraisal. **Journal of Property Tax Assessment and Administration**, v.1, n.1, p.5-25, Sept. 1994.

THRALL, G. I. GIS applications in real estate and related industries. **Journal of Housing Research**, v.9, n.1, p.33-60, 1998.

TOIVONEN, H. Sampling large databases for association rules. In: 22th International Conference on Very Large Data Bases (VLDB'96), 1996, Bombay. **Proceedings...** San Francisco: Morgan Kaufmann, 1996.

TUNG, A. K. H.; HOU, J.; HAN, J. Spatial clustering in the presence of obstacles. In: International Conference on Data Engineering (ICDE'01), 2001, Heidelberg. **Proceedings...** Piscataway (NJ): IEEE, 2001, p.359-367.

VALLADARES, L. P. (Org). **Habitação em questão**. 2ed. Rio de Janeiro: Zahar, 1981.

VARSANO, R. O imposto predial e territorial urbano: receita, eqüidade e adequação aos municípios. **Pesquisa e Planejamento Econômico**, v.7, n.3, p.581-622, Dez. 1977.

VASCONCELLOS, M. A. S.; ALVES, D. (Eds.) **Manual de econometria**: Nível intermediário. São Paulo: Atlas, 2000.

VEGNI-NERI, G. B. D. **Avaliação de imóveis urbanos e rurais**: Método prático e moderno. 4ed. São Paulo: Nacional, 1979.

VEGNI-NERI, G. B. D. **Prática de avaliação de imóveis**. 2ed. São Paulo: Legislação Brasileira, 1968.

VENTOLO Jr., W. L.; WILLIAMS, M. R. **Técnicas del avalúo inmobiliário**: Guía completa para vendedores, corretores, administradores, inversionistas y valuadores de propiedades. Chicago: Real Estate Education Company, 1997.

WACHS, P. Implementation of computerized real estate assessment. **Journal of American Institute of Planners**, v.44, n.1, p.60-68, Jan. 1978.

WALLER, B. D. The impact of AVMs on the appraisal industry. **Appraisal Journal**, p.287-292, July, 1999.

WANG. H.; BELL, D.; MURTAGH, F. Feature subset selection based on relevance. **Vistas in Astronomy**, v.41, n.3, p.387-396, 1997.

WARD, R. D.; WEAVER, J. R.; GERMAN, J. C. Improving CAMA models using geographic information systems/response surface analysis location factors. **Assessment Journal**, n.6, p.30-38, Jan. 1999.

WATSON, I. **Applying case-base reasoning**: Techniques for enterprise systems. San Francisco: Morgan Kaufmann, 1997.

WATSON, I.; ABDULLAH, S. Diagnosing building defects using case-based reasoning. In: Tokyo Symposium on Strategies and Tecnnologies for Maintenance and Modernisation of Buildings (CIB W70), 1994, Tokyo. **Proceedings...** Paris: International Council for Building Research Studies and Documentation (CIB), 1994.

WATSON, I.; MARIR, F. Case-based reasoning: A review. **Knowledge Engineering Review**, v.9, n.4, 1994.

WATSON, I.; OLIVEIRA, L. Virtual reality as an environment for CBR. In: 4th European Workshop on Case-Based Reasoning (EWCBR-98), 1998, Dublin (Ireland). **Proceedings...** Berlin: Springer, 1998, p.448-459.

WEBER-LEE, R. **Pesquisa jurisprudencial inteligente**. 1998. Tese (Doutorado em Engenharia de Produção) – Programa de Pós-Graduação em Engenharia de Produção, Universidade Federal de Santa Catarina, Florianópolis.

WEBER-LEE, R.; BARCIA, R. M.; COSTA, M. C. da; RODRIGUES FILHO, I. W.; HOESCHL, H. C.; BUENO, T. C. D. A.; MARTINS, A.; PACHECO, R. C. A large case-based reasoner for legal cases. In: 2nd International Conference on Case-Based Reasoning (ICCBR-97), 1997, Providence (USA). **Proceedings...** Berlin: Springer, 1997, p.190-199.

WEBER-LEE, R.; BARCIA, R. M.; KHATOR, S. K. Case-base reasoning for cash flow forecasting using fuzzy retrieval. In: 1st International Conference on Case-Based

Reasoning (ICCBR-95), 1995, Sesimbra (Portugal). **Proceedings...** Berlin: Springer, 1995, p.510-519.

WEIMER, A. M. History of value theory for the appraiser. **Appraisal Journal**, p.8-22, Jan. 1953. (In: **Selected readings in the real estate appraisal**. Chicago: American Institute of Real Estate Appraisers, 1953).

WEIMER, A. M.; HOYT, H. **Principles of urban real estate**. rev.ed. New York: Ronald Press, 1948.

WEISBERG, S. **Applied linear regression**. 2ed. New York: John Wiley, 1985.

WEISS, S. M.; INDURKHYA, N. **Predictive data mining**: A practical guide. San Francisco: Morgan Kaufmann, 1998.

WENDT, P. F. **Real estate appraisal** – Review and outlook. Athens: University of Georgia Press, 1974.

WERNA, E.; ABIKO, A.K.; COELHO, L.O.; SIMAS, R.; KEIVANI, R.; HAMBURGER, D.S.; ALMEIDA, M.A.P. **Pluralismo na habitação**. São Paulo: Annablume, 2001.

WESTPHAL, C.; BLAXTON, T. **Data mining solutions**: Methods and tools for solving real-world problems. New York: Wiley Computer Publishing, 1998.

WHITE, H. **Artificial neural netwoks**. Approximation and learning theory. Cambridge (USA): Blackwell, 1992.

WHITE, J. R. Relationship of real estate cost and value. **Appraisal Journal**, p.243-251, Apr. 1950. (In: **Selected readings in the real estate appraisal**. Chicago: American Institute of Real Estate Appraisers, 1953).

WILDERODE, D. J. van. Operações interligadas: Engessando a perna de pau. In: ROLNIK, R.; CYMBALISTA, R. (Orgs.). **Instrumentos urbanísticos contra a exclusão social**. São Paulo: Pólis, 1997.

WILKINSON, R. K.; ARCHER, C. A. Measuring the determinants of relative house prices. **Environment and Planning**, v.5, n.3, p.357-367, May, 1973.

WINCOTT, R. D.; MUELLER, G. R. Market analysis in the appraisal process. **Appraisal Journal**, Jan. 1995.

WITTEN, I. H.; ALISTAIR, M.; BELL, T. C. **Managing gigabytes**: Compressing and indexing documents and images. 2ed. San Francisco: Morgan Kaufmann, 1999.

WITTEN, I. H.; FRANK, E. **Data mining**: Practical machine learning tools and techniques with java implementations. San Francisco: Morgan Kaufmann, 2000.

WOLPERT, D. H.; MACREADY, W. G. No free lunch theorems for optimisation. **IEEE Transactions on Evolutionary Computation**, v.1, n.1, p.67-82, Apr. 1997.

WOLPERT, D. H.; MACREADY, W. G. No free lunch theorems for search. **Working Paper SFI-TR 95-02-010**. The Santa Fe Institute: Santa Fe (NM), 1995.

WOOD, S. Combining forecasts to predict property values for single-family residences. **Land Economics**, v.52, n.2, p.221-229, May, 1976.

WORZALA, E. M.; LENK, M. M.; KINNARD Jr., W. N. How client pressure affects the appraisal of residential property. **Appraisal Journal**, Jan. 1998.

WORZALA, E.; LENK, M.; SILVA, A. An exploration of neural networks and its application to real estate valuation. **Journal of Real Estate Research**, v.10, n.2, p.185-201, 1995.

WYATT, P. The development of a property information system for valuation using a geographical information system (GIS). **Journal of Property Research**, v.13, p.317-336, 1996a.

WYATT, P. Using a geographical information system for property valuation. **Journal of Property Valuation and Investment**, v.14, n.1, p.67-79, 1996b.

XU, X.; ESTER, M.; KRIEGEL, H-P. SANDER, J. Clustering and knowledge discovery in spatial databases. **Vistas in Astronomy**, v.41, n.3, p.397-403, 1997.

YAGER, R. R. Case based reasoning, fuzzy systems modeling and solution composition. In: 2nd International Conference on Case-Based Reasoning (ICCBR-97), 1997, Providence (USA). **Proceedings...** Berlin: Springer, 1997, p.633-642.

YANG, J.; SOH, C. K. Structural optimization by genetic algorithms with tournament selection. **Journal of Computing in Civil Engineering**, v.11, n.3, p.195-200, July, 1997.

YANG, Y.; SOH, C. K. Fuzzy logic integrated genetic programming for optimization and design. **Journal of Computing in Civil Engineering**, v.14, n.4, p.249-254, Oct. 2000.

YAO, Z. Building expert system for real estate. **Modeling, Measurement and Control - D**, v.9, n.1, p.53-64, 1994.

ZADEH, L. A. Fuzzy sets. **Information and control**, v.8, p.338-352, 1965.

ZANCAN, E. C. **Avaliação de imóveis em massa para efeitos de tributos municipais**. Florianópolis: Rocha, 1996.