FINANÇAS

VARIÁVEIS EXPLICATIVAS E A SUA IMPORTÂNCIA NA FORMAÇÃO DO PREÇO DE UM APARTAMENTO EM PORTUGAL: UMA ABORDAGEM COM REDES NEURONAIS ARTIFICIAIS.

Maria Cristina Canavarro Teixeira (ccanavarro@ipcb.pt)
Instituto Politécnico de Castelo Branco
Escola Superior Agrária
Quinta da N. Srª de Mércules, Apartado 119
6001-909 Castelo Branco (Portugal)

José Maria Caridad y Ocerín (ccjm@uco.es)
Nuria Ceular Villamandos (tdcevin@uco.es)
Universidade de Córdoba (Espanha)

RESUMO:

Conhecer o preço da habitação com objectividade, através da medida das suas características é um assunto que ocupa diversos investigadores. A selecção das variáveis explicativas foi validade através dos testes de diagnóstico usuais na modelação econométrica. Este estudo descreve uma investigação sobre a estimação de uma rede neuronal artificial, para o preço de venda de um apartamento numa cidade Portuguesa. Apresentam-se os resultados obtidos na estimação de várias redes neuronais artificiais, em relação ao erro relativo do conjunto de teste, com três tipos de partições de dados diferentes. A partição que funcionou melhor foi a 90% vs 10%, respectivamente para o conjunto de aprendizagem e de teste. Também foi analisada a possibilidade da ordem de importância das variáveis explicativas do preço, estar dependente do tipo de partição utilizada, tendo-se concluído que a área útil, é sempre a variável explicativa que surge com maior importância, independentemente do tipo de partição utilizada.

PALAVRAS CHAVE: Preço da habitação, variáveis explicativas, redes neuronais artificiais, erro relativo, conjunto de aprendizagem e de teste.

ABSTRACT:

The estimation of real estate prices has been the object of many studies, using the characteristics of each dwelling as exogenous variables. The selection of explanatory variables has been validated through the usual diagnostics tests in econometric modeling. This study describes the results of artificial neural network as an alternative to classical hedonic modeling. A case study is described for a medium size Portuguese city. The errors obtained with the test set, using different artificial neural networks and with three types of different data partitions are compared. The partition of the sample that produces the best performance was 90% versus 10%, respectively, for the learning and the test set. Also, the possibility of being dependent of the partition type used was analyzed, to evaluate its importance on the price estimation. As expected, the variable size of the dwelling is always the most important explanatory variable, independently of partition type set used.

KEY WORDS: price of housing, exogenous variables, neural networks, relative error, training and testing set.

1. INTRODUÇÃO

O estudo da evolução dos preços no mercado da habitação é importante por diversos motivos. Em primeiro lugar, porque pode constituir um ponto final a uma série de ligações que determinam o acesso da população ao mercado imobiliário (mediante uma comparação entre o preço médio e a renda familiar disponível). De facto, a razão fundamental que explica a dificuldade de acesso (compra) de habitação é o seu preço elevado.

Em segundo lugar, a habitação é um activo, que em comparação com outros bens de primeira necessidade, constitui o bem de primeira necessidade com preço mais elevado na nossa sociedade, sem dúvida, o mais importante (e o mais pesado) na carteira das famílias. O aumento do preço da habitação, supõe um efeito riqueza que pode ter consequências importantes sobre o equilíbrio macroeconómico, já que incrementos da riqueza, em teoria, darão lugar a aumentos de consumo das famílias e na procura conjunta.

Em terceiro lugar, o sector da habitação residencial apresenta períodos de forte expansão, seguidos de anos de recessão mais acentuada que qualquer outro sector económico, já que o preço da habitação e o ciclo económico estão estritamente relacionados. Por conseguinte, a sua evolução também tem consequências sobre o mercado de trabalho ligado à construção e sobre o mercado de materiais de construção.

Desta forma, o investimento no mercado de habitação não é o mesmo que investir num activo sem risco nenhum, pois a rentabilidade deste investimento, também pode chegar a ser negativa ou substancialmente inferior à gerada por outros activos. Existem múltiplos exemplos que ilustram esta afirmação: no Reino Unido, durante o *boom* do imobiliário no começo dos anos 70 (1970 – 1973) e até finais de oitenta (1986 – 1989) os preços experimentaram taxas de crescimento anuais, que em alguns casos, superaram os 20%. Este crescimento, foi seguido de uma contracção nos preços que caíram à roda dos 40% entre 1973 e 1977, enquanto entre 1989 e 1992 a caída foi de aproximadamente 30%; nos Estados Unidos o preço da habitação nova aumentou durante os anos 70 em 30%, enquanto na recessão do princípio de 90, os preços da costa oeste de EE.UU. chegaram a ter aumentos de 40%; outros casos mais recentes são o Japão e Hong Kong, ou no nosso enquadramento europeu mais próximo, o caso da Alemanha, Áustria ou Espanha.

Evidentemente que devemos assinalar a existência de mercados locais especiais, devido à sua localização geográfica. Em Portugal, por exemplo, existem consideráveis diferenças entre as diferentes regiões (interior e costa), grandes cidades e também dentro das mesmas.

A comparação do preço da habitação com outros bens de primeira necessidade constitui uma tarefa difícil. O preço da habitação, como o seu custo, apresenta uma grande heterogeneidade. O preço varia em função da localização, tamanho, tipo de habitação (moradias, apartamentos em bloco, etc.) qualidade de construção, etc. Além disso as características das habitações também variam com o tempo. Consequentemente a simples evolução do preço médio das habitações compradas e vendidas em cada período pode não ser, o indicador mais adequado para observar a evolução. As séries estatísticas disponíveis, recorrem, na maior parte dos casos, a correcções dos factores diferenciais mais óbvios, como o tamanho, utilizando a medição do preço médio por metro quadrado (embora esta correcção seja apenas superficial se tivermos em conta que a relação preço, área não é linear), isto é, do preço médio das habitações de um determinado tamanho (área).

Também se devia ter em conta a composição do agregado da habitação que se utiliza como representativo do parque total, isso é, o preço do metro quadrado obtido como resultado dependerá do número de habitações novas e usadas, assim como, se na nossa amostra existirem habitações para fins sociais.

Uma característica que o nosso mercado imobiliário apresenta, à semelhança de outros, é que os preços de transacção dos imóveis são variáveis não observadas, no sentido de que o que se dispões é da informação prestada pelo comprador e vendedor sobre o preço do imóvel transaccionado.

Este facto assume especial importância no mercado imobiliário português dado o conhecido fenómeno de "fraude e evasão fiscal", com o consequente risco de uma menor credibilidade da informação disponível sobre preços de transacção dos imóveis. Efectivamente, tenderá a haver uma sub estimação dos preços de transacção declarados, de modo a suportar menores encargos fiscais com a aquisição do imóvel (comprador) ou menos carga fiscal sobre os lucros ou mais-valias da venda (vendedor).

Por outro lado, relativamente aos valores da avaliação pelas entidades de crédito com vista à concessão ao empréstimo bancário, poderá existir uma sob estimação, devido à percentagem do valor de empréstimo ser inferior a 100%. Assim, com uma maior avaliação do imóvel, o valor concedido no empréstimo cobre o montante do valor real de compra, neste caso inferior, ao valor da avaliação.

As dificuldades e limitações na disponibilidade e acesso a dados sobre transacções e características dos imóveis são comuns a vários países.

2. ALGUMA REVISÃO DE LITERATURA

A partir dos anos 90, a previsão com redes neuronais artificiais (RNA's) teve um tremendo avanço. Desde os trabalhos pioneiros de Borst em 1991, que os modelos com redes neuronais artificiais se têm tornado uma alternativa muito atractiva aos modelos econométricos tradicionais. A vantagem principal destas técnicas é a capacidade de lidar com as relações não lineares, e com formas funcionais inicialmente desconhecidas. A literatura mostra que há um misto de sucesso e insucesso com este método, provavelmente devido a diferentes variáveis de entrada e de diferentes condições de mercado.

Pouco tempo depois de Borst ter apresentado o seu trabalho, Do, Quang e Grudnitski em 1993, utilizam as RNA's para mostrar que o valor de uma casa desce significativamente com a sua idade, durante os primeiros 16 a 20 anos, resultado da deterioração física. A partir dessa idade, não só a diminuição do valor pára (devida ao tempo), mas também a casa começa a experimentar uma apreciação relacionada, em parte, pelo tamanho do seu lote. Neste trabalho foi usada informação proveniente de agentes imobiliárias, em San Diego, Califórnia, relativas a 242 moradias vendidas, no período de Janeiro de 1991 até Setembro de 1991.

Testes robustos de modelos de RNA's requerem a separação do conjunto de dados. É necessário definir um modelo de formação e um novo conjunto de dados para testar os modelos (Rossini, 1997). Esta metodologia foi aplicada para vários países, utilizando conjuntos de dados com as características específicas de cada local. Por exemplo em 1991, Borst usou as RNA's para conjuntos de dados de residências familiares na Nova Inglaterra, Do e Grudnitiski (1992) usaram dados de um serviço de listagem múltipla na Califórnia, enquanto Evans et al. (1993) trabalharam com habitação no Reino Unido. Worzala et al., (1995), Borst (1995), Borst et al., (1996), e McCluskey et al., (1997), usaram múltiplos conjuntos de aprendizagem, e compararam os resultados obtidos pelas RNA's e pelos modelos de Regressão Múltipla.

Segundo Borst (1995), a precisão das RNA's, torna-as rivais dos métodos de Regressão Linear Múltipla. O autor considera que estas, podem ser utilizadas na avaliação em massa, bem como num controlo de qualidade sobre os valores estimados por outros métodos. Em 1996, Borst e McCluskey atestam que as capacidades de previsão das RNA estão perfeitamente bem definidas através de estudos de investigação.

No ano seguinte, Rossini, fundamentado nos trabalhos dos seus antecessores, aplicou esta técnica para os dados do Sul Australiano. Para tal foram usados os dados relativos às vendas registadas pelo Department of Environment and Natural Resources (DENR) no Sul da Austrália, e acedidos através do sistema de vendas recuperadas UPmarket. O DENR recolhe os detalhes de todas as vendas ocorridas no Sul da Austrália, e torna-as acessíveis em formato digital. Um vasto conjunto de informação está disponível para cada propriedade incluindo detalhes da venda, valores de avaliação, informações sobre o local e características físicas no caso de se tratar de propriedades residenciais. No seu estudo, Rossini, utilizou três procedimentos para comparar os modelos de RNA's na estimação do valor do mercado imobiliário, com o modelo de regressão linear múltipla.

Neste seu trabalho, Rossini chegou à conclusão de o uso dos modelos de Regressão, é preferível em vez das RNA's, advertindo contudo que estes resultados não são completamente conclusivos (Rossini, 1997). Embora tenha chegado a esta conclusão (com o referido conjunto de dados), acredita que num futuro muito próximo, com o aumento das ferramentas computacionais, que as RNA's se venham a tornar uma poderosa ferramenta. Convicto desta certeza, em 1999, o autor apresenta conjuntamente com outro investigador outro trabalho com RNA's. Kershaw e Rossini (1999), usaram uma série de dados de casas para desenvolver *Constant Quality House Price Indices*, usando redes neuronais e técnicas econométricas. Neste trabalho ficou provado que as RNA's podem ser uma séria alternativa aos métodos econométricos.

Entretanto, Zhang, et al., (1998) apresentam o estado de arte da aplicação das RNA's em previsão. O objectivo destes autores foi o de sintetizar a investigação nesta área, o conhecimento profundo nas técnicas de modelar as RNA's, e sugerir a direcção futura de investigação. Há mais de dez anos, os investigadores ainda não tinham

certezas sobre o efeito dos factores chave no desempenho das RNA's em previsão. Estes autores chegam à conclusão que as RNA's têm um desempenho satisfatório na previsão, em todos os campos. A sua adaptabilidade, a não linearidade e a sua capacidade de mapeamento de função arbitrária (*arbitrary function mapping ability*), são características únicas das RNA's que fazem com que estas sejam completamente adequadas e úteis nas tarefas de previsão. As descobertas foram inconclusivas em relação a onde e quando as RNA's são melhores do que os métodos de previsão clássicos. Um número considerável de investigadores trabalhou no sentido de tentar chegar a uma conclusão. Há vários factores que podem afectar o desempenho das RNA's. Contudo, não existe uma investigação sistemática sobre este assunto (Zhang et al., 1998).

Mas se por um lado as redes neuronais são uma grande promessa no campo da investigação em previsão, por outro lado incorporam muita incerteza.

A análise da regressão linear múltipla, apresenta grandes dificuldades em lidar com a complexidade do mercado imobiliário, especialmente devido à correlação espacial e ao desconhecimento de forma funcional (González e Formoso, 2000). Estes autores consideram que de todos os atributos, o mais importante é a localização, relacionada com a fixação espacial do produto (imobilidade). O valor de localização está relacionado com a acessibilidade (oferta e qualidade de vias e meios de transporte) e com as características da vizinhança, ou seja, do uso do solo na envolvente próxima do imóvel. Medir estes efeitos é muito difícil, pois não são quantificáveis directamente, sendo medidos através de variáveis *proxy*, tais como a distância ao centro comercial/histórico da área urbana.

Em 2001, Nguyen e Cripps, recolheram um total de 3906 observações de residencias uni-familiares vendidas, em Rutherford, através do site Multiple Listing Service for the Rutherford Country, Tennessee, para um período de 8 meses entre 1 Janeiro de 1993 e 30 Junho de 1994. Neste estudo, os autores conseguiram provar que o desempenho das RNA's é superior ao do MRLM, e dão uma explicação plausível para esse facto, enquanto que em investigações anteriores, estes resultados não eram assim tão consistentes (Nguyen e Cripps, 2001).

Um outro estudo com o mesmo objectivo, foi apresentado em 2004 por Limsombunchai, Gan e Lee. Para tal, foi usada uma amostra de 200 casas de Christchurch na Nova Zelândia, tendo sido aleatoriamente seleccionada através do site Harcourt em 2003. Os resultados também mostraram que as redes neuronais têm um melhor desempenho do que os modelos hedónicos, contudo este trabalho apresenta algumas limitações e uma grande desvantagem, que é a utilização do valor estimado da casa, e não do verdadeiro valor de mercado, isto é, os dados não correspondem a vendas efectivas. Este problema, referem os autores deve-se à grande dificuldade de obtenção dos dados reais de mercado.

O interesse em métodos não convencionais para estimação do preço do mercado imobiliário tem vindo a crescer, principalmente na última década. A maioria utiliza as Redes Neuronais, e os resultados obtidos não são consensuais, mas no entanto é notório o crescimento do interesse nestes métodos (Worzala et al., 1995; McGreal et al., 1998; Nguyen et al., 2001; Connellan et al., 1998; Bee-Hua, 2000; Lokshina, et al., 2003).

Já em 2006, o trabalho de Zurada, Levitan e Guan, apresenta os resultados do uso de dois métodos não convencionais, a lógica fuzzy e o raciocínio baseado na memória (memory-based reasoning) na avaliação dos valores dos bens imóveis residenciais, tendo sido estes modelos aplicados a um conjunto de dados reais. Este artigo também compara os resultados obtidos com estes dois métodos e com os modelos de regressão múltipla e as RNA's. Métodos de tratamento prévio de dados, como a Análise de Componentes Principais e selecção de variáveis, também foram utilizados com o objectivo de melhorar os resultados finais. Contudo, os resultados indicam que nenhum dos dois modelos foi consistentemente superior, quando aplicado a este conjunto de dados, quando comparado com os resultados obtidos com as RNA's e a Regressão Linear Múltipla.

O trabalho mais recente, conhecido até à data, para previsão do preço de venda de casas construídas, desenvolveu um novo modelo baseado na lógica fuzzy (Kusan et al., 2010). Este sistema, considera o plano da cidade, a proximidade a edifícios culturais, médicos, desportivos e de educação, os transportes públicos e outros factores ambientais, assim como a crescente tecnologia associada à construção de novos edifícios. Estes factores foram tomados como variáveis de entrada no modelo construído, segundo o objectivo do trabalho. Toda a informação foi proveniente de agências imobiliárias. Os valores da previsão e os valores de venda das casas foram comparados, tendo sido obtida uma muito boa precisão de ajustamento (Kusan et al., 2010).

Num estudo anterior também para a Turquia, Selim, comparou um modelo hedónico com um modelo de RNA's para determinar o preço da habitação na Turquia (Selim, 2009). O autor usou os dados de 2004 constantes do Household Budget Survey Data for Turkey (Instituto de Estatística Turco) num total de 5741 observações

englobando casas urbanas e rurais, contendo 46 variáveis caracterizadoras de cada habitação. As variáveis incluem o tipo de casa, a idade do edifício, o tipo de edifício, o número de quartos, a área, o sistema de aquecimento, entre outras características estruturais, e engloba também uma variável caracterizadora do local. Relembre-se que os dados são referentes a um país. No entanto factores ambientais não puderam ser considerados porque não constam da base de dados. No fim do seu estudo provou que, dada a não linearidade dos modelos de regressão hedónicos, as RNA's podem ser uma melhor alternativa de modelação dos preços das casas na Turquia.

Em 2008, Noelia García, Matías Gámez e Esteban Alfaro, introduziram um sistema automático de avaliação de imóveis, que combina o uso de redes neuronais artificias com um sistema de informação geográfico. A opção dos autores assenta no facto de que ambas as ferramentas já demonstraram a sua utilidade potencial no campo da investigação económica. Em 2002, Thurston declarou, que uma RNA ligada a um GIS pode ser usada para simular como é que o cérebro humana processa problemas de dados espaciais. Há imensas aplicações em que uma RNA acoplada ao GIS se tornou muito útil. Por exemplo, podemos mencionar, o uso do solo, a oceanografia, a floresta, o movimento dos consumidores, avaliação do ruído gerado por um aeroporto, entre outros. Thurston, conseguiu mostrar como alguns modelos de redes neuronais e um sistema de informação geográfica podem ser combinados, constituindo uma poderosa ferramenta na área da economia. García, et al, em 2008, afirmam que qualquer que seja a abordagem utilizada, a análise pode ser melhorada através da integração de um sistema de informação geográfica. No seu trabalho, estes investigadores usaram os modelos de RNA's, perceptrão multi-camada, funções de bases radiais e os mapas de Kohonen´s. Os dois primeiros modelos são uma interessante alternativa aos modelos tradicionais de regressão, enquanto que os mapas de Kohonen´s (SOM), está especialmente vocacionado para tarefas de *clustering*. Por isso os dois primeiros modelos foram usados para estimar o preço dos imóveis, enquanto que os mapas de Kohonen's foram usados para tarefas intermédias relacionadas com a estimação de valores faltantes para variadas variáveis qualitativas, tais como a qualidade da propriedade.

Neste interessante trabalho, os autores combinaram as RNA's com um sistema de informação geográfica, num sistema automático para estimação do preço de uma casa em Albacete, Espanha. Através de um simples *click*, em cima do mapa desta cidade, e após fornecer ao sistema as características da casa pretendida, o sistema devolve a estimativa para o preço da casa. Os resultados de desempenho dos modelos foram comparados, tendo-se obtido uma melhor precisão com as RNA's, na determinação da estimativa do preço total com um R^2 de 92% e um erro médio relativo de 5,56%. Os autores suspeitam que a razão para estes resultados seja a dimensão da amostra disponível (591), pequena para os requerimentos da rede de funções de base radiais. A análise de sensibilidade mostrou que a variável mais importante foi a distância ao distrito central de negócios, com um declive negativo de acordo com o pressuposto monocêntrico. Outro resultado importante do estudo de Garcia et al., de 2008, foi a não linearidade existente na relação entre o efeito da idade no preço da casa. Neste sentido, vale a pena mencionar a capacidade dos modelos neuronais em detectar relações não lineares, que não podem ser detectadas através dos modelos mais tradicionais.

Outros estudos foram feitos para Espanha, utilizando diferentes áreas urbanas. Caridad et al, 2009, usaram uma amostra com mais de 10000 transacções, recolhidas em trabalhos anteriores entre 2002 e 2006, para a cidade de Córdoba, permitindo também fazer comparações temporais. No contexto que se vive actualmente, urge encontrar formas objectivas de determinar o verdadeiro valor das propriedades. Em Espanha, os preços dos imóveis são recolhidos pelo INE e pelos municípios, com fins fiscais, não estando concentrados na verdadeira avaliação da propriedade individual. Assim, foi através de inquéritos e de entrevistas aos mediadores imobiliários que os dados para este estudo foram recolhidos. Usaram uma RNA multi-camadas para modelar o preço, com seis variáveis de entrada escolhidas através de diversas técnicas não descriminadas pelos autores. A variável explicativa do preço mais importante foi a área, seguida do índice de localização, e das despesas com o condomínio. Os anos do edifício, o índice de extras (medido através da existência de arrecadação e garagem) e o índice de qualidade (janelas, chão, mobílias e cozinha), demonstraram ter menos peso, no entanto se retiradas do modelo, o resultado final é menos bom. Caridad et al, 2009, consideram que o uso das RNA's é mais flexível do que os modelos clássicos econométricos, quando um conjunto de dados suficiente está disponível.

Mas apesar da contribuição desses modelos, os resultados não são unânimes sobre a existência de modelos capazes de predizer a variação do valor dos imóveis com o tempo. Como tal, há uma necessidade de modelos práticos e automatizados que ajudem a alcançar este importante objectivo (Khalafallah, 2008).

Resumindo, desde os anos 90 em que as Redes Neuronais Artificias se começaram a aplicar na área do imobiliário, tem-se assistido ao surgimento de variados modelos de valorização do mercado imobiliário, em

diversas regiões do planeta. É notória a procura crescente ao longo do tempo, de novos e melhores algoritmos de funcionamento das redes neuronais.

São também muitos os estudos que estabelecem uma comparação entre os sistemas de inteligência artificial e os métodos tradicionais de avaliação de imóveis, especialmente com a Regressão Múltipla. Para o efeito, é geralmente calculada a percentagem de erro de um sistema de IA e de outro de Regressão Múltipla aplicando-os a uma amostra representativa do mercado para a qual se conhece o preço de venda dos imóveis. As vantagens que os sistemas de IA trouxeram relativamente aos métodos tradicionais podem resumir-se basicamente em dois:

- Os sistemas de IA, apresentam nas provas, erros médios que se situam entre 5 e 10%, enquanto os modelos de Regressão Múltipla têm erros entre os 10 e os 15%. Há no entanto que ressalvar que em algumas experiências os resultados das experiências de ambos são semelhantes, quando se trata de amostras homogéneas (Couto, 2007).

- A segunda vantagem de um sistema de IA é a sua capacidade para estimar o valor das propriedades que apresentam características significativamente diferentes das que estão nas proximidades (valores extremos ou *outliers*), dado que este tipo de sistemas submete as amostras a processos matemáticos muito mais complexos que o modelo de Regressão Múltipla. Por outro lado, em determinados estudos, os sistemas de IA apresentam dificuldades em estimar com precisão os valores das propriedades com características especiais, *outliers*.

Os sistemas de IA já funcionam em Espanha em determinadas áreas, como por exemplo no sistema desenvolvido pela Agência Tributária para detecção da fraude e evasão fiscal, no imposto sobre o valor acrescentado, IVA.

Em Espanha, e dentro da valorização do mercado imobiliário, podemos destacar as contribuições de Caridad e Ceular (2001), García Rubio (2004), Gallego (2004) e Lara (2005), com aplicação às cidades de Córdoba, Albacete, Madrid e Jaén respectivamente. Actualmente, a Direcção Geral do Cadastro está a desenvolver um projecto para elaborar a estimação do valor de cada imóvel em preços de mercado, com o fim de perseguir a fraude imobiliária utilizando redes neuronais.

Em qualquer aplicação econométrica à realidade portuguesa, a obtenção de dados fidedignos representa parte substancial do trabalho a realizar. No caso concreto do mercado imobiliário habitacional as dificuldades encontradas são acrescidas, devido ao facto de não existir uma série temporal para os preços de "venda" da habitação, suficientemente longa, e que contemple os diferentes atributos do bem residencial (Carvalho, 1995).

Paula Couto, na sua tese de Doutoramento, faz entre outras, uma aplicação com redes neuronais a uma base de dados recolhida pelo INE com o apoio do software JavaNNS. Neste trabalho, não é apresentada qualquer interpretação da RNA, assim como da sua capacidade de estimação. No entanto, a autora deste trabalho acaba por concluir que o modelo de avaliação de apartamentos, obtido por regressão linear múltipla, corresponde a um bom modelo para avaliação em massa para Portugal continental, com vista à obtenção dos respectivos valores patrimoniais tributáveis (Couto, 2007).

3. A AMOSTRA E O AGRUPAMENTO DE VARIÁVEIS QUALITATIVAS

Castelo Branco, está situada na zona centro de Portugal, na região raiana da Beira Interior profundamente esvaziada, e com os sectores agrícolas tradicionais em crise. A cidade demonstrou, apesar disso, um dinamismo apreciável, fruto do investimento industrial e da dotação de equipamentos e serviços de âmbito regional.

Nos Censos de 2001, os números apontavam para 31 mil habitantes na Freguesia de Castelo Branco registando-se uma variação positiva de aproximadamente 15% face aos censos anteriores. Relativamente aos alojamentos, Castelo Branco registava em 2001, 16 607 alojamentos familiares clássicos e 93 de outro tipo de alojamento, dos quais 44 eram barracas. Dos 16 607 alojamentos familiares clássicos, 67% eram de residência habitual, 24% de residência secundária e apenas 2% estavam vagos para venda (INE, 2001).

A base de dados, foi constituída com a boa vontade de alguns Agentes Imobiliários da cidade, e a informação diz respeito a mais de 200 apartamentos vendidos na cidade de Castelo Branco, entre 2005 e 2009, através da intervenção destes agentes. O facto de os apartamentos terem sido efectivamente vendidos, permite-nos o uso da palavra preço, uma vez que a venda ocorreu com um determinado valor conhecido.

Aproximadamente metade dos apartamentos amostrados são novos, tendo-se registado para cada um deles, o seu verdadeiro valor de transacção, isto é, o preço, assim como um vasto conjunto de características, parte das quais se podem ver na tabela seguinte (Tabela 1).

Tabela 1. Definição das variáveis

Variável	Definição
t	1 a 5, consoante o ano de venda seja 2005 a 2009 respectivamente
Estado	0 se é novo, 1 se é usado
Área	área útil habitável (metros quadrados)
Nºasso	nº de assoalhadas
Nº WC	nº total de casas de banho
Varanda	1 se tem varanda; 0 caso contrário
Lareira	1 se tem lareira; 0 caso contrário
Ar Condicionado	1 se tem pré-instalação; 2 se tem aparelhos; 3 se é central; 0 c.c.
Aquec_central	1 se tem aquecimento central; 0 caso contrário
Janelas	2 se estão em bom estado; 1 estado regular e 0 em mau estado
Electrodomésticos	1 se tem electrodomésticos na cozinha; 0 caso contrário
Arrecadação	1 se tem arrecadação; 0 caso contrário
Garagem	1 se tem lugar estacionamento; 2 se tem garagem individual; 0 c.c.
Condomínio	Valor de prestação mensal de condomínio, em euros
Preço	Preço de venda, em euros
Andares	Nº de andares do prédio
Elevador	1 se tem elevador; 0 c.c.
Videovigilância	1 se tem sistema de vídeo-vigilância; 0 c.c.
Localização	Zona da cidade

Nota: Foram recolhidas outras variáveis como o número de casas de banho com banheira e sem banheira; armários embutidos; estores eléctricos; vidros duplos; soalho; móveis de cozinha; gás canalizado; pintura interior; hidromassagem; antiguidade; estado de conservação do prédio e rampa de acesso ao prédio.

A existência de um grande número de atributos qualitativos na base de dados, levou-nos ao agrupamento dessas variáveis, permitindo por um lado a utilização de mais informação com menos variáveis, e por outro, a possibilidade de estas variáveis qualitativas serem tratadas de forma quantitativa.

Constituíram-se cinco índices, englobando as características qualitativas amostradas, nomeadamente um índice de Conforto, de Anexos, de Conservação, um Interno e um Externo. O índice de Conforto, é formado pela existência de varanda, ar condicionado, aquecimento central, lareira, vidros duplos e estores eléctricos. O índice de Anexos é relativo à existência de arrecadação e garagem que pode ser individual, ou lugar de estacionamento. O índice de Conservação reflecte o estado geral do apartamento, bem como o estado da pintura, e o estado geral das janelas. O índice Interno reflecte a existência de armários embutidos, soalho, móveis de cozinha, electrodomésticos na cozinha e gás canalizado e o índice Externo está relacionado com as características do prédio onde se encontra o apartamento, e engloba a existência de elevador, o estado geral de conservação do prédio, a existência de rampa de acesso, e o sistema de vídeo vigilância.

Estes índices variam todos entre 0 e 1, para que sejam o mais homogéneo possível e para que todos tenham à priori a mesma importância relativa. Se o valor do índice se aproxima do um, então significa que as variáveis que o compõem se encontram no óptimo, se estiver perto de zero significa que as variáveis que o compõem se encontram em situação desfavorável.

Podemos encontrar na literatura autores que defendem o recurso a estes índices (Richardson, 1973; Saura, 1995; Jaén e Molina, 1995; Caridad e Ceular, 2001; Tabales, 2007). Contudo, a formação destes índices tem algum grau de subjectividade ao atribuir um valor numérico a um bem qualitativo. Assim, para que a constituição dos índices e dos seus valores fosse o mais assertivo possível, foram feitas várias validações, nomeadamente junto dos Agentes Imobiliários.

4. ESTIMAÇÃO DA REDE NEURONAL ARTIFICIAL

Para estimar o preço de um apartamento em Castelo Branco, através de um modelo de redes neuronais, usámos o software estatístico SPSS, v.18. As variáveis incluídas no modelo, foram anteriormente propostas e usadas para estimar o preço da habitação em Castelo Branco – Portugal, através de metodologias hedónicas. Neste artigo, os autores sugerem um modelo para o preço da habitação com as cinco variáveis, a saber a área útil, o índice de conforto, o índice de anexos, a interacção entre o t e o estado da casa e a interacção entre o índice de conservação e o estado da casa (Teixeira et. al., 2010). Para além destas variáveis, decidimos incluir no presente estudo, uma variável que quantifica a localização do imóvel, i.e. o índice de localização. Sendo um apartamento um bem imóvel, acreditamos que a sua localização afecta o seu valor (Kiel e Zabel, 2008).

4.1. MEDIDAS DE PRECISÃO E PARTIÇÃO DOS DADOS

O grau de ajuste uma rede neuronal, permite-nos avaliar a sua capacidade de generalização. O SPSS calcula duas medidas, a soma dos quadrados do erro (SSE) e o erro relativo. A melhor rede, será a que tem menor erro relativo no conjunto de teste, cuja fórmula pode ser observada a seguir:

$$\frac{\sum_{m=1}^{M}\left(y^{(m)}-\hat{y}^{(m)}\right)^2}{\sum_{m=1}^{M}\left(y^{(m)}-\overline{y}\right)^2}$$

Onde $y^{(m)}$ é o valor observado para o caso m;

$\hat{y}^{(m)}$ é o valor estimado para o caso m;

$\overline{y}$ é a média dos valores observados e,

M é o conjunto de casos do teste.

A partir do erro relativo, podemos calcular a eficiência da rede através da seguinte fórmula (Pulido-Calvo et. al, 2007):

$$E = 1 - \frac{\sum_{m=1}^{M}\left(y^{(m)}-\hat{y}^{(m)}\right)^2}{\sum_{m=1}^{M}\left(y^{(m)}-\overline{y}\right)^2}$$

Note-se que para garantir a capacidade generalizadora da rede, o conjunto de observações da amostra deve ser dividido de forma aleatória em dois subconjuntos: o de aprendizagem e o de teste. Em situações onde se dispõe de um vasto conjunto de informação que não tenha sido usado nem no conjunto de aprendizagem nem no de teste, é ainda possível fazer uma validação adicional à rede com este conjunto de dados (Costa, 2003).

O SPSS disponibiliza a opção de selecção aleatória das observações que vão constituir estes conjuntos de dados, assim como as diferentes percentagens. Por defeito o programa sugere a divisão em 70%, 30% e 0% respectivamente, para os conjuntos de aprendizagem, de teste e de validação (holdout). É importante ressalvar que as observações do conjunto de aprendizagem, não pertencem nem ao conjunto de teste, nem no conjunto de validação caso este exista. Assim, os dados utilizados no conjunto de validação, não são utilizados em nenhuma fase do processo de aprendizagem da rede, nem para actualizar pesos, nem para determinar as arquitecturas, sendo de grande importância para evitar a escolha de uma rede que sobre-ajuste os dados do conjunto de validação.

Para além disso, não há uma regra que dite qual é a percentagem adequada para os diferentes conjuntos. Há redes em que 70% dos dados é suficiente para a aprendizagem, mas há outras que precisam de uma percentagem maior para aprender. Não obstante, a literatura oferece um pequeno guia sobre a escolha das dimensões destes conjuntos; veja-se por exemplo Nguyen & Cripps (2001). Outros autores seleccionam-nos baseados nas regras empíricas de 90% vs 10%, 80% vs 20% ou 70% vs 30%, respectivamente para aprendizagem e teste, mas na maior parte dos casos, esta selecção deve ser feita com base no problema em concreto (Zhang, Patuwo & Hu, 1998). Quer a dimensão da amostra, quer o tipo de dados, vão influenciar grandemente a estimação da rede, e por isso o investigador deve experimentar diferentes percentagens para a constituição da partição dos dados, cujos resultados o guiaram na sua escolha. Esta tarefa está grandemente facilitada no SPSS, porque o utilizador

pode fazer as combinações que entender, bastando alterar as percentagens das partições e estimar uma nova rede, que poderá avaliar pelo resultado do erro devolvido para cada um desses conjuntos juntamente com a rede estimada.

Vários testes foram efectuados com diferentes partições aleatórias geradas pelo SPSS. Os resultados das estatísticas descritivas do erro relativo para os três tipos de partição utilizadas encontram-se na Tabela 2.

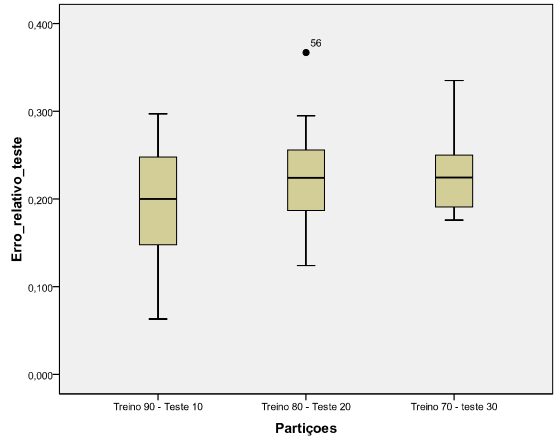
Tabela 2: Estatísticas descritivas do erro relativo do conjunto de teste, para as três partições

Descriptives

	Partiçoes			Statistic	Std. Error
Erro_relativo_teste	Treino 90 - Teste 10	Mean		,19933	,011161
		95% Confidence Interval for Mean	Lower Bound	,17651	
			Upper Bound	,22216	
		Median		,20000	
		Std. Deviation		,061133	
		Minimum		,063	
		Maximum		,297	
	Treino 80 - Teste 20	Mean		,21933	,009656
		95% Confidence Interval for Mean	Lower Bound	,19958	
			Upper Bound	,23908	
		Median		,22400	
		Std. Deviation		,052887	
		Minimum		,124	
		Maximum		,367	
	Treino 70 - Teste 30	Mean		,22910	,007967
		95% Confidence Interval for Mean	Lower Bound	,21281	
			Upper Bound	,24539	
		Median		,22450	
		Std. Deviation		,043635	
		Minimum		,176	
		Maximum		,335	

Embora o erro médio relativo do conjunto de teste para as três partições geradas não seja estatisticamente diferente para 5% de significância (ANOVA F = 2,456; Sig. = 0,092), podemos verificar que de uma maneira geral a estimação da rede neuronal para este conjunto de dados funciona melhor com um conjunto de treino (ou aprendizagem) de 90% (Gráfico 1).

Gráfico 1: Box-plot do erro relativo do conjunto de teste para as três partições

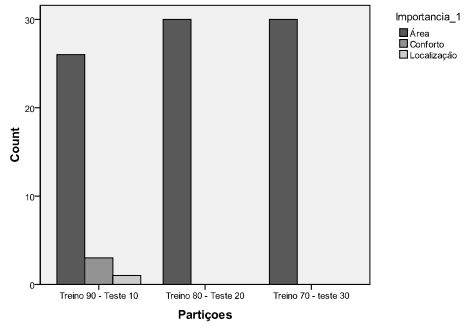


4.2. IMPORTÂNCIA RELATIVA DAS VARIÁVEIS E PARTIÇÃO DOS DADOS

Em relação à importância relativa das variáveis utilizadas na estimação da rede neuronal, para explicar o preço de um apartamento na cidade de Castelo Branco, investigámos se a variável com mais peso era dependente da partição do conjunto de dados utilizada.

Chegamos à conclusão que independentemente do tipo de partição usada na estimação, a variável com mais importância na determinação do preço, é a área útil do apartamento, medida em metros quadrados (Qui-quadrado = 8,372; sig. (2-sided) = 0,079)[1]. Este resultado pode ser comprovado graficamente no seguinte gráfico de barras para as três partições diferentes usadas (Gráfico 2).

Gráfico 2: Gráfico de barras para a variável com mais importância na estimação do preço, nas três partições



Já em relação à segunda variável com mais importância nos modelos estimados, observamos a variável índice de Conforto predomina em todas as partições, embora no caso das partições 80% vs 20% e 70% vs 30%, esta seja seguida das variáveis t x Estado e Localização, e t x Estado, respectivamente, enquanto no caso da partição 90% vs 10%, se observa que as variáveis Área útil e Índice de Anexos figuram também como segunda variável mais explicativa (Qui-quadrado = 16,894; sig. (2-sided) = 0,031)[2].

Gráfico 3: Gráfico de barras para a segunda variável com mais importância na estimação do preço, nas três partições

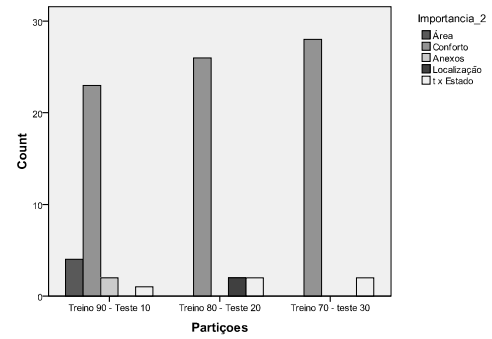
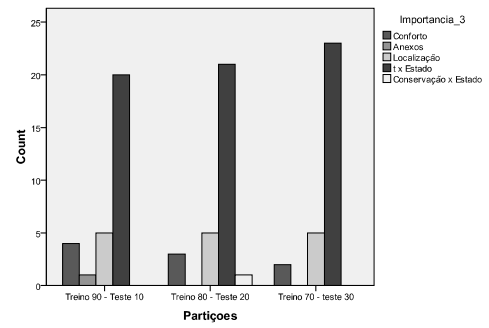


Gráfico 4: Gráfico de barras para a terceira variável com mais importância na estimação do preço, nas três partições



[1] Veja-se o Anexo I.

[2] Veja-se o Anexo II.

No caso da terceira variável mais representativa, independentemente da partição usada, surge a variável interacção t x Estado, seguida da variável Localização e do Índice de Conforto (Qui-Quadrado = 4,885; sig. (2-sided) = 0,770)[3].

5. CONCLUSÕES

Não há uma regra que dite como deve ser feita a partição dos dados para estimar uma rede neuronal. A melhor rede será obtida mediante uma vasta experimentação, tarefa grandemente facilitada com software apropriado do tipo do SPSS.

A partição que funcionou melhor neste conjunto de dados, foi aquela em que o conjunto de treino/aprendizagem tinha mais observações, isto é, 90%. De todas as redes neuronais estimadas com este tipo de partição, obteve-se uma com um erro relativo de apenas 0,063 (Tabela 3). Segundo Pulido e Calvo esta rede tem uma eficiência de aproximadamente 93,7%. Trata-se de uma rede com três neurónios na camada intermédia como se pode ver no gráfico seguinte (Gráfico 5).

Tabela 3: Medidas de erro da RNA estimada no SPSS

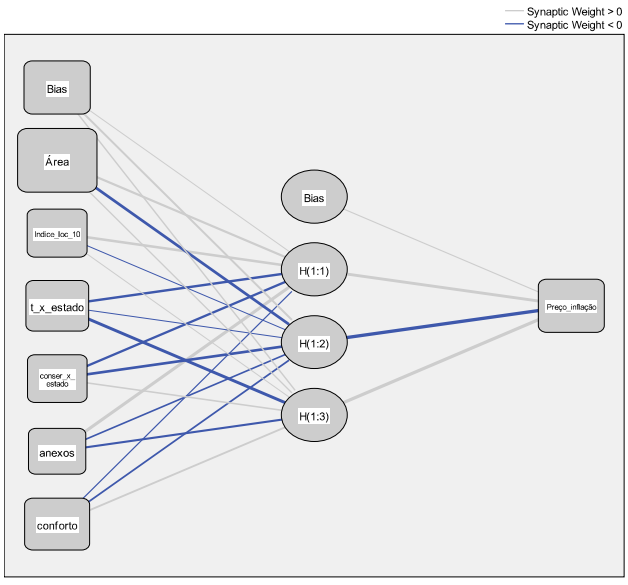
Model Summary

Training	Sum of Squares Error	27,345
	Relative Error	,275
	Stopping Rule Used	1 consecutive step (s) with no decrease in error[a]
	Training Time	0:00:00.141
Testing	Sum of Squares Error	,605
	Relative Error	,063

Dependent Variable: Preço_inflação

a. Error computations are based on the testing sample.

Gráfico 5: Desenho da Rede Neuronal Artificial estimada



Hidden layer activation function: Hyperbolic tangent

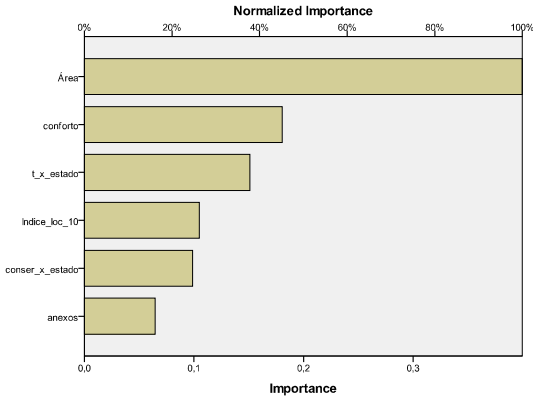
Output layer activation function: Identity

Destacamos que nesta rede, a ordem de importância das variáveis explicativas já foi analisada anteriormente, podendo ser observada no Gráfico 6.

[3] Veja-se o Anexo III.

A variável com mais peso na explicação do preço é a área útil (medida em m^2), seguindo-se o índice de conforto e a interacção entre t e o estado da casa. Observe-se que nesta interacção, a variação do preço ocorre apenas para os apartamentos usados, que entre 2005 e 2009, tiveram um decréscimo acentuado no preço de venda.

Gráfico 6: Importância das variáveis usadas na estimação do preço do apartamento com uma RNA



Independent Variable Importance

	Importance	Normalized Importance
Área útil (metros quadrados)	,400	100,0%
I Localização	,105	26,3%
t x Estado	,151	37,8%
I Conserv x Estado	,099	24,7%
I Anexos	,065	16,2%
I Conforto	,181	45,2%

Em relação às restantes variáveis utilizadas na estimação do preço do apartamento, através da rede neuronal, seguem-se por ordem de importância a localização do apartamento na cidade, o seu estado de conservação no caso de se tratar de um apartamento usado, e por fim, e por fim o índice de anexos que contempla a existência de garagem (individual ou box) e arrecadação.

Julgamos que este grupo de seis variáveis, pela quantidade de informação que englobam, pode ser usado na estimação do preço da habitação em outras cidades do país.

Consideramos que as redes neuronais artificiais constituem um bom método de estimação do preço da habitação em Portugal, e por isso devem ser atendidos como uma séria alternativa às metodologias hedónicas tradicionais.

6. Bibliografia

Bee-Hua, B. (2000): "Evaluating the performance of combining neural networks and genetic algorithms to forecast construction demand: the case of the Singapure residential sector", *Construction Management and Economics*, nº 2, p.209-218.

Borst, R. A. (1991): "Artificial neural networks: The next modellin/calibration technology for the assessment community?", *Property Tax Jiurnal*, IAAO, nº1, p.69-94.

Borst, R.A. e McCluskey (1996): "The Role of Artificial Neural Networks in the Mass Appraisal of Real Estate", *Paper presented to the Third European Real Estate Society Conference*, Belfast, June 26-28.

Borst, R.A. (1995): "Artificial neural networks in mass appraisal", *Journal of Property Tax Assessment & Administration*, nº2, p.5-15.

Caridad y Ocerin, J.M., Núñez Tabales, J., Ceular Villamandos, N. e Millán Vázquez, G. (2009): *27th International Conference. Mathematical Methods in Economics 2009*, Czech Republic, Praga.

Caridad, J. M. e Ceular, N. (2001): "Un análisis del mercado de la vivienda a través de Sistemas de Redes Neuronales", *Revista de Estudios de Economia Aplicada*, nº18, p.67-81.

Carvalho, P. F. G. (1995): *O mercado de habitação em Portugal, Tese de Mestrado, Universidade de Coimbra. Faculdade de Economia, Portugal.*

Connellan, O. e H. James (1998): "Estimated realization price by neural networks: forecasting commercial property values", J*ournal of Property Valuation & Investiment*, nº1, p.71-86.

Couto, P. (2007): *Avaliação Patrimonial de Imóveis para Habitação, Tese de Doutoramento, Laboratório Nacional de Engenharia Civil – Lisboa, Portugal.*

Do, A.Q. e Grudnitski, G. (1993): "A neural network analysis of the effect of age on housing values", *J. Real Estate Res*., nº2, p.253-264.

Evans A., James H. e Collins A. (1993): "Artificial Neural Networks: an Application to Residential Valuation in the UK", *Journal of Property Valuation & Investment*, nº11, p.195-204.

Gallego Mora-Esperanza, J. (2004): "La inteligencia artificial aplicada a la valoración de inmuebles. Un ejemplo para valorar Madrid", *Revista CT/Catastro*, nº50, p.51-67.

García Rubio, N. (2004): *Desarrollo y aplicación de redes neuronales artificiales al mercado inmobiliario: aplicación a la ciudad de Albacete, Tesis Doctoral, Universidad de Castilla – La Mancha.*

González, M. A. S. e Formoso, C. T. (2000): "Análise conceitual das dificuldades na determinação de modelos de formação de preços através de análise de regressão", *Centro de Engenharia Civil da Universidade do Minho, Revista de Engenharia Civil*, nº8, p.65-75.

Zhang G., Patuwo B. E. e Hu M. Y. (1998): "Forecasting with artificial neural networks: The state of the art", *International Journal of Forecasting*, nº14, p.35–62.
INE (2001): *Censos 1991 e 2001, resultados definitivos, Instituto Nacional de Estatística de Portugal.*
Jaén, M. e Molina, A. (1995): *Modelos econométricos de tenencia y demanda de vivienda, Editorial Universidad de Almería.*
Kershaw, P. e Rossini, P. (1999): "Using Neural Networks to Estimate Constant Quality House Price Indices", F*ifth Annual Pacific-Rim Real Estate Society Conference*, Kuala Lumpur, Malaysia.
Khalafallah, A. (2008): "Neural Network Based Model for Predicting Housing Market Performance", *Tsinghua Science and Technology*, Number S1, p.325-328.
Kusan H., Aytekin O. e Ozdemir, Í. (2010): "The use of fuzzy logic in predicting house selling price", *Expert Systems with Applications*, nº37, p.1808-1813.
Lara Cabeza, J. (2005): "Aplicación de las redes neuronales artificiales al campo de la valoración inmobiliária", *Revista Mapping*, nº 104, p.64-71.
Limsombunchai, V., Gan C. e Lee, M. (2004): "House Price Predication: Hedonic Model vs. Artificial Neural Network", *American Journal of Applied Science*, nº3, p.193-201.
Lokshina, Hammerslag, e Insinga (2003): "Applications of artificial intelligence methods for real estate valuation and decision suppor", *In Hawaii international conference on business*, Honolulu.
McCluskey, W. e Borst, R. (1997): "An evaluation of MRA, comparable sale analysis, and ANNs for the mass appraisal of residential properties in North Ireland", *Assessement Journal*, nº1, p.47-55.
McGreal, S., A. Adair, D. McBurney, e D. Patterson (1998): "Neural Networks: the predicition of residential values", *Journal of Property Valuation & Investment*, nº1, p.57-70.
Nguyen, N. e Cripps, Al. (2001): "Predicting housing value: a comparison of multiple regression analysis and artificial neural networks", *The Journal of Real Estate Research*, nº3, p.313-336.
García N., Gámez M. e Alfaro E. (2008): "ANN+GIS: An automated system for property valuation", *Neurocomputing*, nº71, p.733-742.
Pulido-Calvo, I., Montesinos, P., Roldán, J. e Ruiz-Navarro, F. (2007): "Linear regressions and neural approaches to water demand forecasting in irrigation districts with telemetry systems", *Biosystems Engineering*, nº 97, p.283 – 293.
Richardson, H. W. (1973): *Teoria de la localización, estructuras urbanas y crecimiento regional, Economía Regional, 1a. ed. Vicens-Vives, España.*
Rossini, P.A. (1997): "Artificial Neural Networks versus Multiple Regression in the Valuation of Residential Property", *Australian Land Economics Review*, nº1.
Saura, P. (1995): "Demanda de características de la vivienda en Murcia", *Secretariado de Publicaciones de la Universidad de Murcia*, 1ª ed.
Selim, H. (2009): "Determinants of house prices in Turkey: Hedonic regression versus artificial neural network", *Expert Systems with Applications*, nº36, p.2843-2852.
Tabales, J. M. N. (2007): *Mercados inmobiliarios: Modelización de los Precios, Tesis Doctoral, Departamento de Estadística, Econometria, I. O. y Organización de Empresas – Universidad de Córdoba, Espanha.*
Teixeira, M.C.C., Villamandos, N. C. y Caridad, J.M. (2010). Factores Formadores do Preço da Habitação em Portugal. VIII Colóquio Ibérico de Estudios Rurales: del desarrollo local al desarrollo territorial. Cáceres 21 e 22 de Outubro de 2010, Complejo Cultural San Francisco.
Thurston, J. (2002): "GIS & Artificial Neural Networks: Does your GIS Think?", *GISVision Magazine*.
Worzala, E., M. Lenk, and A. Silva (1995): "An exploration of neural networks and its application to real estate valuation", *The Journal of Real Estate Research*, nº2, p.185-201.
Zurada, J. M., Levitan, A. S. e Guan, J. (2006): "Non-Conventional Approaches to Property Value Assessment", *Journal of Applied Business Research*, nº3.

ANEXO I.

Tabela 1.1. Teste de independência do Qui-quadrado para a variável com mais importância na estimação do preço, nas três partições

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	8,372[a]	4	,079
Likelihood Ratio	9,167	4	,057
Linear-by-Linear Association	4,603	1	,032
N of Valid Cases	90		

a. 6 cells (66,7%) have expected count less than 5. The minimum expected count is ,33.

Fonte: Output do SPSS 17

ANEXO II.

Tabela 2.1. Teste de independência do Qui-quadrado para a segunda variável com mais importância na estimação do preço, nas três partições

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	16,894[a]	8	,031
Likelihood Ratio	18,512	8	,018
Linear-by-Linear Association	,665	1	,415
N of Valid Cases	90		

a. 12 cells (80,0%) have expected count less than 5. The minimum expected count is ,67.

Fonte: Output do SPSS 17

ANEXO III.

Tabela 3.1. Teste de independência do Qui-quadrado para a terceira variável com mais importância na estimação do preço, nas três partições

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	4,885[a]	8	,770
Likelihood Ratio	5,291	8	,726
Linear-by-Linear Association	1,180	1	,277
N of Valid Cases	90		

a. 9 cells (60,0%) have expected count less than 5. The minimum expected count is ,33.

Fonte: Output do SPSS 17

Agradecimentos

Aos responsáveis das Agências Imobiliárias GRADUZ, LING, IMOFACTOR e SGH de Castelo Branco, os nossos sinceros agradecimentos pela disponibilização de toda a informação fornecida, que tornou este estudo possível.