SISTEMÁTICA PARA PRECIFICAÇÃO DE IMÓVEIS ATRAVÉS DE REGRESSÃO LINEAR MÚLTIPLA

Guilherme Nardino Enck (UFRGS) – guilherme.enck@gmail.com

Michel José Anzanello (UFRGS) – anzanello@producao.ufrgs.br

Resumo

A ausência de metodologias científicas de suporte à tomada de decisão impõe ineficiências à indústria imobiliária que comprometem a competitividade dos players do setor. Os processos de precificação de apartamentos seguem critérios empíricos, e consistem em um potencial foco de ineficiência. A regressão linear múltipla é uma metodologia estatística de previsão de valores que permite explicar a relação entre o comportamento de uma variável de resposta Y e de variáveis explicativas X. Este artigo propõe uma sistemática para precificação de ativos imobiliários a partir da análise de um conjunto de dados de preços e características de imóveis da cidade de Porto Alegre, visando a contribuir para a adoção de métodos científicos no setor. Esta sistemática consiste na aplicação de análise de regressão juntamente de um método de seleção de variáveis. Para o conjunto de dados testados, a maior precisão preditiva foi gerada com um modelo contendo apenas 5 das 10 variáveis inicialmente testadas.

Palavras-chave: imóveis, precificação imobiliária, análise de regressão, modelo hedônico, seleção de variáveis.

1 Introdução

A indústria da construção civil possui um impacto relevante no desenvolvimento das cidades, na densificação urbana e no progresso econômico e social. A demanda reprimida por moradias (o déficit residencial brasileiro, de acordo com os resultados preliminares do último levantamento da Fundação João Pinheiro, corresponde a 5,846 milhões de residências (VIANA; SANTOS, 2015)) favorece a venda dos imóveis produzidos e torna praticamente inexistente um cenário de estocagem de imóveis. Este fato, somado ao crédito barato e farto

que caracterizou o setor nos últimos anos, gerou uma pressão compradora que resultou em lucros substanciais aos investidores da indústria de incorporação imobiliária.

Este paradigma de lucros facilitados permitiu que as companhias incorressem em ineficiências operacionais e financeiras na execução dos projetos, fato que afeta não apenas a lucratividade da empresa, mas também a geração de riqueza da economia como um todo, visto que o custo das ineficiências é repassado ao consumidor final. Neste contexto, uma das maiores fontes geradoras de ineficiências é a utilização de métodos empíricos de suporte à tomada de decisão, em detrimento a procedimentos científicos (AZMI et al., 2015). As decisões acerca do melhor local de construção de um empreendimento imobiliário, as características do empreendimento e a precificação das unidades - entre outros fatores - são comumente tomadas com base na experiência e intuição do incorporador, não garantindo a otimização na alocação de recursos e a consequente maximização dos lucros. Além das perdas financeiras, as companhias ainda estão sujeitas a danos à imagem e reputação no caso de procederem com a construção de um empreendimento mal sucedido comercialmente.

Dentro deste contexto, a implementação de técnicas e ferramentas que gerem uma predição acurada do valor de um empreendimento podem representar significativa mudança de paradigma para as incorporadoras, investidores da indústria envolvida e até mesmo para políticas públicas (KOSCHINSKY et al., 2014). Por se tratar de um setor intensivo em capital, um incremento marginalmente baixo na acurácia preditiva já justifica estudos nesta área.

Este artigo apresenta uma sistemática para precificação de imóveis através de um modelo de regressão múltipla para a construção civil. Objetiva-se (i) identificar as variáveis que impactam no preço de unidades imobiliárias e quantificar a sua influência, (ii) selecionar as variáveis independentes que melhor explicam a variável de resposta e (iii) gerar uma equação que permita prever o valor comercial de um empreendimento imobiliário baseado nas suas características de construção. Para tanto, atributos e dimensões (também chamadas de variáveis independentes da regressão) que impactam no valor de venda de um imóvel (variável dependente da regressão) serão coletados e analisados em termos de consistência e relevância para o estudo. Na sequência, modelos de regressão relacionando as variáveis independentes e dependentes serão gerados. Através da aplicação do método de seleção de variáveis *leave one variable out at a time*, será identificado o conjunto de variáveis que apresentam o menor erro de predição da variável dependente.

2 Referencial Teórico

A utilização de métodos estatísticos para embasar a precificação imobiliária vem sendo utilizada por várias décadas. O cálculo de um índice de preços para imóveis a partir da combinação dos preços de repetidas vendas de propriedades, proposta deste artigo, pode ser realizado através de técnicas de regressão (BAILEY et al., 1963). Esta seção traz os fundamentos de regressão múltipla linear, seguido por uma revisão das principais variáveis que impactam no contexto imobiliário (e que poderão fazer parte do modelo de regressão).

2.1 Regressão múltipla linear

Análise de regressão é uma técnica que permite criar um modelo que explique a relação entre variáveis baseado em dados de padrão de comportamento (MONTGOMERY et al., 2012; MONSON, 2009). A técnica possui três objetivos: descrever, controlar e prever (KUTNER et al., 2005). Para executar estes três objetivos, é estudado o impacto de diversos fatores no fenômeno que se pretende prever, e tal estudo é feito através de estimativas e testes a partir de dados experimentais ou observacionais. Para determinar este impacto, Montgomery et al. (2012) comenta a utilidade de variadas ferramentas de regressão, enquanto Seltman (2015) cita o método dos mínimos quadráticos como uma dessas ferramentas, destacando a simplicidade e objetividade do método. A partir dessa análise, são estimados os parâmetros das diferentes variáveis preditoras, e é elaborado um modelo de regressão.

Assim sendo, pode-se definir o modelo de regressão como uma representação matemática da tendência de uma variável de resposta em relação a variáveis preditoras (KUTNER et al., 2005). Uma vez que o modelo de regressão é constituído, é possível realizar estimativas do comportamento de uma variável de resposta em diferentes contextos. Problemas de regressão começam com a determinação de potenciais variáveis preditoras, podendo estas serem medidas contínuas (tal como altura de um objeto ou distância), dados discretos ou categóricos (como, por exemplo, cor do olho ou grupo populacional) (WEISBERG, 2005). Essas variáveis serão os termos X da equação de regressão, conforme apresentado na Equação (1).

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_n X_{in} + \epsilon_i \qquad (1)$$

O objetivo desta equação é descrever como a variável dependente Y varia em função das variáveis independentes X. Os parâmetros β descrevem a intensidade com que cada variável independente influencia a variável dependente Y, e serão estimados a partir de dados (RAWLINGS et al., 1998; KUTNER et al., 2005); ϵ_i representa o erro do modelo. Tal estimativa pode ser realizada a partir do método *Least Squares Estimation,* que busca encontrar o melhor ajuste para um conjunto de dados através da menor soma possível dos quadrados da diferença entre valores observados e os valores estimados para a variável de resposta Y (RAWLINGS et al., 1998; WEISBERG, 2005; KUTNER et al., 2005; MONTGOMERY et al., 2012). Para a análise de regressão múltipla, são utilizados modelos matriciais, assim permitindo o cálculo de extensos sistemas de equações (KUTNER et al., 2005; RAWLINGS et al., 1998).

2.2 Métodos de seleção de variáveis

Identificar quais variáveis preditoras são relevantes e quais são irrelevantes para um modelo é uma solução para trabalhar com problemas com muitos potenciais preditores (WEISBERG, 2005; MEHMOOD et al., 2012). A acurácia de estimativas e predições sempre tenderá a ser maior em modelos que contenham apenas variáveis relevantes (WEISBERG, 2005). Existem duas abordagens principais para a seleção de variáveis: (i) seleção com base na amostra coletada, sendo todas as predições futuras baseadas nas mesmas variáveis; (ii) executar um processo de seleção para cada amostra individual, de forma que cada amostra possa apresentar diferentes variáveis independentes para se realizar uma predição da variável dependente (THOMPSON, 1978). Esta segunda abordagem tende a ser a ótima (THOMPSON, 1978).

Dentre os métodos de seleção de variáveis alguns se destacam. O critério S_p seleciona as variáveis a partir da comparação entre o valor estimado e o valor observado minimizando o erro quadrático médio de predição (MSEP) (THOMPSON, 1978). A seleção através da abordagem Bayesiana - conforme descrito por O'Hara & Sillanpää (2009), nas bases de Lindley & Smith (1972), tendo sua implementação prática ilustrada por Chipman et al. (2001) - e suas variações são largamente utilizadas. Procedimentos sequenciais também são muito frequentemente aplicados. Entre eles estão os métodos de *forward selection,* através do qual variáveis são paulatinamente introduzidas na equação até que um ajuste satisfatório seja obitido; e *backward elimination,* o qual funciona de forma contrária ao anterior, sendo o número de variáveis no modelo reduzido desde o seu conjunto total até atingir uma variável

única. Em ambos os testes, a entrada ou saída de uma nova variável é determinada a partir de teste F (THOMPSON, 1978). Também podem ser adotados métodos multicriteriais, conforme aplicado por Anzanello et al. (2011) em estudo para a seleção e classificação de variáveis de lotes de produtos. O método conhecido como Jackknife é um processo iterativo através do qual é feita a estimativa do parâmetro a partir da amostra completa (composto por k variáveis) e então cada variável é retirada do modelo, uma de cada vez, recalculando-se o parâmetro, para produzir uma estimativa do valor do R^2 ajustado. (ADBI & WILLIAMS, 2010). De metodologia similar, a técnica *leave one variable out at a time,* conforme utilizado por Anzanello & Fogliatto (2011), consiste em (i) Testar a acurácia preditiva de um modelo com k variáveis; (ii) testar a sua acurácia com k-1 variáveis, retirando uma variável por vez; (iii) a partir do conjunto de variáveis que apresentaram a melhor acurácia, retirar uma por vez (restando k-2 variáveis) e avaliar os resultados; (iv) repetir o procedimento até encontrar o conjunto de variáveis independentes que maximizam o potencial de predição.

A performance de um modelo pode ser avaliada pela expectativa do seu erro, sendo este estimado a partir da utilização de um banco de teste independente do banco de treino (RIVALS & PERSONNAZ, 2003). Uma variação interessante da análise de sensibilidade é a utilização do erro obtido pela validação cruzada leave-one-out, substituindo a função objetivo por esta métrica (GUYON & ELISSEEFF, 2003).

2.3 Características do setor imobiliário e a aplicação de ferramentas de regressão

Para o setor imobiliário e, mais especificamente, para a precificação de imóveis, a ferramenta da regressão linear múltipla é aplicada através de modelos conhecidos como preços hedônicos. Os modelos de precificação hedônica são utilizados para medir a influência de cada característica de um imóvel no seu preço de transação a partir dos coeficientes gerados pela análise de regressão, bem como para criar uma ferramenta de previsão de preços de mercado (MONSON, 2009). Estes modelos comumente seguem os princípios microeconômicos estabelecidos por Rosen (1974), que serviram de base para os estudos posteriores. Cebula (2009) encontrou, através do uso destes modelos, uma série de atributos físicos que impactam positivamente o valor do imóvel, tais como número de lareiras, de andares e inclusive a presença de um sistema de *sprinkler* subterrâneo. Inúmeras outras variáveis de influência foram encontradas por outros autores, conforme demonstrado a seguir.

Kaplanski e Levy (2011) demonstra como os preços dos imóveis são influenciados por efeitos de sazonalidade de forma persistente, e comenta que os preços podem ser até 3,75%

maiores no verão. Os resultados dos estudos de Miller et al. (2012) também revelam que os preços dos imóveis atingem o seu pico durante os meses de verão e o seu menor valor no inverno. Saiz (2010) pesquisou o efeito econômico da geografia e de outras restrições de construção na oferta de bens imóveis. Os seus estudos evidenciam que a incorporação imobiliária é reduzida em locais onde os terrenos possuem alta inclinação, e que na maioria das áreas nas quais a oferta de imóveis é considerada inelástica, há a presença de fortes restrições geográficas. Segundo Saiz (2010), cidades que apresentam maiores restrições geográficas tendem a apresentar maiores preços dos imóveis (*ceteris paribus*).

Alguns autores também apontaram a influência do setor público no preço dos imóveis. Para Schill (2005), o excesso de regulação estatal no setor da construção contribui para o aumento dos custos de construção e, por conseguinte, dos preços praticados. Saiz (2010) concorda, afirmando que restrições de crescimento urbano de natureza regulatória possuem o efeito de reduzir a elasticidade da oferta de casas e apartamentos. Shi et al. (2015), por sua vez, estudaram os efeitos da manipulação do sistema de preços através de subsídios governamentais, demonstrado como estes também impactam o comportamento de precificação de forma inversa – quanto maiores os subsídios, menor os preços. Baseado em estudo realizado em Beijing, China, a partir do uso de um modelo de preço hedônico, Shi et al. (2015) também relatam que entre os fatores que impactam o preço de unidades imobiliárias novas estão a volatilidade do mercado (com relação direta), características físicas dos apartamentos, a condição financeira da incorporadora (com relação inversa) e fatores conjunturais de mercado.

Alves et al. (2011) também citam a grande importância relativa de variáveis de mercado e de crédito na precificação dos imóveis, como por exemplo o spread dos juros bancários. A partir da aplicação de modelos hedônicos sobre uma amostra de 1860 apartamentos, Fávero et al. (2008) demonstram que, à medida que aumenta o perfil sócio-demográfico, a importância relativa da variável área do imóvel também aumenta, e que para o perfil médio cresce a importância dada às áreas de lazer do empreendimento. Os autores ainda citam a proximidade com lojas, supermercados, shopping centers, parques, hospitais, o número de vagas na garagem, de quartos e de banheiros como fatores com peso importante no processo de *valuation* das unidades imobiliárias (FÁVERO et al., 2008). Através da utilização de ferramentas de regressão linear, Guo e Wu (2013) demonstram que a taxa de aluguel e o crescimento do Produto Interno Bruto foram as únicas variáveis que impactaram de forma significativa o preço de imóveis em Xangai. Já os estudos de Xu e Chen (2012) indicam que

taxas de juro em queda, aumento da oferta monetária e redução do pagamento antecipado da hipoteca aceleram a escalada de preços no setor imobiliário chinês.

Segundo Sirmans et al. (2006), as características imobiliárias encontradas com maior frequência em modelos de preços hedônicos são metragem, tamanho do lote, idade do imóvel, número de quartos, número de banheiros, existência de garagem e número de vagas, existência de piscina, de lareira e de ar condicionado. Sirmans e Macpherson (2005) concordam que tais 9 itens são os mais frequentemente mencionados, e ainda citam que, em alguns modelos hedônicos, a metragem pode ter impacto negativo e a idade do imóvel pode ter impacto positivo nos preços, ao contrário do que se infere intuitivamente. Com efeito, algumas das variáveis citadas nestes dois estudos também foram utilizadas por Cortright (2009): tamanho, número de quartos, número de banheiros e idade do imóvel. Os resultados encontrados por Owusu-Ansah (2012) reforçam ainda mais estas variáveis: o autor destaca o efeito significativo das variáveis 'número de quartos', 'número de banheiros', 'idade do imóvel', 'vagas de garagem' além de itens de infraestrutura condominial, como piscina e jardins. Para a variável de localização, os modelos hedônicos normalmente utilizam o conceito de distância ao CBD, que consiste na distância entre imóvel e a região onde está concentrado os negócios e o comércio da cidade; He et al. (2010) inclui esta variável ao seu modelo. Ottensmann et al. (2008) propõem uma revisão da variável de localização do método dos preços hedônicos, e demonstram como a utilização de múltiplos centros de emprego, em detrimento da tradicional premissa de um centro de emprego por cidade, incrementam a performance do modelo. Por fim, Bover e Velilla (2002) propõem um método alternativo aos tradicionais modelos hedônicos, que considera, além das características internas do imóvel, características externas não observáveis, tais como localização, transporte, tráfego e proximidade com serviços.

As variáveis elencadas nesta seção de revisão (variáveis independentes) terão sua contribuição para estimativa do preço de imóveis (variável dependente) avaliada nos modelos de regressão a serem gerados na seção 3 deste artigo.

3 Metodologia

Esta seção apresenta os procedimentos metodológicos adotados para a condução do trabalho. Será caracterizado o método de pesquisa quanto à sua natureza, abordagem,

objetivos e procedimentos utilizados. O método de trabalho será especificado, seguido então pelas etapas de operacionalização do método.

3.1 Método de pesquisa

Este trabalho utiliza técnicas de regressão linear aplicadas ao contexto da precificação imobiliária, e é classificada como de natureza aplicada tendo em vista que possui, como objetivo final, a aplicação prática (SILVA & MENEZES, 2005) de um modelo estatístico. É utilizada uma abordagem quantitativa, visto que o modelo de regressão será obtido através da análise de dados de preço e atributos físicos numéricos de unidades imobiliárias, utilizando assim de quantificação tanto no levantamento dos dados quanto na aplicação destes em modelos estatísticos (BOAVENTURA, 2004). Uma pesquisa descritiva visa, dentre os seus objetivos, o estabelecimento de relação entre variáveis (GIL, 2002). Por pretender descrever a correlação dos atributos de uma unidade imobiliária com a variável dependente (preço do imóvel), os objetivos do estudo são classificados como descritivos. Quanto aos procedimentos de pesquisa, este trabalho é definido como experimental, uma vez que, diante da determinação de um objeto de estudo e da seleção das variáveis que podem influenciá-lo, são observados os efeitos dessas variáveis no objeto (GIL, 2002).

3.2 Método de trabalho

As etapas que compõem a execução deste trabalho estão explicitadas na Figura 1.

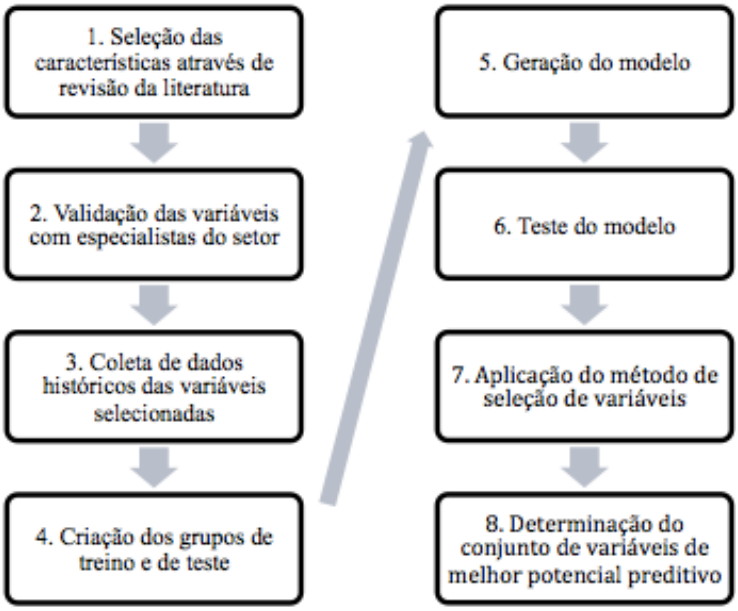

Figura 1 – Fluxograma das etapas para execução do estudo

Foi feito um levantamento dos estudos da literatura que focaram na geração de modelos de precificação hedônica para determinar as variáveis mais utilizadas e cujos resultados tenham demonstrado maior impacto nos modelos. Tais variáveis dizem respeito a aspectos de localização geográfica dos imóveis, características da construção e aspectos de vizinhança, dentre outras.

Após o levantamento das variáveis de maior potencial preditivo baseado em revisão da literatura existente sobre o tema, foram realizadas reuniões presenciais individuais com três especialistas do setor da construção civil: um incorporador, um arquiteto projetista e urbanista e um diretor de uma imobiliária. Nestas reuniões, foram apresentadas as variáveis retiradas da literatura, as quais foram arguidas em termos da sua pertinência no modelo de regressão a ser gerado. O intuito desta etapa era garantir que as variáveis escolhidas para o modelo - definidas em caráter preliminar com base em estudos realizados sob vários contextos culturais e geográficos - também eram aplicáveis à realidade da região deste estudo. Somente as variáveis consideradas relevantes pelos especialistas foram incorporadas ao modelo de regressão.

Para cada variável do modelo foram coletados dados de apartamentos. Cada variável estava relacionada a uma característica física do imóvel, além da variável dependente do modelo (preço de venda anunciado). Para garantir a diversidade de características, o único pré-requisito para seleção dos imóveis era que estivesse localizado na região urbana da cidade de Porto Alegre. Esta coleta foi realizada através das informações contidas nas páginas das 3 maiores imobiliárias da cidade na internet. Os dados foram compilados da forma ilustrada no Quadro 1.

Observação	Preço do imóvel	Variável 1	Variável 2	...	Variável *k*
Imóvel 1	-	-	-	-	-
Imóvel 2	-	-	-	-	-
...	-	-	-	-	-
Imóvel *n*	-	-	-	-	-

Quadro 1 – Estrutura do banco de dados

Para avaliar a capacidade preditiva do modelo, os dados foram divididos em dois grupos: grupo de treino e grupo de teste. O primeiro foi utilizado para a determinação dos coeficientes da regressão (construção do modelo), enquanto os dados do segundo grupo foram utilizados para validar a acurácia do modelo em prever a variável dependente de observações que não estavam presentes no banco no momento da criação do modelo.

Gerou-se então o modelo de regressão múltipla linear utilizando as observações do grupo de treino através de aplicativo estatístico. Nesta etapa, objetivou-se determinar os coeficientes do modelo de regressão para todas as *k* variáveis contidas no Quadro 1. Neste procedimento, não foram consideradas interações entre as variáveis independentes, bem como eliminou-se o termo independente da regressão.

O modelo de regressão gerado na etapa anterior foi aplicado ao conjunto de dados do grupo de teste, visando a avaliar a acurácia preditiva do modelo. A estimativa do preço de cada imóvel gerada pelo modelo foi comparada com o preço real, e foi calculado o erro absoluto percentual $100 * (\frac{|y-\hat{y}|}{y})$ para cada observação. Calculou-se então, o erro absoluto percentual médio (MAPE) da aplicação a partir da média aritmética do erro absoluto percentual das observações do banco de teste, conforme a fórmula $\frac{\sum_i^n 100*(\frac{|y-\hat{y}|}{y})}{n}$.

Por fim, procedeu-se à seleção de variáveis do tipo *leave one out at a time*, conforme Anzanello & Fogliatto (2011); utilizou-se como métrica comparativa o MAPE de cada conjunto de variáveis. A partir das variáveis iniciais, foi omitida uma variável por vez, gerando novos coeficientes para cada modelo composto pelas *k*-1 variáveis remanescentes. Os coeficientes gerados por cada modelo foram aplicados ao banco de teste, e comparou-se o valor previsto pelos coeficientes com o valor real de cada observação; a partir desta comparação, foi calculado o MAPE gerado por cada rodada. Após a determinação das *k*-1 variáveis que minimizavam esta métrica (ou seja, identificação da variável que quando omitida conduziu ao menor MAPE), o procedimento foi repetido, retirando-se uma variável de cada vez, até encontrar o conjunto de *k*-2 variáveis que conduz ao menor MAPE. Esta sequência foi levada à cabo até identificar o conjunto de variáveis independentes que, quando utilizadas no modelo de regressão, apresentavam o menor MAPE e, portanto, a melhor capacidade preditiva.

4 Resultados

Esta seção detalha a aplicação da sistemática descrita na seção anterior e os resultados encontrados pelo modelo de regressão.

A primeira etapa da sistemática proposta consistiu na definição das variáveis independentes do modelo de regressão. Para tanto, fez-se uso da pesquisa já realizada na seção 2 do presente artigo. Com base nas recomendações de Sirmans & MacPherson (2005), Sirmans et al. (2006), Cortight (2009), Owusu-Ansah (2012) e Ottensmann et al. (2008) foram listadas as potenciais variáveis independentes do modelo de regressão no Quadro 2.

#	Variável	Descrição
1	Tamanho do imóvel	Área privativa do imóvel ($m^{2)}$
2	Tamanho do lote	Tamanho do terreno onde o imóvel está construído (m^2)
3	Idade do imóvel	Anos desde a construção
4	N° de quartos	-
5	N° de banheiros	-
6	Vagas de garagem	Número de vagas de garagem
7	Ar condicionado	Presença de sistema de ar condicionado central ou de sistema de ar condicionado Split (Dummy; 1=sim)
8	Lareira	(Dummy; 1=sim)
9	Infraestrutura	Existência de infraestrutura condominial (Piscina e salão de festas) (Dummy; 1=sim)
10	Localização	Distância ao CBD

Quadro 2 – Variáveis selecionadas a partir de revisão da literatura

Após esta etapa, as variáveis foram validadas com especialistas do setor para garantir a sua aplicabilidade ao contexto do estudo. As variáveis aprovadas pelos especialistas foram ‘tamanho do imóvel’, ‘idade do imóvel’, ‘número de quartos’, ‘número de banheiros’, ‘vagas de garagem’, ‘ar condicionado’, ‘localização’ e ‘infraestrutura’. A variável ‘tamanho do lote’ foi considerada irrelevante para este estudo pois, segundo os profissionais, esta característica é relevante quando da análise de dados referentes a casas térreas, nas quais o tamanho do lote impacta não apenas a área privativa, mas também a presença de jardins e piscina, entre outros itens. Sob sugestão dos especialistas, a variável ‘Lareira’ (relevante sobretudo em estudos em locais de temperatura reduzida) foi substituída pela presença de churrasqueira.

Os especialistas apontaram ainda a necessidade de uma variável que representasse a importância da localização do imóvel. O arquiteto e urbanista, assim como o incorporador, criticaram a utilização da metodologia do CBD, comentando que não seria aplicável a Porto Alegre pelo fato de a cidade não possuir um centro de emprego único e bem definido. Além disso, os especialistas observaram que os apartamentos do centro histórico de Porto Alegre (região cujas características mais se assemelham a um centro de emprego) não apresentam uma valorização que permita inferir que a proximidade com o centro de emprego de fato represente o impacto da variável de localização neste contexto. O fato de utilizar um único centro de emprego também foi motivo de críticas à metodologia de distância ao CBD por Ottensmann et al. (2010). Após esta análise, foi sugerido que se substituísse a metodologia do CBD pela pontuação de *walkability* de cada imóvel. W*alkability* (ou "caminhabilidade") diz respeito à medida na qual os aspectos do ambiente construído promovem ou inibem caminhar na região (GLAZIER et al., 2012), e este conceito aplicado contribui para a saúde, bem-estar e qualidade de vida (TONG et al., 2016).

Indicadores de *walkability* são índices numéricos que combinam várias características do ambiente para avaliar o quão caminhável é a região (GLAZIER et al., 2012). Levam em conta a existência de estabelecimentos comerciais associados a compras do dia-a-dia (supermercados, padarias), localidades associadas a atividades sociais (parques, bares, centros culturais), densidade populacional e conectividade das ruas (CORTRIGHT, 2009; GLAZIER et al., 2012; FAVERO et al., 2008). O Walk Score® é o mais consagrado índice de *walkability,* tendo sido utilizado por Duncan et al. (2011) e Cortright (2009), e foi o escolhido para representar a variável de localização no presente estudo. O algoritmo do Walk Score® considera estabelecimentos de 13 categorias e associa uma pontuação a cada estabelecimento posicionado a até 1600 metros do imóvel em estudo. As categorias são: comércio de alimentos (super-mercados e mini-mercados, dentre outros), restaurantes, cafés, lojas, bares, cinemas, escolas, parques, bibliotecas, livrarias, academias, farmácias, lojas de ferramentas e lojas de roupas e de música (CORTRIGHT, 2009). Em essência, o Walk Score® é a medida da proximidade de uma residência a bens e serviços convencionais (CORTRIGHT, 2009). Os estudos de Duncan et al. (2011) e de Carr et al. (2010) sugerem que o índice é uma medida válida e confiável do deslocamento por meio de caminhada de uma região, e de fato há uma forte correlação entre os preços de imóveis residenciais e o índice de *walkability* determinados pelo algoritmo Walk Score® (CORTRIGHT, 2009).

Os especialistas consultados ainda sugeriram a adição das variáveis 'preço do condomínio' - argumentando a possibilidade de que taxas condominiais maiores demandariam preços de venda menores para efetivar a alienação - e 'presença de portaria 24 horas', citando a preocupação popular com a violência urbana. Ambas as variáveis foram incorporadas. Após constatar a dificuldade de se coletar dados para a variável 'idade do imóvel', optou-se por suprimir esta variável do modelo por considerar que a sua ausência não causaria prejuízos ao desenvolvimento da sistemática. Ao final da etapa de consulta aos especialistas, completou-se a lista de variáveis que compuseram o modelo. Tais variáveis estão listadas no Quadro 3.

#	Variável	Descrição
1	Tamanho do imóvel	Área privativa do imóvel ($m^{2)}$
2	N° de quartos	-
3	N° de banheiros	-
4	Vagas de garagem	Número de vagas de garagem
5	Ar condicionado	Presença de sistema de ar condicionado central ou de sistema de ar condicionado Split (dummy; 1=sim)
6	Churrasqueira	(Dummy; 1=sim)
7	Infraestrutura	Existência de infraestrutura condominial (Piscina e salão de festas) (Dummy; 1=sim)
8	Preço Condomínio	Valor da taxa condominial
9	Portaria	Presença de segurança 24 horas (Dummy; 1=sim)
10	Localização	Pontuação do Walk Score®

Quadro 3 – Variáveis validadas pelos especialistas

Selecionadas as variáveis para a composição do modelo hedônico, foram coletadas 120 observações (imóveis descritos pelas variáveis acima), compilando-se em um arranjo semelhante ao Quadro 1 os dados referentes a cada um dos atributos contemplados pelas variáveis independentes da equação de regressão. As observações foram separadas em grupo de treino e grupo de teste, alocando-se 100 e 20 observações respectivamente para os grupos. A análise dos dados através de software estatístico determinou os coeficientes da regressão para as k=10 variáveis, e estes coeficientes foram então aplicados às respectivas variáveis independentes das observações do banco de testes. Foram calculados os erros absolutos

percentuais entre os valores da variável dependente calculados pela análise de regressão e os valores reais para cada observação deste banco, bem como o erro absoluto percentual médio (MAPE) dos dados. O MAPE encontrado foi de 31,53%.

Deu-se início à seleção de variáveis através do processo *leave one variable out at a time*, descrito na seção anterior. Eliminou-se do banco de dados primeiramente a variável 'tamanho do imóvel' e gerou-se o modelo de regressão; o modelo gerado conduziu a um MAPE de 37,55% na porção de teste. Como o erro foi maior do que o erro encontrado quando do uso de todas as variáveis, percebe-se que tal variável oferece informação relevante para predição do valor do imóvel. De tal forma, a variável 'tamanho do imóvel' foi reintegrada ao banco de treino e a variável 'número de quartos', segunda variável do banco, foi eliminada. Sucessivamente, suprimiu-se uma variável por vez e recalculou-se o erro absoluto percentual médio ocasionado pela aplicação ao banco de teste do modelo gerado por cada combinação de variáveis. A Tabela 1 apresenta na sua segunda coluna os valores de MAPE encontrados para cada modelo gerado quando da ausência da variável da primeira coluna.

Variável suprimida	MAPE
Tamanho do imóvel	37.55%
N° de quartos	32.12%
N° de banheiros	31.51%
Vagas de garagem	32.24%
Ar condicionado	29.72%
Churrasqueira	31.22%
Infraestrutura	28.34%
Preço Condomínio	26.93%
Portaria	30.13%
Localização	31.52%

Tabela 1 – Valores de MAPE para cada variável suprimida

Conforme a Tabela 1 apresenta, o modelo apoiado em *k*-1 variáveis que apresentou o menor MAPE quando aplicado ao banco de testes foi aquele composto pelas variáveis 'tamanho do imóvel', 'nº de quartos', 'nº de banheiros', 'vagas de garagem', 'ar condicionado', 'churrasqueira', 'infraestrutura', 'portaria' e 'localização' (ou seja, para um modelo composto por *k*-1 variáveis, o erro é mínimo quando é suprimida a variável 'preço do condomínio'). O passo seguinte consistiu em testar todos os modelos de *k*-2 variáveis

utilizando como base o modelo adotado na etapa anterior, repetindo-se o processo de eliminar uma variável por vez. Os MAPEs são apresentados na Tabela 2.

Variável suprimida	MAPE
Tamanho do imóvel	45.49%
N° de quartos	29.81%
N° de banheiros	26.78%
Vagas de garagem	28.98%
Ar condicionado	25.38%
Churrasqueira	26.17%
Infraestrutura	25.39%
Portaria	28.74%
Localização	27.36%

Tabela 2 – Valores de MAPE para cada variável suprimida (*k*-2)

O melhor modelo composto por *k*-2 variáveis é aquele no qual está ausente a variável 'ar condicionado'. Com um MAPE de 25,38%, este modelo é superior ao modelo adotado para *k*-1, e é, portanto, aquele de melhor eficiência preditiva até esta etapa. As variáveis que compõem o modelo são: 'tamanho do imóvel', 'n° de quartos', 'n° de banheiros', 'vagas de garagem', 'churrasqueira', 'infraestrutura', 'portaria' e 'localização'. A Tabela 3 apresenta os resultados da repetição do procedimento *leave one out at a time* para o novo modelo.

Variável suprimida	MAPE
Tamanho do imóvel	43.66%
N° de quartos	29.45%
N° de banheiros	25.16%
Vagas de garagem	29.42%
Churrasqueira	24.02%
Infraestrutura	24.41%
Portaria	26.72%
Localização	26.03%

Tabela 3 – Valores de MAPE para cada variável suprimida (*k*-3)

A nova variável suprimida é 'churrasqueira'. O modelo de *k*-3 variáveis substitui o modelo anteriormente adotado, uma vez que apresenta valor de MAPE inferior. Os resultados dos testes dos modelos para *k*-4 variáveis estão contidos na Tabela 4.

Variável suprimida	MAPE
Tamanho do imóvel	42.34%
Nº de quartos	29.02%
Nº de banheiros	24.14%
Vagas de garagem	28.44%
Infraestrutura	23.43%
Portaria	25.59%
Localização	24.93%

Tabela 4 – Valores de MAPE para cada variável suprimida (*k*-4)

A variável 'infraestrutura' é removida do modelo. A Tabela 5 demonstra os resultados da sequência do procedimento.

Variável suprimida	MAPE
Tamanho do imóvel	43.88%
Nº de quartos	28.98%
Nº de banheiros	22.89%
Vagas de garagem	26.94%
Portaria	26.30%
Localização	23.96%

Tabela 5 – Valores de MAPE para cada variável suprimida (*k*-5)

'Nº de banheiros' foi a variável descartada nesta etapa, atingindo MAPE de 22,89%. Nova iteração é realizada para *k*-6 variáveis, cujos resultados podem ser observados na Tabela 6.

Variável suprimida	MAPE
Tamanho do imóvel	42.12%
Nº de quartos	28.59%
Vagas de garagem	27.31%
Portaria	25.55%
Localização	23.27%

Tabela 6 – Valores de MAPE para cada variável suprimida (k-6)

O modelo de *k*-6 de menor MAPE 23,27%, é obtido com a supressão da variável 'localização'. Observa-se, portanto, que o modelo composto por *k*-5 variáveis apresentou o menor erro dentre todos os modelos testados (MAPE=22,89%). As iterações subsequentes foram conduzidas da mesma forma, mas os MAPEs apresentaram valores superiores aos obtidos até então. Desta forma, optou-se por suprimir sua apresentação, visto que o melhor modelo foi localizado. Além disso, especialistas entendem que modelos apoiados em menor número de variáveis reduzem a possibilidade de avaliação subjetiva do impacto das variáveis

sobre a precificação. A Figura 2 apresenta a curva de MAPE mínimo para cada número de variáveis; o menor MAPE é encontrado quando 5 variáveis são mantidas no modelo.

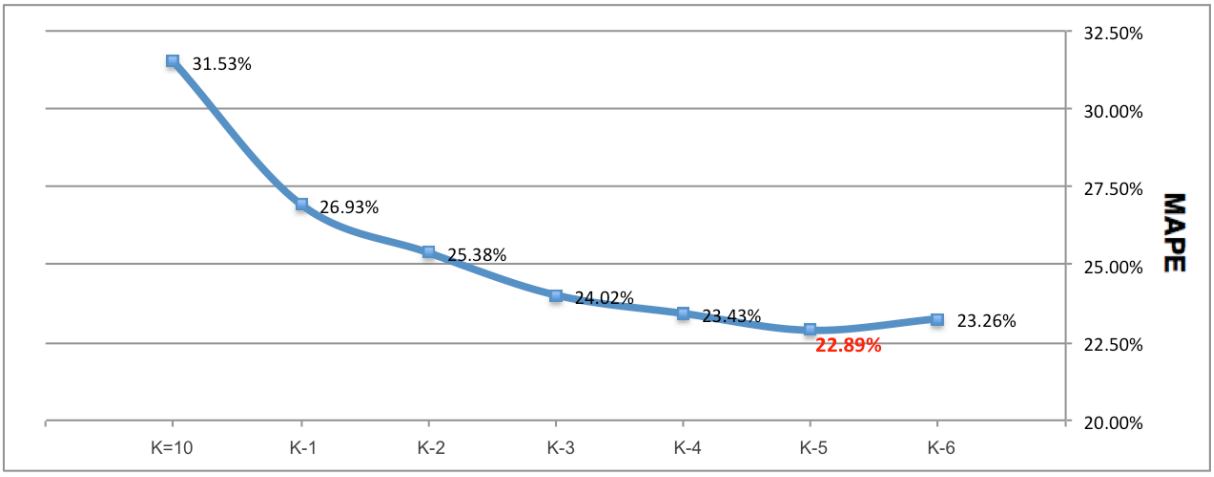

Figura 2 – Erro absoluto percentual médio (MAPE) mínimo por quantidade de variáveis do modelo

Assim sendo, finalizando a aplicação da sistemática, define-se que, para este grupo de dados, o modelo de melhor predição é aquele que contém as variáveis 'tamanho do imóvel', 'n^{o} de quartos 'vagas de garagem', 'portaria' e 'localização', apresentando MAPE de 22,89% quando da aplicação dos seus coeficientes ao banco de teste. A Equação 2 demonstra a equação de regressão deste modelo.

$$Y = 5{,}47X_1 - 85{,}19X_2 + 150{,}46X_3 + 97{,}43X_4 + 0.149X_5 \qquad (2)$$

Na qual:
X_1= Tamanho do imóvel
X_2 = N^{o} de quartos
X_3= Vagas de garagem
X_4= Portaria
X_5= Localização

5 Considerações Finais

A tomada de decisão na indústria da construção civil é historicamente orientada por métodos empíricos. Dentro de um contexto de mercados cada vez mais competitivos, colher as vantagens de ganhos de eficiência torna-se decisivo para a sobrevivência e desenvolvimento do negócio. Em busca destes potenciais ganhos de eficiência, este artigo apresentou uma sistemática de predição do preço de imóveis, com o objetivo de fornecer uma ferramenta científica de suporte à decisão empresarial.

Foi apresentada uma sistemática que se utiliza de regressão linear múltipla para, a partir de um banco de dados, determinar o modelo de regressão de melhor predição dentro de um universo amostral específico. Para a seleção das variáveis preditoras para compor o modelo, procedeu-se com a validação das variáveis com especialistas, seguida do método *leave one variable out at a time*, visando a testar a acurácia de todas as possíveis combinações de características. O modelo responsável pela predição de maior acurácia apresentou MAPE de 22,89%.

Um apontamento interessante é o impacto da variável 'tamanho do imóvel' na acurácia preditiva dos modelos. Ao longo das etapas de seleção, quando a variável foi suprimida, os valores de MAPE atingiram patamares consideráveis: esta métrica não superou os 40% em apenas uma ocasião. Este fato comprova o já esperado impacto substancial da metragem quadrada do imóvel no seu preço. Das 4 variáveis adicionadas mediante sugestão dos especialistas, a pontuação do Walk Score® e a variável 'portaria' permaneceram no modelo de melhor ajuste. A variável 'preço do condomínio', também incorporada devido à sugestão dos profissionais, foi eliminada ainda na primeira sequência de iterações do processo *leave-one-variable-out-at-a-time.*

Para o primeiro modelo – gerado a partir de todas as k variáveis – foi encontrado um erro de 31,53%, enquanto que o erro encontrado pelo melhor modelo – de k-5 variáveis – foi de 22,89%, possibilitando medir a eficácia do método de seleção de variáveis. Não obstante, o MAPE do melhor modelo ainda é relevantemente alto, fato que evidencia a necessidade de se buscar outras variáveis que possam ser adicionadas ao modelo para explicar de forma mais completa os valores da variável dependente.

É importante referir que os resultados encontrados para os coeficientes de regressão de cada conjunto de variáveis, para o coeficiente de erro e para o conjunto de variáveis definido como ótimo, tendem a ser específicos aos dados da amostra coletada. Em outras palavras, os valores dos parâmetros β – que representam a quantificação do impacto gerado por cada característica do imóvel no seu preço –, assim como as características X escolhidas para compor o modelo final não necessariamente podem ser diretamente aplicados a dados diferentes dos aqui utilizados. Entretanto, a sistemática proposta é recomendada para futuras aplicações, sendo a sua metodologia e o seu algoritmo de execução perfeitamente replicáveis a qualquer conjunto amostral.

Sugere-se também para aplicações futuras que sejam consideradas outras variáveis que representem o impacto da localização de um imóvel no seu preço para serem utilizadas em conjunto com a pontuação de *walkability*, assim como variáveis que contemplem a posição solar e a vista do apartamento. Além disso, recomenda-se também incorporar à análise um método de identificação e remoção de *outliers*. Por fim, a geração de modelos não-lineares também aparece como alternativa promissora.

Referências

ADBI, Hervé, WILLIAMS, Lynne J. Jackknife. In: SALKIND, Neil J. *Encyclopedia of Research Design.* Thousand Oaks: Sage. 2010. 1776 páginas.

ALVES, Denisard Cneio de Oliveira, YOSHINO, Joe Akira, PEREDA, Paula Carvalho, AMREIN, Carla Jucá. Modelagem dos Preços de Imóveis Residenciais Paulistanos. *Revista Brasileira de Finanças,* v. 9, n. 2, p. 167-187, 2011.

ANZANELLO, Michel J., ALBIN, Susan L., CHAOVALITWONGSE, Wanpracha A. Multicriteria Variable Selection for Classification of Production Batches. *European Journal of Operational Research,* v. 218, n. 1, p. 97-105. 2011.

ANZANELLO, Michel J., FOGLIATTO, Flavio S. A Review of Recent Variable Selection Methods in Industrial and Chemometrics Applications. *International Journal of Production Economics,* v. 130, n. 2, p. 268-276. 2011.

AZMI, Ahmad S. M., NAWAWI, Abdul H., LATIF, Siti N. F. A., LING, Nur L. F. J. Knowledge Management Obstacles in Real Estate (Valutation) Organisations: Towards quality property services. *Procedia – Social Behavioural Sciences,* v. 202, p. 159-168, 2015.

BAILEY, Martin J., MUTH, Richard F., NOURSE, Hugh O. A Regression Method for Real Estate Price Index Construction. *Journal of the American Statistical Association,* v. 58, n. 304, p. 933-942, 1963.

BOAVENTURA, Edivaldo M. *Metodologia da Pesquisa: Monografia, Dissertação e Tese*. São Paulo: Atlas. 2004. 160 páginas.

BOVER, Olympia, VELILLA, Pilar. Hedonic House Prices Without the Characteristics: The Case of New Multiunit Housing. *European Central Bank, Working Paper Series,* n.117, 2002.

CARR, Lucas J., DUNSIGER, Shira I., MARCUS, Bess H. Walk Score™ As a Global Estimate of Neighborhood Walkability. *American Journal of Preventive Medicine,* v. 39, n. 5, p. 460-463. 2010.

CEBULA, Richard J. The Hedonic Pricing Model Applied to the Housing Market of the City of Savannah Historic Landmark District. *The Review of Regional Studies,* v. 39, n. 1, p. 9-22, 2009.

CHIPMAN, Hugh, GEORGE, Edward I., MCCULLOGH, Robert E. The Practical Implementation of Bayesian Model Selection. *Institute of Mathematical Statistics: Monograph Series,* v. 38, n.1, p. 65-116. 2001.

CORTRIGHT, Joe. How Walkability Raises Home Values in U.S. Cities. *Walking the Walk,* CEO For Cities. 2009. 30 páginas.

DUNCAN, Dustin T., ALDSTADT, Jared, WHALEN, John, MELLY, Steven J., GORTMAKER, Steven L. Validation of Walk Score® for Estimating Neighbourhood Walkability: An Analysis of Four US Metropolitan Areas. *International Journal of Environmental Research and Public Health,* v. 8, n. 1, p. 4160-4179. 2011.

FÁVERO, Luiz Paulo Lopes, BELFIORE, Patrícia Prado, FRANCO DE LIMA, Gerlando A. S. Modelos de Precificação Hedônica de Imóveis Residenciais na Região Metropolitana de São Paulo: uma abordagem sob as perspectivas da demanda e da oferta. *Estudos Econômicos*, v. 38, n. 1, 2008.

GIL, Antonio Carlos. *Como Elaborar Projetos de Pesquisa*. 4ª edição. São Paulo: Atlas. 2002. 176 páginas.

GLAZIER, R. H., WEYMAN, J.T., CREATORE, M. I., GOZDYRA, P., MOINEDDIN, R., MATHESON, F. I., DUNN, J. R., BOOTH, G. L. Development and Validation of an Urban Walkability Index for Toronto, Canada. *Toronto Community Health Profiles,* 2012.

GUO, Mingzhen, WU, Qing. The Empirical Analysis of Affecting Factors of Shanghai Housing Prices. *International Journal of Business and Social Science,* v. 4, n. 14, 2013.

GUYON, Isabelle, ELISSEEFF, André. An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research,* v. 3, n. 1, p.1157-1182. 2003.

HE, Chengjie, WANG, Zhen, GUO, Huaicheng, SHENG, Hu, ZHOU, Rui, YANG, Yonghui. Driving Forces Analysis for Residential Housing Price in Beijing, *Procedia Environmental Sciences,* v. 2, n. 1, p. 925-936. 2010.

KAPLANSKI, Guy, LEVY, Haim. Real Estate Prices: an international study of seasonality's sentiment effect. *Journal of Empirical Finance,* v. 19, n. 1, p. 123-146, 2012.

KOSCHINSKY, Julia, LOZANO-GRACIA, Nancy, PIRAS, Gianfranco. The Welfare Benefit of a Home's Location: an empirical comparison of spatial and non-spatial model estimates. *Journal of Geographical Systems,* v. 14, n. 3, p. 1-38. 2012.

KUTNER, Michael H., NACHTSHEIM, Christopher J., NETER, John, LI, William. *Applied Linear Statistical Models.* 5ª edição. Nova York: McGraw-Hill. 2005. 1396 páginas.

LINDLEY, D.V., SMITH, A. F. M. Bayes Estimates for the Linear Model. *Journal of the Royal Statistical Society, Series B (Methodological),* v. 34, n. 1, p. 1-41. 1972.

MATTOS VIANNA, Raquel, SALLES SOUZA SANTOS, Maria Aparecida. *Déficit Habitacional no Brasil em 2013: Resultados Preliminares.* Nota Técnica, Fundação João Pinheiro**.** 2015.

MEHMOOD, Tahir, LILAND, Kristian H., SNIPEN, Lars, SÆBØ, Solve. A Review of Variable Selection Methods in Partial Least Squares Regression. *Chemometrics and Intelligent Laboratory Systems*, v.118, n. 1, p. 62-69. 2012.

MILLER, Norm, SAH, Vivek, SKLARZ, Michael, PAMPULOV, Stefan. Correcting for the Effects of Seasonality on Home Prices. *The Appraisal Journal,* v. 80, n. 1, p.46-53, 2012.

MONSON, Matt. Valuation Using Hedonic Pricing Models. *Cornell Real Estate Review,* v. 7, n. 1, p. 62-73, 2009.

MONTGOMERY, Douglas C., PECK, Elizabeth A., VINING, Geoffrey G. *Introduction to Linear Regression Analysis.* 5ª edição. Nova Iorque: John Wiley & Sons. 2012. 660 páginas.

O'HARA, R. B., SILLAMPÄÄ, M. J. A Review of Bayesian Variable Selection Methods: what, how and which, *Bayesian Anal,*v. 4, n. 1, p. 85-117. 2009.

OTTENSMANN, John R., PAYTON, Seth, MAN, Joyce. Urban Location and Housing Prices within a Hedonic Model. *The Journal of Regional Analysis & Policy,* v. 38, n. 1, p. 19-35, 2008.

OWUSU-ANSAH, Anthony. Examination of the Determinants of Housing Values in Urban Ghana and Implications for Policy Makers. *Journal of African Real Estate Research,* v. 2, n. 1, p. 58-85. 2012.

RAWLINGS, John O., PANTULA, Sastry G., DICKEY, David A. *Applied Regression Analysis: A Research Tool.* 2ª edição. Nova York: Springer-Verlag. 1998. 658 páginas.

RIVALS, Isabelle, PERSONNAZ, Léon. Neural Network Construction and Selection in Nonlinear Modeling. *IEEE Transactions on Neural Networks,* v. 14, n.4, p. 804-819. 2003.

ROSEN, Sherwin. Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition. *Journal of Political Economy,* v. 82, n. 1, p. 34-55, 1974.

SAIZ, Albert (2010). The Geographic Determinants of Housing Supply. *The Quarterly Journal of Economics*, v. 125, n. 3, p. 1253-1296, 2010.

SCHILL, Michael H. Regulations and Housing Development: What We Know. *Cityscape: A Journal of Policy Development and Research*, v. 8, n. 1, p. 5-20, 2005.

SELTMAN, Howard J. *Experimental Design and Analysis.* Carnegie Mellon University. 2015. 414 páginas.

SHI, Song, YANG, Zan, TRIPE, David, ZHANG, Huan. Uncertainty and New Apartment Price Setting: A Real Options Approach. *Pacific-Basin Finance Journal,* v. 35, pt. B, p. 574-591, 2015.

SILVA, Edna Lúcia da, MENEZES, Estera Muszkat. *Metodologia da Pesquisa e Elaboração de Dissertação.* 4ª edição. Florianópolis: UFSC. 2005. 138 páginas.

SIRMANS, Stacy G., MACDONALD, Lynn, MACPHERSON, David A., ZIETZ, Emily Norman. The Value of Housing Characteristics: A Meta Analysis. *The Journal of Real Estate Finance and Economics,* v. 33, n. 3, p. 215-240, 2006.

SIRMANS, Stacy G., MACPHERSON, David A. The Composition of Hedonic Pricing Models: A Review of Literature. *Journal of Real Estate Literature,* v. 13, n. 1, p. 1-44, 2005.

THOMPSON, Mary L. Selection of Variables in Multiple Regression: Part I. A Review and Evaluation. *International Statistical Review,* v. 46, n. 1, p. 1-19. 1978.

TONG, Xin, WANG, Yaowu, CHAN, Edwin H.W. International Research Trends and Methods for Walkability and Their Enlightment in China. *Procedia Environmental Sciences,* v. 36, p. 130-137. 2016.

WEISBERG, Sanford. *Applied Linear Regression.* 3ª edição. New Jersey: Wiley-Interscience. 2005. 336 páginas.

XU, Xiaoqing, CHEN, Tao. The Effect of Monetary Policy on Real Estate Price Growth in China. *Pacific-Basin Finance Journal,* v. 20, n. 1, p. 62-77, 2012.