3
Metodologia

3.1.
A epistemologia e o tipo da pesquisa

Na presente tese, utilizou-se o método hipotético-dedutivo com o teste das hipóteses formuladas e a busca das causalidades. A pesquisa foi quantitativa na sua essência, uma vez que, a maior parte das características comumente utilizadas nas regressões hedônicas são variáveis quantitativas. São também empregadas as variáveis dicotômicas, quando é necessária a inclusão de algum atributo de qualidade (por exemplo, a variável da posição) ou de indicação de data (assinalando o período em que ocorreu a transação).

A pesquisa também foi quantitativa no sentido de estabelecer um modelo matemático que permitisse construir um índice de preços para apartamentos. Foi fundamental não só que as características valorizadas pelos compradores fossem identificadas, mas também e que estivessem disponíveis em uma base de dados. Como o objetivo principal foi estabelecer uma nova metodologia para construção de índices de imóveis residenciais, também foram definidos os processos necessários a tal construção.

A pesquisa adotou as taxionomias apresentadas por Gil (1996) e Vergara (2003), que dividem a pesquisa em dois aspectos: quanto aos fins e meios. Quanto aos fins, a pesquisa foi exploratória, explicativa e metodológica. Exploratória, porque muito pouco existe publicado no Brasil sobre a questão da construção de índices de preços de imóveis. Explicativa, pois foram identificados alguns dos fatores que contribuem para as variações de preços. Metodológica, pois o objetivo final foi construir um índice de preços de imóveis que pudesse ser replicado com relativa facilidade por uma prefeitura brasileira. Quanto aos meios de investigação, foi uma investigação documental, bibliográfica.

3.2.
Etapas da pesquisa

Para a construção de um índice de preços de apartamentos prontos, a partir de dados obtidos nas prefeituras das cidades brasileiras, foram revistos os artigos acadêmicos de publicações com ênfase no mercado imobiliário, no Brasil e no exterior. Hardin, Liano e Chan (2006) fizeram uma extensa pesquisa para identificar os mais influentes periódicos, instituições e pesquisadores na área de mercado imobiliário (*Real Estate*). A partir dos resultados obtidos por estes autores, a pesquisa bibliográfica foi complementada para a tese.

A revisão bibliográfica envolveu artigos sobre a construção de índices de preços de imóveis residenciais publicados por autores brasileiros, americanos, europeus e asiáticos, em periódicos influentes. Uma grande parte dos artigos publicados sobre índices de preços nos periódicos americanos indexados, no entanto, se refere ao método de vendas repetidas.

Com base nessa revisão, foi construído um modelo com as variáveis existentes e formuladas as hipóteses da pesquisa. Optou-se pelo modelo hedônico.

A pesquisa desenvolveu-se da seguinte forma:

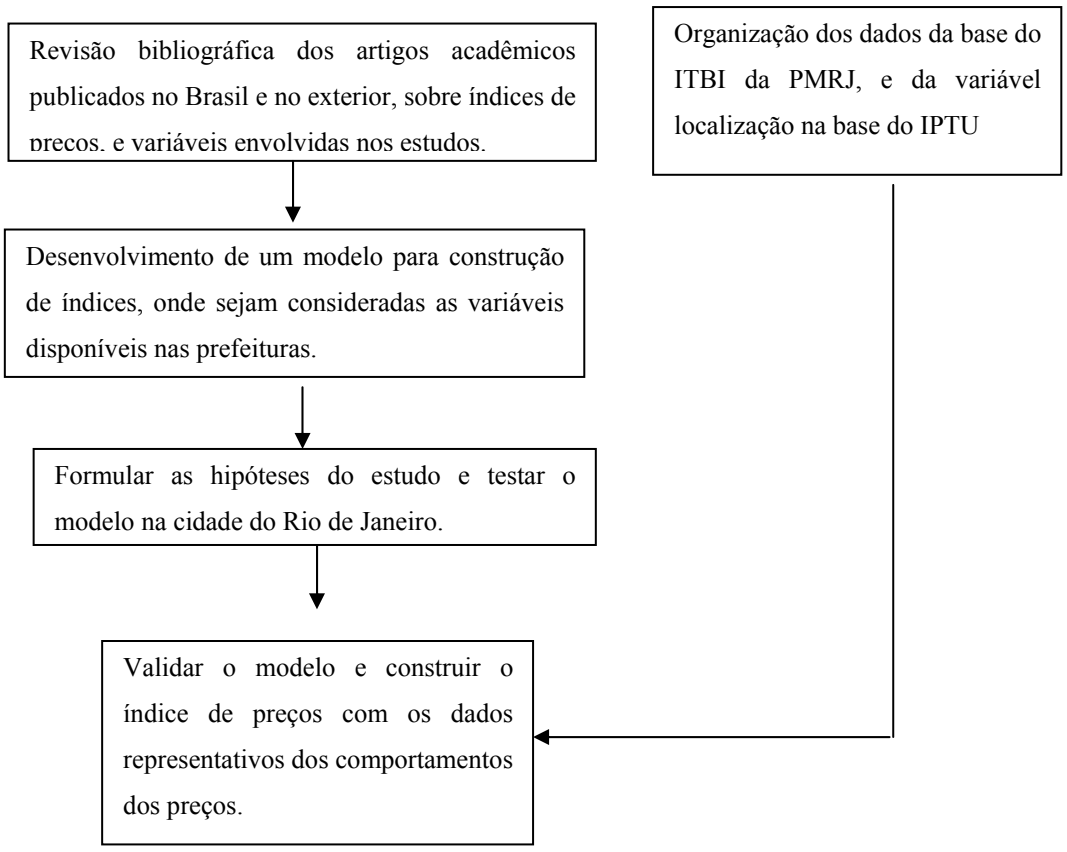


Figura 1- Etapas da pesquisa

3.3.
Universo, amostra e seleção de sujeitos

O universo das unidades amostrais foi a base de dados do ITBI da PMRJ com os preços praticados nas vendas de imóveis na cidade do Rio de Janeiro no período de 1999 a 2007. A seleção, obtida por uma amostragem estratificada que dividiu a população em grupos homogêneos, consiste de unidades residenciais prontas para serem habitadas (apartamentos), negociadas no período mencionado. Para o período do estudo (1999.4 a 2007.2), existiam quase trezentos e dez mil registros espalhados por trinta e quatro Regiões Administrativas, que respondem pela administração pública dos cento e setenta e sete bairros do Município do Rio de Janeiro.

Foram excluídas as casas, pois apresentavam dois componentes na sua formação de preço, o terreno e a área interna da casa. A base de dados do ITBI não permite obter, separadamente, uma área privativa da casa e a área do terreno. Também foram excluídos os terrenos, os endereços genéricos, tais como Av. Projetada e logradouros ainda não reconhecidos pela PMRJ (por exemplo, Rua 2 do PAL 5343). As transações observadas em prédios com numeração suplementar para outro logradouro, de apartamentos de cobertura ou do tipo dúplex também não foram consideradas.

Para redução ou mesmo eliminação de erros de lançamento, limitamos a área privativa no intervalo entre trinta e trezentos metros quadrados e a idade em trinta anos.

O objetivo de propor uma metodologia para a construção do índice de preço de unidades residenciais (apartamentos) levou em consideração a facilidade de replicação da metodologia por qualquer prefeitura. Foram adotados os procedimentos da nova norma brasileira de avaliação de imóveis urbanos (NBR-14653-2).

3.4.
Coleta dos dados

Os dados sobre os preços de imóveis foram obtidos na Prefeitura do Rio de Janeiro, que mantém um setor de acompanhamento dos preços do mercado imobiliário, responsável por manter atualizado o valor dos imóveis para atribuir

valores ao ITBI (Imposto de Transmissão de Bens Imobiliários). O valor para o VR de cada logradouro foi obtido no texto da lei municipal número 2585/97.

3.5. Tratamento dos dados

Índices de preço construídos a partir de modelos com regressões hedônicas dependem tanto das especificações das variáveis independentes, quanto do método utilizado para o cálculo do índice. A NBR 14653-2, da Associação Brasileira de Normas Técnicas (ABNT), com validade a partir de 30/06/04, foi denominada "Avaliação de Bens Parte 2: Imóveis urbanos". Nessa norma, no Anexo A, constam os procedimentos recomendados para o uso e a validação dos modelos de regressão linear. As regressões utilizadas no modelo atenderam às principais recomendações da norma. O modelo obtido foi considerado apto ao uso por qualquer órgão público.

3.5.1. A base de dados do ITBI

Imposto sobre a Transmissão de Bens Imóveis e de direitos a eles relativos realizado inter-vivos, por ato oneroso, esse é o nome da taxa municipal mais conhecida por ITBI. A obrigação do pagamento cabe ao adquirente e o ITBI, no Município do Rio de Janeiro, é cobrado no valor de 2% (dois por cento) sobre uma base de cálculo, que é o valor declarado pelo contribuinte. Se a autoridade fiscal (PMRJ) discordar, é feito um arbitramento, mas sujeito a recurso pelo contribuinte. O arbitramento, segundo a PMRJ, é feito a partir de critérios tecnicamente reconhecidos pela norma vigente de avaliação de imóveis.

O gerente do Setor de ITBI da PMRJ, em uma das muitas reuniões durante o processo de autorização da utilização da base de dados, assim descreveu (extra-oficialmente) a atuação dos responsáveis pela coleta de preços:

> *"Além da base de dados permanentemente atualizada, alimentada constantemente pelas guias emitidas, os técnicos, engenheiros e estagiários coletam e verificam os preços junto aos anúncios de jornais, junto às incorporadoras e corretoras".*

O próprio pesquisador, por várias ocasiões já discutiu, pessoalmente, com os responsáveis pelo arbitramento (nos anos 1980 e 1990) alguns valores que tinham sido recusados. Nessas ocasiões, quando informado da avaliação e colocado a par dos dados do mercado, não houve alternativa a não ser pagar o imposto conforme arbitrado.

Como o imposto é devido por ocasião das escrituras definitivas de compra e venda, os cartórios de notas só lavram os instrumentos com a apresentação da guia paga do ITBI. O imóvel também pode ter sido adquirido por uma promessa de compra e venda, por instrumento particular, em um lançamento imobiliário ou de um vendedor qualquer. Nessas situações, o imposto é devido após 30 dias da data do último pagamento estipulado. Nas circunstâncias descritas ocorre o seguinte:

a) No caso de uma promessa de compra e venda (PCV), o adquirente (promitente comprador) declara o preço da PCV e o ITBI, e se o responsável pelo lançamento do ITBI não concorda, arbitra o preço do imóvel (avaliação). O mesmo acontece em transações particulares. Como a legislação também prevê multas por omissão do pagamento do ITBI na data (30 dias após o ultimo pagamento), são muito poucos os casos de atraso de pagamento.

b) Em um lançamento imobiliário, a venda de um apartamento em construção quase sempre prevê a contratação pelo comprador de um financiamento em longo prazo.

Nessa última situação, por ocasião da entrega das chaves, é lavrada uma escritura, onde o preço é atualizado e o ITBI pago sobre o novo preço de face ou por uma avaliação. A data do fato gerador em uma promessa de compra e venda é a data da apresentação de solicitação da guia e, assim, não há efeito pelo tempo decorrido. Quando na base de dados do ITBI é apresentada uma "data de fato gerador" além de uma "data de recolhimento", ocorre um atraso no pagamento da guia, com conseqüente cobrança de encargos. Isso também acontece por qualquer outro motivo fora do normal.

Para evitar distorções, foram eliminadas todas as transações com multas (significando terem sido apresentadas após o tempo hábil). Dessa forma, entre a data de avaliação pela PMRJ e a inclusão da transação na base de dados, provavelmente não existe lapso de tempo. Na realidade, quando a escritura é

definitiva, a guia só é paga depois da escritura. O cartório fica com o cheque nominal à Prefeitura e providencia o pagamento. De outra forma, no caso de algum contratempo ou uma desistência, o processo de recuperação de um imposto seria uma tarefa demorada e árdua.

O comprador do imóvel é responsável pelo pagamento do ITBI (imposto municipal) no valor de 2% (dois por cento)[3] sobre o preço declarado na escritura ou arbitrado. O vendedor será taxado em 15% (quinze por cento) pelo lucro apurado conforme a legislação vigente do imposto sobre a renda.

Por esse motivo, ambos poderiam em algumas situações, omitir o verdadeiro valor da transação, registrando na escritura um valor menor. Tal prática vem sendo combatida pela Secretaria de Receita Federal através da obrigação de fornecimento de informações, a declaração de operação imobiliária (DOI) pelos cartórios (obrigatória desde 1999) e a declaração de informação sobre atividades imobiliárias (DIMOB, obrigatória desde 2003) pelos participantes indiretos das transações, no caso, as incorporadoras, corretoras e administradoras. O cruzamento de informações com as declarações emitidas pelos cartórios e pelas partes envolvidas permite o rastreamento das transações com omissão de preço. Já a PMRJ, desde os anos 1980, antes de emitir as guias de ITBI, verifica o valor de mercado, que só é utilizado no caso de o valor apresentado à tributação estar abaixo daquele valor.

Dessa forma, podem ocorrer as seguintes situações:

a) O valor informado para a transação está acima do valor de mercado.

b) O valor informado para a transação é igual ao valor de mercado.

c) O valor informado para a transação está abaixo do valor de mercado segundo os critérios de avaliação da PMRJ.

No caso da ocorrência do último item, a PMRJ aplica, para cálculo do ITBI, o valor de mercado. O comprador pode recorrer desde que apresente uma avaliação, conforme a norma vigente, elaborada por um profissional legalmente habilitado.

3- sobre a parte financiada pelo SFH, o ITBI é de 0,5% (meio por cento)

Se tal avaliação for aceita pela PMRJ, provavelmente o valor foi baixo por algum motivo que tenha depreciado o imóvel, tal como a necessidade de reformas ou a existência de encargos condominiais e tributários em atraso. Nesses casos, é normal uma dedução no preço.

Quando é analisado um grande número de transações, uma das preocupações é quanto à ocorrência de elementos estranhos. Normalmente, o pesquisador estabelece limites e remove tudo o que estiver fora dos limites. Conforme Laferrère (2003), o critério adotado foi eliminar os sextos superiores e inferiores. Entretanto, em outro estudo, Gouriéroux e Laferrère (2006), só eliminaram os quartos superiores e inferiores dos preços por metro quadrado. O estudo de Fletcher, Gallimore e Mangan, (2000) propôs eliminar as transações com mais de dois desvios padrão.

Optou-se nesta pesquisa pela eliminação dos sextos, por ser o método aplicado na cidade de Paris (LAFERRÈRE, 2003).

Ao eliminarmos um sexto com os menores preços praticados (por metro quadrado de área privativa) para cada região estudada, provavelmente os casos onde os imóveis têm algum problema são eliminados.

Ao eliminarmos a sexta parte com os maiores preços (por metro quadrado de área privativa), provavelmente foram eliminadas as transações cujos preços são fora do padrão, pela ocorrência de atributos difíceis de avaliar tais como, uma vista privilegiada (mar, lagoa, parques), benfeitorias adicionadas e não relatadas, tais como uma cozinha planejada, armários embutidos, ar condicionado, vagas de automóvel adicionais e também unidades situadas em prédios com amenidades extraordinárias.

3.5.2.
O atributo de localização (VR)

Em 14/11/1997, a câmara municipal da cidade do Rio de Janeiro aprovou a lei de número 2585, na qual foi publicado um anexo com a última planta genérica de valores (PGV) para os efeitos da cobrança do IPTU.

A PGV contém todas as ruas do município do Rio de Janeiro classificadas por bairros e atribui o VR (valor unitário padrão residencial, em UFIR da época) do metro quadrado de avaliação imobiliária para um apartamento de frente para o

logradouro. Tais valores é que vão formar o valor venal (base de cálculo IPTU) de um imóvel. Na realidade, são os indicadores da diferença de preço decorrente da localização. O VR foi considerado como um atributo de localização.

Embora a PGV tenha sido projetada para o uso no ano base de 1998, ficaram preservadas as características de cada logradouro em relação aos demais. Conferimos alguns valores de ruas conhecidas e o VR pode ser utilizado como indicador do atributo localização, comumente chamado de ponto, no mercado imobiliário.

A base de dados do ITBI forneceu poucas variáveis explicativas das variações de preços: localização física (endereço e bairro), área da unidade, posição em relação ao logradouro (frente ou não), idade do imóvel. Um pesquisador que atue mercado imobiliário pode usar seus conhecimentos para avaliar os atributos de cada endereço (em relação aos demais vizinhos) que podem explicar variações no preço. Isso pode ser feito facilmente um pequeno trecho de um bairro, mas é impossível, para uma só pessoa, analisar uma cidade inteira.

Pressupomos corretas, até porque podem ser contestadas, as avaliações representativas dos preços de mercado efetuadas pela equipe do ITBI da PMRJ.

3.5.3.
Variáveis explicativas disponíveis e a variável dependente

Ramos, Silva, Loch (2000) propuseram um modelo para avaliação em massa de imóveis, com vistas à atualização da base de cálculo dos tributos imobiliários municipais (ITBI e IPTU), usando a base de dados do município de Blumenau (SC). Dividiram os municípios em áreas homogêneas e utilizaram oito variáveis independentes, sendo quatro delas variáveis dicotômicas. Em noventa e três observações, removeram nove *outliers*. O R^2 ajustado foi de 0, 8581.

Silva, Loch e Ramos (2006), em estudo na cidade de Blumenau (SC), ressaltaram o valor do atributo *localização* em um estudo sobre o emprego de análises espaciais na avaliação em massa e concluíram pela sua viabilidade, desde que a base de dados do cadastro técnico municipal disponibilize informações para o processamento de dados geográficos e também apresente mais informações relevantes, que, de outra forma, teriam de ser obtidas individualmente.

Já Meese e Wallace (1997) construíram índices para Oakland e Fremont (CA) com seis variáveis independentes (números de banheiros, número de quartos, área, condições da propriedade e idade) para a estimativa do preço.

Os mercados imobiliários municipais no Brasil, especialmente na cidade do Rio de Janeiro, apresentam alguns atributos que são responsáveis pela percepção de valor por parte dos compradores, tais como área privativa, localização, idade do imóvel, acessos, amenidades, segurança, vista.

Tais características, principalmente localização e acessos, estudadas em Sirmans et al.(2005) são algumas das mais percebidas pelos compradores. Não se deveria descartar, contudo, no período considerado (1999-2007), a influência das tensões sociais e da criminalidade sobre os preços. Esta tese tem como objetivo principal a construção de um índice de preços com as variáveis disponíveis para cada endereço individualizado. As estatísticas de criminalidade, que no caso do Estado do Rio de Janeiro correspondem a várias delegacias, são publicadas por região, conforme a abrangência das AISP- Área Integrada de Segurança Pública. Nas cidades, as PGV das prefeituras contêm os valores dos VR que já levam em consideração sua proximidade em relação a comunidades menos favorecidas. Carvalho e Lemme (2005) dimensionaram perdas no valor de apartamentos em relação à sua distância dessas comunidades. Uma mesma rua apresenta vários VR que aumentam à medida que os apartamentos se afastam destas comunidades. O VR por atribuir um valor a uma localização (ao nível de endereço) é também provavelmente um indicador adequado para a insegurança local, associada a comunidades. Um modelo hedônico com o atributo *segurança* foi estudado em Clark e Cosgrove (1990), que constataram a disposição do comprador de maior renda em pagar mais por uma localização com menos incidência de crimes (em nível de cidade).

Nos modelos com regressões hedônicas, são comumente utilizadas as seguintes características (SIRMANS et al. 2005): o número de quartos, a área privativa, o tipo da unidade, o número de vagas de garagem, equipamentos de aquecimento e refrigeração, a idade da unidade, o padrão da construção, a vizinhança, a distância ao centro comercial, a qualidade das escolas públicas, a distância à rodovia principal, o clima, as áreas de lazer, a vista.

A especificação das características hedônicas depende, em parte, da experiência do pesquisador e do conhecimento do funcionamento do mercado.

Usamos as variáveis que estavam disponíveis nas bases de dados da prefeitura da cidade do Rio de Janeiro. As bases de dados do ITBI são provavelmente as únicas que podem ser objeto de estudos que envolvam um grande número de transações.

As variáveis independentes utilizadas foram:

VR	Valor relativo do atributo localização
Andar	A altura do apartamento é valorizada
Idade	A idade da unidade influencia na negociação do preço
Área	Tamanho interno (área privativa)
Posição	Indica se o apartamento é de frente (1) ou não (0)

Tabela 2- Variáveis Independentes

A pesquisa de Abramo (2003) concluiu que 74% das causas de mudanças (mobilidade residencial) tiveram como motivação a localização, contra 26% que declararam como causa o próprio imóvel. Um bom modelo hedônico deve sempre incluir uma variável de localização, quanto mais específica melhor, que represente as características da vizinhança (TSE, 2002). A localização é uma variável importante em qualquer estudo sobre avaliação de imóveis, mas dados sobre transações com identificação precisa da unidade são quase impossíveis de se obter no Brasil, a não ser na base de dados do ITBI. É possível obter o preço médio e algumas das características das unidades na base de dados da ADEMI-RJ, mas apenas para lançamentos imobiliários. Na cidade do Rio de Janeiro, existe apenas a publicação da base do SECOVI-RJ, em convênio com o jornal O Globo, que dispõe de preços médios de imóveis ofertados por tipologia (FORTUNATO et al., 2007). Essa base de dados é coletada por levantamentos de anúncios publicados em jornais de grande circulação e apresenta como principais problemas a não identificação do preço real da transação e das características dos apartamentos, indicando unicamente uma média de preços.

No modelo descrito em Laferrère (2003), a variável dependente foi o valor do metro quadrado de área privativa. Não foi usada a mesma variável para o Brasil, por causa da tendência, observada por especialistas do mercado imobiliário, de valorizar o preço do metro quadrado em unidades pequenas e o desvalorizar nas grandes unidades. Por exemplo, em Ipanema, apartamentos de dois quartos com 80m² poderiam apresentar um preço por metro quadrado de R$ 4.000,00 (quatro mil reais), enquanto outro apartamento de 200m² na mesma rua

(com o mesmo valor do VR) pode ter um preço observado de R$3.500,00 (três mil e quinhentos reais).

O objetivo desta tese foi a construção de um índice e não a avaliação de unidades. Foi considerado mais adequado adotar como variável dependente a avaliação da unidade. A suposição é de que a uniformidade de tratamento dos preços submetidos a uma prefeitura (solicitação da guia do ITBI) leva a uma segura observação das variações dos preços, resultando em um índice adequado.

O acesso ao real valor da transação imobiliária, no Brasil, é muito difícil. Os cartórios dos registros de imóveis, onde é possível obter uma certidão com os valores pagos, não mencionam a área privativa ou idade do imóvel, que são características muito importantes para o pesquisador. Como a certidão é obtida por requerimento do interessado a um custo tabelado, para um determinado imóvel, o pesquisador teria de saber qual o imóvel objeto da transação para, assim, solicitar a informação.

Os lançamentos imobiliários podem servir de base para indicação de preços e de características dos imóveis, mas não há um preço de transação individualizado e sim uma média das unidades ofertadas com os preços de tabela. Pesquisadores podem ter acesso a esses dados pela base de dados da ADEMI (no caso, a do Rio de Janeiro) mediante solicitação.

Como as transações imobiliárias são feitas por instrumentos públicos ou particulares e os cartórios não fornecem informações no Brasil, não há qualquer outra fonte que opere com dados primários e disponíveis em número suficiente

Tendo em vista o objetivo da construção de um índice, a solução foi utilizar a base de dados do ITBI das prefeituras, com as limitações existentes, principalmente, quanto ao pequeno número de variáveis explicativas.

Os países da OCDE que já dispõem de índices de preços são indicados na tabela 3 a seguir :

EUA	Nationwide single family house price index
Japão	Nationwide urban land price index
Alemanha	Index for total Germany, total resales
França	Indice de prix des logements anciens
Itália	Media13area urbane numeri indice dei prezzi medi di abitazioni, usate
Inglaterra	Mix-adjusted house price index
Canada	Multiple listing series, average price in Canadian dollars
Austrália	Index of a weighted average of 8 capital cities
Dinamarca	Index of one-family house sold
Espanha	Precio medio del m2 de la vivienda, mas de un año de antiguedad
Finlândia	Housing prices in metropolitan area, debt free, price per m2
Irlanda	Second hand houses
Coréia da Sul	Nationwide house price index
Holanda	Existing dwellings
Noruega	Nationwide index for dwellings
Nova Zelândia	Quotable value index for dwellings (new and existing)
Suécia	Sweden One and two dwelling buildings
Suiça	Single-family home

Tabela 3- Índices de Preços na OCDE
Fonte: GIROUARD et al. (2006)

3.5.4.
Redução da base de dados (idade, área, sextos)

Gonzalez (2002) considera ser necessária uma redução quando a base de dados é excessiva. Recomenda uma primeira limpeza de erros grosseiros e também uma redução realizada com base nos atributos.

Harding, Knight e Sirmans (2003), em estudo sobre a "barganha" nas negociações de preços no mercado imobiliário, usaram base de dados de duas cidades (Modesto, CA e Baton Rouge, La). Para tornar homogêneas as amostras, também reduziram a base de dados por preço de unidade de área e por características fora do padrão (ar condicionado central e terrenos pequenos).

A redução da base de dados, nesta tese, em linha com os estudos de Gonzalez (1997), Gonzalez (2002), Harding, et. al (2003), Laferrère (2003), Maurer et al (2004), foi feita conforme os seguintes critérios:

a) Retirada das transações de imóveis com idade igual ou maior de 31 anos, por causa da provável necessidade de reformas e da legislação que obriga a existência de vagas de automóveis. Uma unidade de mais de 31 anos poderia ter sido projetada em uma época em que a legislação não obrigava a previsão de vagas de automóvel.

b) Retirada das transações de imóveis com áreas inferiores a 30m² e superiores a 300m² para eliminação de tendências.

c) Ordenação por preço do metro quadrado e retirada da sexta parte das transações com o menor preço e a sexta parte das transações com o maior preço, na pressuposição de que grande parte das transações com características não observadas que possam influir no preço sejam removidas.

d) Exame de cada uma das transações remanescentes, seguido de eliminação daquelas que apresentem erros, anotações, duplo logradouro, apartamentos de cobertura, do tipo duplex, com falta de numeração ou com localização provisória (rua projetada)

3.5.5. Divisão em regiões homogêneas (RH)

Foi necessário dividir a cidade do Rio de Janeiro em regiões homogêneas, quanto à valorização ou à desvalorização imobiliária. Para isso, foi considerada a estrutura existente das regiões administrativas da cidade do Rio de Janeiro. Essa divisão funcional estabelecida pela PMRJ permite a alocação de recursos em conjuntos de bairros semelhantes, bem como o estabelecimento de planos de obras e intervenções. A escolha do bairro ou bairros representantes de cada região homogênea foi feita com base na representatividade e preferência dos incorporadores para novas construções.

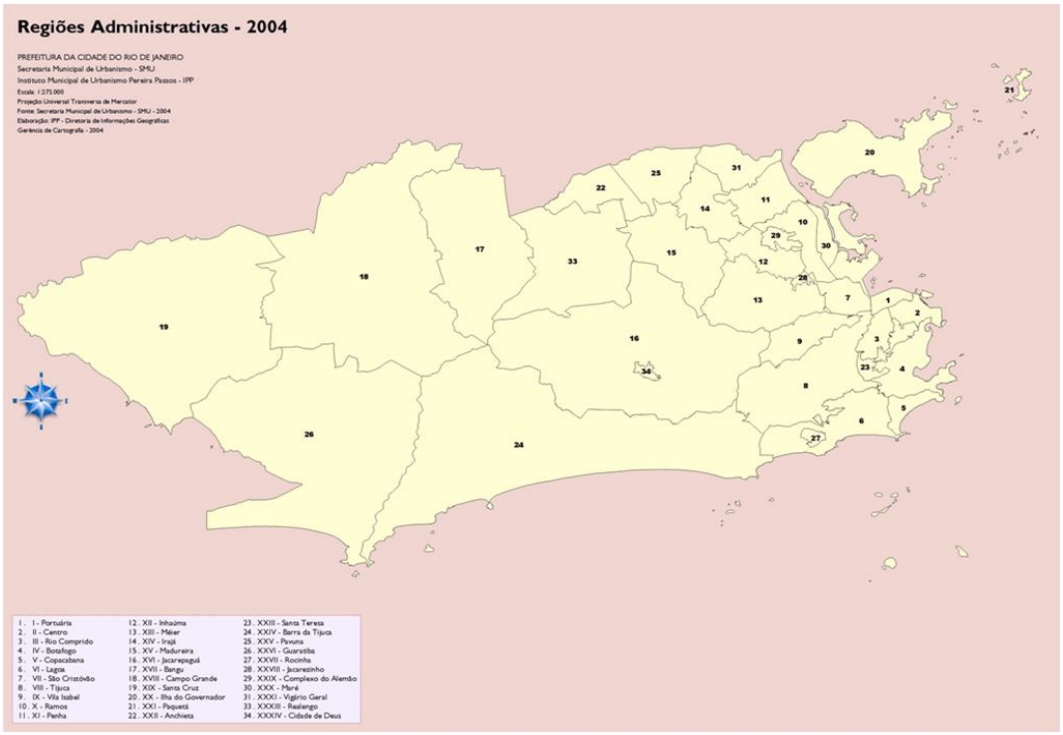


Figura 2- regiões administrativas da cidade do Rio de Janeiro
Fonte: Prefeitura Municipal da cidade do Rio de Janeiro

A Tabela 4 mostra a divisão da cidade do Rio de Janeiro em regiões homogêneas e seus representantes. Alguns bairros de algumas regiões administrativas não foram considerados para participar do índice. Os complexos de comunidades de baixa renda não participaram pela falta de titulação nesses endereços, o que desobriga (informalmente) ao pagamento do ITBI, e também pela inexistência de parâmetros de avaliação pelo setor do ITBI. Alguns bairros apresentam muito poucas transações e não têm tido lançamentos imobiliários há muito tempo. Também existem bairros onde a construção de edifícios é rara, com predominância de casas, como Santa Teresa.

Região Administrativa	Bairros	Critério Adotado (RH)	Representante
I Administração Regional - Portuária	Caju, Santo Cristo, Saúde e Gamboa.	N	N
II Administração Regional - Centro	Aeroporto, Castelo, Centro, Fátima, Lapa e Praça Mauá.	N	N
III Administração Regional - Rio Comprido	Catumbi, Cidade Nova, Estácio e Rio Comprido.	N	N
IV Administração Regional - Botafogo	Botafogo, Catete, Cosme Velho, Flamengo, Glória, Humaitá e Laranjeiras.	1	Botafogo
V Administração Regional - Copacabana	Copacabana e Leme.	2	Ipanema Leblon Copacabana
VI Administração Regional - Lagoa	Gávea, Ipanema, Jardim Botânico, Lagoa, Leblon, São Conrado e Vidigal.	2	Ipanema Leblon Copacabana
VII Administração Regional - São Cristóvão	Benfica, São Cristóvão, Triagem e Vasco da Gama.	N	N
VIII Administração Regional - Tijuca	Alto da Boa Vista, Praça da Bandeira e Tijuca.	3	Tijuca
IX Administração Regional - Vila Isabel	Andaraí, Grajaú, Maracanã e Vila Isabel.	3	Tijuca
X Administração Regional - Ramos	Bonsucesso, Olaria e Ramos.	4	Méier,Olaria
XI Administração Regional - Penha	Brás de Pina, Penha e Penha Circular.	4	Méier,Olaria
XII Administração Regional - Inhaúma	Del Castilho, Engenho da Rainha, Inhaúma, Higienópolis, Maria da Graça e Tomaz Coelho	4	Méier,Olaria
XIII Administração Regional - Méier	Abolição, Água Santa, Cachambi, Consolação, Encantado, Engenho de Dentro, Engenho Novo, Jacaré, Lins de Vasconcelos, Méier, Piedade, Pilares, Riachuelo, Rocha,	4	Méier,Olaria
XIV Administração Regional - Irajá	Colégio, Irajá, Vicente de Carvalho, Vila da Penha, Vila Kosmos e Vista Alegre.	4	Méier,Olaria
XV Administração Regional - Madureira	Bento Ribeiro, Campinho, Cascadura, Cavalcante, Engenheiro Leal, Honório Gurgel, Madureira, Marechal Hermes, Osvaldo Cruz, Quintino Bocauiuva, Rocha	4	Méier,Olaria
XVI Administração Regional - Jacarepaguá	Anil, Curicica, Freguesia, Gardênia Azul, Jacarepaguá, Pechincha, Praça Seca, Tanque, Taquara e Valqueire	5	Freguesia
XVII Administração Regional - Bangu	Bangu, Gericinó, Padre Miguel, Santíssimo (Bangu) e Senador Camará.	6	Campo Grande
XVIII Administração Regional - Campo Grande	Campo Grande, Cosmos, Inhoaíba e Senador Augusto Vasconcelos.	6	Campo Grande
XIX Administração Regional - Santa Cruz	Paciência e Santa Cruz	6	Campo Grande
XX Administração Regional - Ilha do Governador	Bancários, Cacuia, Cidade Universitária, Cocotá, Freguesia (Ilha), Galeão, Jardim Carioca, Jardim Guanabara, Moneró, Pitangueiras, Portuguesa, Praia da Bandeira, Ribeira, Tauá	7	Jardim Guanabara Portuguesa
XXI Administração Regional - Paquetá	Paquetá.	N	N
XXII Administração Regional - Anchieta	Anchieta, Guadalupe, Parque Anchieta e Ricardo de Albuquerque.	N	N
XXIII Administração Regional - Santa Teresa	Santa Teresa.	N	N
XXIV Administração Regional - Barra da Tijuca	Barra da Tijuca, Camorim, Grumari, Itanhangá, Joá, Recreio dos Bandeirantes, Vargem Grande e Vargem Pequena.	8	Barra da Tijuca
XXV Administração Regional - Pavuna	Acari, Barros Filho, Coelho Neto, Costa Barros, Parque Colúmbia e Pavuna.	N	N
XXVI Administração Regional - Guaratiba	Barra de Guaratiba, Guaratiba, Pedra de Guaratiba e Sepetiba	N	N
XXVII Administração Regional - Rocinha	Rocinha.	N	N
XXVIII Administração Regional - Jacarezinho	Jacarezinho e Vieira Fazenda.	N	N
XXIX Administração Regional - Complexo do Alemão	Complexo do Alemão.	N	N
XXX Administração Regional - Maré	Baixa do Sapateiro, Conjunto Pinheiros, Marcílio Dias, Maré, Nova Holanda, Parque União, Praia de Ramos, Roquete Pinto, Rubens Vaz, Timbaú, Vila do João, Vila	N	N
XXXI Administração Regional - Vigário Geral	Cordovil, Jardim América, Parada de Lucas e Vigário Geral.	N	N
XXXIII Administração Regional - Realengo	Campo dos Afonsos, Cordovil, Deodoro, Jardim América, Magalhães Bastos, Parada de Lucas, Realengo, Sulacap e Vigário Geral, Vila Militar	N	N
XXXIV Administração Regional - Cidade de Deus	Cidade de Deus.	N	N

Tabela 4- Regiões, bairros e seus representantes
Fonte: Elaborada pelo autor com dados da PMRJ

3.6.
Construção do índice de uma região homogênea (RH)

A Figura 3 indica a preparação dos dados para construção do índice de uma RH:

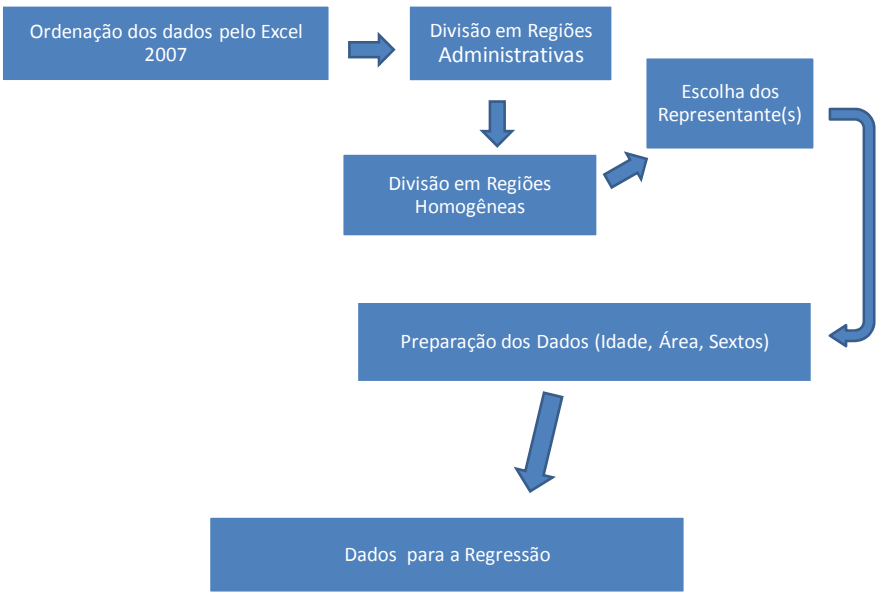


Figura 3- Preparação dos dados
Fonte: elaborado pelo autor

Assim como adotado em Clapp et al (1991), foi utilizada uma equação para cada período de três anos, com variáveis dicotômicas para representar um determinado período trimestral. Determinou-se, assim, por média aritmética para o ano de 1999 (primeiro ano-base), o valor dos atributos, área, idade, VR, andar, posição, para uma hipotética unidade padrão (HUP) em cada uma das RH. Com uma unidade padrão, teremos atributos padronizados, isto é, qualidade constante.

Para os períodos trimestrais de 1999.4 até 2002.4 obteve-se uma regressão múltipla com variáveis dicotômicas para cada trimestre. Quando os atributos da unidade padrão de 1999 foram aplicados aos coeficientes da equação de regressão, obteve-se um valor de avaliação para cada um dos trimestres. Se ao valor obtido para 1999.4 for dado o valor de 100, os demais valores comporão a evolução do índice de preços. O conceito de unidade padrão por média aritmética também foi usado em Maurer et al. (2004). Já Laferrère (2003) utilizou o conceito de média geométrica.

O processo repete-se para 2002.4 a 2005.4, com ano-base de 2002 e para 2005.4 a 2007.2, com ano-base de 2005. Os cálculos do índice em 2002.4 e 2005.4, com duas bases diferentes, são necessários para se proceder à mudança de base de 1999 para 2002 e, depois, para 2005.

As etapas podem ser resumidas da seguinte forma:

1- Divisão do Rio de Janeiro em regiões homogêneas (RH).

2- Escolha de um ou mais bairros representantes de cada uma dessas áreas.

3- Para os trimestres de 1999 (4º trimestre) até 2002 (4º trimestre), obtenção da equação de regressão e do valor trimestral da unidade padrão.

4- Obtenção, no ano base de 2002 (três anos decorridos), de uma nova unidade padrão e repetição do procedimento anterior. Repetição de novo em 2005.

5- Obtenção de um índice para cada uma das RH do período de 1999.4 a 2007.2, na base de 2005.

6-Agregação dos índices das nove RH e construção de um índice para o Município do Rio de Janeiro.

A Figura 4 exibe o número de transações remanescentes após cada eliminação até chegar ao número final de 40.789 transações utilizadas nas regressões.

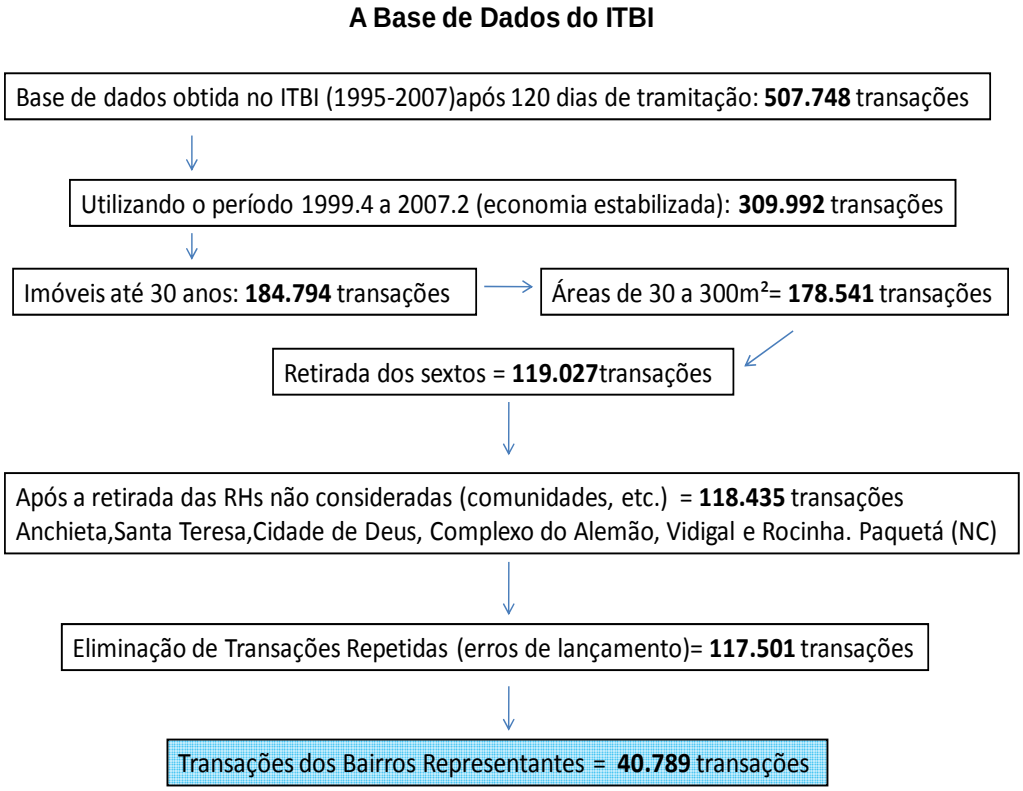


Figura 4- Base de dados do ITBI
Fonte: elaborado pelo autor

3.6.1.
A forma funcional da regressão múltipla

O modelo hedônico proposto estuda os preços dos apartamentos a partir de cinco variáveis explicativas, que podem ser obtidas nas bases de dados das prefeituras. Conforme Malpezzi, Chun, e Green (1998), os modelos parcimoniosos reduzem a variância dos coeficientes hedônicos. Para os autores, tais modelos só poderiam ser um problema se o objetivo fosse a obtenção de preços implícitos das características (*estimate implicit prices*), mas não no caso de previsão de preços de unidades residenciais (*predict house prices*). Mathews (2006) alertou para que os modelos hedônicos parcimoniosos sejam específicos na questão da localização.

Um dos importantes fundamentos teóricos para a equação hedônica é a sua forma. Para cada uma das nove regiões homogêneas, em cada um dos três períodos, foi obtida uma regressão múltipla, com as transformações necessárias e validadas pelos testes.

A forma genérica antes das transformações foi:

$$\textbf{Aval} = \beta_0 + \textbf{Idade}\beta_1 + \textbf{VR}\beta_2 + \textbf{posição}\beta_3 + \textbf{Andar}\beta_4 + \textbf{Área}\beta_5 + \varepsilon$$

Mc CLAVE et al. (2001, p.535) propõem a seguinte construção para o modelo de regressão múltipla:

a) Escolha da forma funcional.
b) Estimativa dos parâmetros a partir da amostra original.
c) Realização das modificações necessárias para atender aos pressupostos da regressão.
d) Avaliação estatística do modelo.
e) Utilização o modelo para previsão ou estimação.

Para especificação do modelo, recomendam análise gráfica dos resíduos contra cada variável independente.

Chatfield (1995) considerou que a análise dos dados deve envolver a seguinte estratégia:

- exclusão, ponderação ou ajuste de pontos atípicos;
- transformação de variáveis para obtenção da normalidade e da variância constante dos resíduos.

Sugeriu também que os pesquisadores devem optar por uma abordagem mais pragmática, na qual um modelo parcimonioso que obtenha uma adequada aproximação com os dados existentes talvez seja melhor do que a difícil tarefa de se encontrar um modelo verdadeiro.

Gujarati (2000, p.463), por sua vez, propõe que se examinem no modelo os resultados obtidos (R^2 ajustado, t estimados, sinais esperados dos coeficientes, estatística de Durbin- Watson) e se os diagnósticos forem bons, seria possível declarar o modelo como válido.

3.6.2.
Testes de verificação dos pressupostos da regressão múltipla

Para a verificação dos pressupostos básicos da equação de regressão, os testes aplicados foram:

a) Linearidade - gráficos da variável dependente x cada variável independente.

b) Normalidade - gráficos de densidade em relação à normal e dos resíduos.

c) Pontos Atípicos - verificação e remoção justificada.

d) Homocedasticidade - teste de Breusch-Pagan / Cook-Weisberg

e) Verificação da autocorrelação - pela aplicação do teste de Durbin-Watson e se necessário, uso do gráfico da variável dependente x resíduos padronizados.

f) Colinearidade e Multicolinearidade - observação dos VIF (*variance inflation factors*)

Quanto à significância (das variáveis e da regressão), foram usados os testes *t* e *F* e, para o poder de explicação, a observação do R^2 ajustado.

Para cada variável independente, foi usado o *gladder* e, quando necessário, o *ladder of powers* (STATA10). Este apresenta nove resultados das transformações possíveis de uma determinada variável em relação a sua distribuição normal (com *p-values*) e aquele mostra os nove gráficos das mesmas transformações. Pelos gráficos das variáveis transformadas, com o apoio do *ladder of powers*, foi possível escolher a transformação que apresentasse a melhor forma (do tipo normal). Tais escolhas, no caso de dúvida na interpretação, foram ratificadas por outros gráficos do tipo *pnorm* e *qnorm*. A normalidade dos resíduos é necessária para que os *p-value* dos testes *t* e o resultado do teste *F* sejam considerados válidos. O *pnorm* é um gráfico da probabilidade normal padronizada sensível à normalidade no intervalo médio dos dados, enquanto o *qnorm* é sensível à normalidade perto das extremidades.

A linearidade, após escolhida a transformação, foi confirmada pelos gráficos da dependente contra cada independente (comando *acprplot* X, *lowess*).

A remoção de pontos atípicos, a forma adotada (*pooled cross sections*), a heterocedasticidade, a multicolinearidade e os erros de especificação foram discutidos em separado nos próximos itens.

Uma vez definida a nova forma funcional, obteve-se uma regressão com a análise dos testes (*F-test*), seus *p-value*, um R^2 ajustado e a função beta, que compara a força relativa dos preditores.

Finalmente, foi testada a especificação com a aplicação do *linktest* (STATA). O *linktest* cria duas novas variáveis, uma delas denominada "*hat*" (variável de previsão) e outra denominada "*hatsq*" (o quadrado da variável de previsão). O modelo é refeito com essas duas variáveis independentes. A variável "*hat*" deve ser significativa, pois é o valor previsto. Já "*hatsq*" não deve ser significativa, pois, se o modelo está corretamente especificado, as previsões ao quadrado não devem ter poder de explicação significativo. Assim, o valor do *p-value* de *hatsq* deve ser maior do que 0,05 para que a regressão seja considerada corretamente especificada.

3.6.3. As variáveis indicativas dos trimestres e a regressão do tipo *"pooled cross section"*

O ciclo de produção imobiliária no Brasil dura em média três anos, período que vai desde os estudos de viabilidade, durante a fase de aquisição do terreno, até a entrega das chaves. Os lançamentos imobiliários são precedidos de pesquisas em relação às preferências do consumidor e o preço pago por novas unidades a serem produzidas influencia o preço das unidades prontas que estão à venda. Essa é a razão de limitar em três anos o período onde são mantidas constantes as características do modelo hedônico. Além das preferências dos consumidores, outros fatores podem influir na mudança dessas características, notadamente uma mudança da legislação municipal da ocupação do solo. Quase sempre as mudanças se traduzem em menor área de pavimento e menor altura da edificação. Uma valorização nas regiões onde uma nova legislação torna a construção mais restrita é provável.

Quando as observações de preços de transações são obtidas por períodos trimestrais, elas são independentes e seus erros não são correlacionados. Elas se diferenciam de uma amostra randômica simples, pois, ao serem obtidas em

diferentes períodos no tempo, podem mostrar observações que não são identicamente distribuídas. São também diferentes de um painel de dados que, mesmo tendo as duas dimensões de seções cruzadas e tempo, coletam informações das mesmas pessoas, famílias, empresas, cidades ou estados em vários pontos do tempo. Quando se juntam as observações obtidas a cada trimestre, há uma combinação de seções cruzadas. A introdução de variáveis dicotômicas para cada trimestre (menos para o primeiro trimestre) resolve o problema das distribuições diferentes em cada trimestre, sem complicações estatísticas e com a vantagem de aumentar o tamanho da amostra (WOOLDRIDGE 2003, pp.409). Não ocorre, pois, uma correlação serial, pois as observações são independentes para cada período trimestral. Existe a probabilidade de ocorrência, nessas equações, de heterocedasticidade, que pode, contudo, ser resolvida com a utilização de uma regressão do tipo WLS (*weighted least squares*) ou com o processo de correção de Huber-White (*robust standard errors*). Li e Prud' Homme (2006) concluíram que o uso de "*pooled cross section*" resulta em um índice parecido com aquele obtido por regressões separadas, a cada trimestre. Como a evolução tecnológica na construção civil é lenta, o uso desse tipo de equações não apresenta os problemas que ocorrem quando as características das unidades mudam rapidamente, como em computadores, por exemplo. A denominação deste tipo de regressões hedônicas como "*pooled cross section"* foi discutida em Conniffe e Duffy (1999).

3.6.4. Justificativa para eliminação de observações

Belsley, Kuh e Welsch (2004) definiram que uma observação influente é aquela que, individualmente ou em conjunto com outras, tem demonstrado grande impacto nos valores dos coeficientes.

Mesmo sem ser um ponto atípico quando o resíduo é grande, um determinado ponto é dito influente se sua simples remoção, ou ainda sua remoção combinada com outros pontos, modifica substancialmente os resultados da regressão. Os pontos atípicos (*outliers*) podem ser identificados por gráficos ou testes e podem ocorrer também nas variáveis explanatórias. São denominados

pontos de alta alavancagem (*high leverage points*), de forma a distingui-los dos pontos atípicos da variável dependente.

Conforme e Hadi (2006, pp.101), a análise dos resíduos pode falhar na detecção de pontos atípicos e influentes e, por essa razão são necessários testes adicionais. Chatterjee e Hadi (1986) resumiram as medidas mais utilizadas para a detecção de pontos atípicos, pontos influentes e de pontos de alta alavancagem. Para a análise dos resíduos foi utilizada na pesquisa a medida do RSTUDENT (Belsley, Kuh e Welsch, 2004, p.20) também chamada de *jackknife residuals* por Atkinson (1982) e *studentized residuals* por Velleman e Welsch (1981). Esses autores preferem examinar os resíduos estudentizados em vez dos padronizados. Hair et al (2005) também indicam a análise do resíduo estudentizado que, embora conceitualmente semelhante ao resíduo padronizado, elimina o impacto sobre o modelo de regressão. Os resíduos estudentizados de uma regressão são os resíduos divididos pelo erro padrão dos resíduos. Um *outlier* potencial é definido como uma observação cujo resíduo estudentizado está fora do intervalo (-3, 3). Se existirem, a investigação deve ser feita quanto a erros na base de dados, ou ainda quanto à omissão de variáveis no modelo, ou à violação dos pressupostos da regressão. Se, com a remoção dos *outliers* algum dos parâmetros muda substancialmente, a remoção deve ser justificada (PARDOE, 2006).

Uma observação com valor extremo em uma variável explicativa é um ponto de alta alavancagem. A alavancagem é a medida do quão longe está uma variável independente de sua média. Pontos de alta alavancagem podem causar efeitos na estimativa dos coeficientes da regressão. Para a medida de alta alavancagem (*high leverage*), foi seguido o modelo de Hoaglin e Welsch (1978), nos quais são considerados de alta alavancagem aqueles com valor superior a 2(k+1)/n. Esses pontos devem ser examinados para uma eventual eliminação.

Outras medidas de pontos influentes dizem respeito à função de influência (IF) proposta por Hampel (1974). A medida utilizada foi a distância de Cook , similar à distância de Welsch-Kuh, denominada como DFFITS. Como a NBR sugere a distância de Cook, esse foi o método utilizado na pesquisa. A distânia de Cook é considerada a medida mais representativa de uma influência sobre o ajuste geral (Hair, 2005 p. 192). Devem ser investigadas as observações com uma distância d> 4/n.

Chatterjee e Hadi (2006) e Gujarati (2005) também consideram, DFFITS e Distância de Cook, como medidas similares. A medida da influência desses pontos sobre cada coeficiente da variável independente (denominada de influência parcial) foi proposta por Belsley, Kuh e Welsch (2004 p.13) sob a denominação de dfbetas, que devem ser examinados se excederem $2/\sqrt{n}$.

Lacour-Little e Malpezzi (2001) e Chave e Thompson (2003) usaram técnicas de detecção e remoção de *outliers* por IQR ("*interquartile range*" dos resíduos), considerando que devem ser eliminadas as observações ($x < Q(25)-3IQR$ ou $x > Q(75) +3IQR$) fora das "cercas" estabelecidas (*outer fences*). O procedimento foi explicado em Sheskin (2004) e os *outliers* fora destas "cercas" são denominados *severe outliers***.**

Os preços de apartamentos prontos dependem de várias características que, mesmo não existentes na base de dados, poderiam ser observadas localmente pelo pesquisador. Dentre as características, as mais importantes seriam as amenidades próximas, a degradação da vizinhança, a poluição (ambiental e sonora), o estado e a infra-estrutura da construção, a distância ao comércio, a distância a escolas. É impossível observar essas características para cada uma das transações. Com poucas características utilizadas (cinco), é provável o aparecimento de grande número de observações caracterizadas como atípicas, influentes e de alta alavancagem.

A investigação de pontos influentes e de pontos atípicos foi feita a partir de considerações necessárias para a justificativa da necessidade de remoção. No modelo construído com a base de dados do ITBI utilizaram-se apenas as variáveis disponíveis e uma variável de localização. Com essas cinco variáveis independentes, o estudo e a remoção de observações influentes devem obedecer aos seguintes conceitos, conforme Hair (2005, p. 164).

a) Para um erro de observação, corrigir ou eliminar.

b) Se a observação é válida, mas excepcional, e explicada por uma situação extraordinária, a remoção é justificada.

c) A observação excepcional, sem explicação, é um problema que deve ser resolvido pelo pesquisador.

d) A observação comum em suas características individuais, mas excepcional no conjunto, deve ser mantida.

O pesquisador deve examinar as observações influentes detectadas pelos métodos propostos. Deve ainda eliminar aquelas justificadas como excepcionais e manter as que possam ser representativas da população (Hair et al, 2005). O uso das técnicas de detecção desses pontos, com os critérios propostos para justificar sua remoção, proporcionou um resultado satisfatório no objetivo da pesquisa, que é a estimação do preço médio de uma unidade padrão a cada trimestre.

Conforme González e Formoso (2000), no mercado imobiliário é muito comum, por suas características, a ocorrência de pontos atípicos. Conniffe e Duffy (1999) mencionam a identificação e exclusão dos pontos atípicos (*outliers*) na formação do índice ESRI (*Irish Permanent*) e do *Halifax Index*. O processo é considerado necessário para a proteção contra erros de dados e erros nas características do imóvel. Os pontos atípicos removidos representavam em torno de 2% (dois por cento) da amostra.

Foram adotadas nesta pesquisa, dentre as propostas mais utilizadas na literatura para exame dos pontos influentes, resíduos estudentizados, alavancagem, *severe outliers (IQR)*, dfbetas e Distância de Cook**.** Também foram listadas as observações consideradas pontos atípicos por cada um dos métodos e removidos conforme os critérios propostos. Muitos desses pontos são comuns a vários métodos. Foi considerada provável a existência de muitos desses pontos pela natureza da base de dados e pela quantidade de observações.

A justificativa para a remoção de tais pontos da base de dados do ITBI foi considerada na seguinte forma:

1) *Localização* – é considerada como preço "diferenciado" aquele observado sob o impacto (positivo ou negativo) da proximidade de uma comunidade, praia, parques e jardins

2) *Área interna* – uma área menor ou igual a A m² e maior ou igual a B m² pode distorcer o preço conforme os valores estipulados para A e B em cada região estudada. Por exemplo, um apartamento com uma área de 200 m² em um bairro popular, provavelmente será uma observação removida.

3) *Idade* – unidades imobiliárias com até dois anos de idade podem ser apartamentos novos. Os valores podem conter distorções provocadas pela preferência dos compradores pela aquisição de unidades nunca habitadas

e, portanto, mais valorizadas. Mas unidades com mais de 25 anos podem ter sido reformadas ou necessitarem de reformas.

4) *Andar* – em determinadas regiões, um andar alto pode conter uma vista privilegiada (Ex:- Botafogo > 12). Ao longo do tempo, modificações na legislação da região produzem projetos arquitetônicos diferenciados. Quando existe a permissão de construção mais alta que a média, algumas unidades (as mais altas) podem ter uma vista privilegiada, que irá se traduzir em um preço mais elevado, um ponto atípico.

REPRESENTANTES	**VR**	**Andar (vista)**	**Idade**	**Área**
Barra da Tijuca I e II	Praia	18 ou mais	Até 2 e mais de 20	50-250
CampoGrande	Comunidade	8 ou mais	Até 2 e mais de 25	50-150
Ipanema,Leblon Copacabana	Praia e comunidade	12 ou mais	Até 2 e mais de 25	50-250
Meier Olaria	Comunidade	8 ou mais	Até 2 e mais de 25	50-150
Botafogo	Praia e comunidade	12 ou mais	Até 2 e mais de 25	50-180
Freguesia(Jacarepaguá)	Comunidade	12 ou mais	Até 2 e mais de 25	50-150
JardimGuanabara Portuguesa	Praia e comunidade	6 ou mais	Até 2 e mais de 25	50-150
Tijuca	Comunidade	12 ou mais	Até 2 e mais de 25	50-180

Tabela 5- Critérios da justificativa para remoção de pontos atípicos
Fonte: Elaborado pelo autor

Outra solução (persistindo a presença de *outliers*) possível, mas que não foi utilizada nesta tese é o uso da correção pelo teste *Jackknife,* que mostra quando as observações têm impacto nos resultados. Esse teste roda a mesma regressão (n-1) vezes, deixando de lado uma observação de cada vez. No caso do teste mostrar algum impacto significativo no resultado, seriam utilizados os coeficientes e os erros padrão fornecidos, que serão mais robustos aos *outliers* que os de uma OLS

normal. Outro método de correção é o *bootstrap,* que roda várias vezes a mesma regressão para diferentes amostras aleatoriamente obtidas na base original.

Cook e Weisberg (1982) recomendam primeiro a identificação e o exame dos pontos influentes e atípicos para posteriormente considerar uma regressão OLS (*Ordinary Least Squares*).

Gonzalez (1997) alerta para a necessidade da verificação da presença de observações diferenciadas nos estudos com dados provenientes do ITBI. Dentre as origens dessas observações na base de dados do ITBI, está a própria natureza da transação imobiliária, que pode gerar distorções se o imóvel está bem conservado e recém-reformado, ou degradado e com vazamentos. Uma localização com amenidades (praças, parques, praia, lagoa), uma vista privilegiada ou uma varanda espaçosa podem distorcer os parâmetros. Até mesmo uma razão afetiva pode produzir uma observação diferenciada quando, por exemplo, os pais fazem questão que os filhos morem perto e se dispõem a valorizar determinado imóvel. Dentre outros motivos que depreciam os imóveis e que não são detectados pelas variáveis explicativas do modelo estão a degradação da vizinhança. Ela costuma ocorrer pela falta de segurança, pelo crescimento de comunidades vizinhas, pela mudança de leis municipais de ocupação do solo, pela execução de obras públicas de impacto negativo (por exemplo, um viaduto) ou ainda por uma queda na qualidade dos serviços existentes no bairro (educação, saúde, transportes). Outra causa provável da presença de observações diferenciadas seria a omissão de valores na escrituração.

3.6.5. Heterocedasticidade

Na regressão linear múltipla, um dos pressupostos básicos é a igualdade das variâncias dos erros (homocedasticidade). Quando isso não ocorre, concluí-se que a regressão apresenta uma heterocedasticidade, que pode ser visualizada, sob a forma de um funil, em um gráfico dos resíduos contra os valores estimados da variável dependente ou de uma das variáveis independentes. A presença de heterocedasticidade causa os seguintes efeitos:

- Incorreta estimação dos erros padrão, geralmente, uma subestimação. A inferência estatística é prejudicada.
- A regressão não é mais a melhor, isto é, a mais eficiente e com menor variância estimadora dos coeficientes. Não existe, contudo, o enviesamento dos coeficientes.

Os erros podem aumentar quando o valor de uma variável independente fica maior. Na pesquisa, por exemplo, a variável andar pode causar heterocedasticidade. Em prédios onde o gabarito é de cinco andares, os preços terão pouca variância devido à influência da altura do apartamento (por exemplo, do 2º para o 4º andar). Mas se o gabarito permite dez andares, diferenças entre o 2º andar e o 9º andar serão muito maiores. Os erros associados a preços observados em prédios de alturas diferentes podem apresentar variâncias diferentes.

Da mesma forma, a variável idade apresenta outro problema, já que apartamentos podem ter sido reformados e o atributo idade é reduzido na percepção do comprador. A base de dados não detecta a "idade aparente" e provavelmente isso vai causar aumento na variância dos erros. Erros associados a apartamentos reformados podem ter maior variância do que aqueles associados a apartamentos que aparentam a idade real ou maior.

No caso do atributo de localização (VR), a distância da unidade até uma amenidade (praia, praça, parque, lagoa) ou a uma comunidade (vizinhança degradada) pode influir na variância dos erros das observações do preço, pois vários prédios e mesmo quadras tem o mesmo VR.

Já o atributo área pode apresentar problemas por causa da tendência, no mercado imobiliário, de não se perceber um mesmo preço uniforme por metro quadrado. Uma unidade com 100 metros quadrados, mantidos constantes os outros atributos (idade, andar, VR, posição) provavelmente terá um preço por metro quadrado maior do que outra unidade com 250 metros quadrados. Tal fenômeno se explica pela maior quantidade de pessoas que podem adquirir um apartamento de 100 metros quadrados.

Todas as variáveis independentes (Área, VR, Idade, Andar) podem, portanto, ser a causa da heterocedasticidade. Isso só não deveria ocorrer em bairros onde os prédios têm a mesma altura (baixa), idades mais uniformes (sem

lançamentos imobiliários), uma área de unidade uniforme e uma localização sem grandes diferenças. Na cidade do Rio de Janeiro, um bairro enquadra-se nessas características, a Urca. Em uma pesquisa futura, seria interessante verificar se há evidências que confirmem tal expectativa.

Quando o termo de erro de uma regressão OLS tem a mesma variância para todos os valores das variáveis independentes, os erros são homocedásticos e os valores estimados para a variável dependente são válidos. Se for considerado, por exemplo, o efeito da localização na variável andar, em um bairro como Campo Grande, onde as localizações têm atributos parecidos (sem amenidades ou comunidades), provavelmente os erros terão a mesma variância e os preços estimados serão válidos.

Por outro lado, em Ipanema, onde os prédios têm alturas diferentes, causadas por mudanças do código de obras ao longo do tempo, a variância dos erros será maior, apresentando heterocedasticidade.

3.6.5.1. Correção pelo método da regressão *weighted least squares*

Dentre os testes disponíveis para grandes amostras, o método de Breusch e Pagan (1980), complementado por Cook e Weisberg (1983), é usado pela maioria dos pacotes estatísticos. No STATA, o comando *hettest* depois da regressão programa aquele teste, cujo resultado deve ser comparado na tabela da distribuição qui-quadrado (onde com 5% de significância o valor é de 3,84). O *hettest* testa a hipótese nula de que a variância dos resíduos é homogênea. Quando o resultado do teste é superior a 3,84, acusando a heterocedasticidade, é preciso corrigir.

Testada a presença de heterocedasticidade na regressão do tipo seção cruzada (*cross-sectional*), uma correção pode ser tentada com a utilização de uma transformação para a variável dependente. A utilização da forma logarítmica, por exemplo, reduz a heterocedasticidade. Essa transformação é de fácil computação, bem mais do que o uso da transformação de Box-Cox, que, além de ser de difícil aplicação (MALPEZZI, CHUN e GREEN, 1998), normalmente rejeita a forma linear, semilog e log-linear (TRIPLETT, 2004). O método Box-Cox não permite

que se testem os coeficientes estimados quanto à significância e apresenta resultados de difícil interpretação (HADDAD e HERMANN, 2005)

Se ainda assim persistir a heterocedasticidade, é recomendado o uso da regressão WLS (*weighted least squares*), conforme Carter e Haloupek (2000) e Yaffee (2002).

O uso de estimadores GLS (*generalized least square*) para a correção da heterocedasticidade, denominados de WLS (*weighted least squares*), minimiza a soma ponderada dos quadrados dos resíduos (uma OLS dá um peso constante para cada observação). O peso atribuído pelo WLS a cada resíduo ao quadrado corresponde a 1/h. Uma explicação sobre esse procedimento pode ser obtida em Wooldridge (2003, p.261). Morais e Cruz (2003), em estudo sobre a demanda de habitações no Brasil, utilizaram um modelo hedônico com variáveis dicotômicas para representação das áreas metropolitanas. O peso utilizado foi a participação das regiões no PNAD (pesquisa domiciliar anual do IBGE em regiões metropolitanas).

Como na pesquisa a variável dependente sempre foi transformada no logaritmo do preço, a primeira tentativa de correção da heterocedasticidade foi feita através da utilização da WLS. Um dos problemas é que os pesos analíticos a serem usados na WLS podem não ser encontrados se não sabemos a causa da heterocedasticidade. Na formação do preço, levaram-se em consideração a localização, o andar, a posição (dicotômica), a idade e a área. Na suposição mais comum de que a heterocedasticidade é função de uma das quatro variáveis explicativas (PARK, 1966), a representação gráfica do quadrado dos resíduos contra cada variável independente pode apresentar padrões nos diagramas de dispersão. Isso pode ajudar na transformação e na identificação da causa da heterocedasticidade.

Gujarati (2005, p.384) propõe algumas hipóteses plausíveis a respeito do padrão de heterocedasticidade quando a variância não é conhecida:

a) A variância do erro é proporcional a X^2 (divisão por x).

b) A variância do erro é proporcional a X (a transformação da raiz quadrada).

c) A variância do erro é proporcional ao quadrado do valor médio de Y.

d) Transformação para logaritmos.

Willett e Singer (1987) consideraram que, geralmente, os pesos para a aplicação de uma regressão do tipo WLS não são conhecidos. Os autores ponderaram ainda que qualquer que seja a abordagem, o objetivo é obter o melhor modelo. Se uma OLS, que pode ser vista como uma WLS de pesos um, é insuficiente pelo fato de seus erros não serem homocedásticos, a WLS é uma solução que envolve a criação de pesos inversamente proporcionais à variância observada nas variáveis explicativas, que por suposição são a causa da heterocedasticidade. Com a WLS, os pontos que estavam mais dispersos agrupam-se em torno da linha de regressão. Espera-se, assim, a obtenção de erros padrão menores. Os autores concluíram que com o uso de pesos $1/X^2$ (inverso do quadrado), a variância pode ser estimada corretamente. O método também é indicado pelo STATA e adotado nesta tese.

Conforme Chatterjee e Hadi (2006) o procedimento pode ser explicado da seguinte forma: na presença da heterocedasticidade, uma das formas de correção mais utilizada é a regressão *weighted least squares* (WLS), quando os parâmetros são estimados minimizando a soma ponderada dos resíduos quadrados, cujos pesos são inversamente proporcionais à variância. O procedimento é diferente da OLS, onde os mesmos parâmetros são estimados, minimizando a soma dos resíduos quadrados com pesos iguais.

Considere-se uma equação que tem a forma $y_i = \beta_0+\beta_1 x_i + ...+ \beta_n x_{in} + \varepsilon_i$ e apresenta erros aleatórios, independentes e normalmente distribuídos com média zero e variância σ^2. Se os resíduos não são consistentes com as premissas da regressão, isto é, não são homocedásticos, existem procedimentos para que se possa obter um modelo melhor que o fornecido por regressão OLS. No caso da pesquisa, se a heterocedasticidade é causada por apenas uma variável independente, um gráfico dos resíduos x uma variável independente pode apresentar padrões que evidenciam a ocorrência de heterocedasticidade.

Assim, se supomos que Var (ε_i) é proporcional a X^2 ou seja,

Var (ε_i)= $k^2 x^2_{i2}$, (onde k> 0), os parâmetros não determinados minimizam a soma:

$$\sum_{i=1}^{n} = \frac{1}{x^2_{i2}} (Y_i - \beta_0-\beta_1 x_i - ....- \beta_n x_{in})^2$$

No STATA, para esse procedimento, o peso analítico será $1 / X^{2}_{i2}$.

A visualização dos gráficos dos resíduos x variável independente para identificação da forma pode apresentar dúvidas na interpretação da variável, que é responsável ou, pelo menos, que mais contribui para a presença da heterocedasticidade.

Como foram consideradas apenas quatro variáveis, preferiu-se, na presente tese, examinar pelo STATA (*hettest*) a variável explicativa que mais contribuiu para a heterocedasticidade. Tal procedimento é muito mais rápido do ponto de vista computacional e elimina problemas de interpretação dos gráficos.

3.6.5.2. Correção de Huber-White (*robust standard errors*)

Na presença de heterocedasticidade de forma desconhecida, uma das formas de correção, aproveitando a OLS, é o procedimento denominado "*robust standard errors*" (RSE), válido para grandes amostras. O procedimento também denominado HUBER-WHITE é baseado no trabalho de White (1980) que segue o trabalho de Huber (1967). Após a obtenção dos "*robust standard errors*" por um programa estatístico, eles serão válidos para uma amostra grande, o que é o caso da tese.

Robust Standard Errors (RSE) ou *White Adjustment* ou "*Huber-White Sandwich*" é aplicado para que os erros padrão de uma OLS sejam corrigidos na presença de heterocedasticidade, quando sua forma não é conhecida e não se pode precisar o valor dos pesos para uma WLS. Se não existe heterocedasticidade, a variância estimada na OLS de um coeficiente qualquer é:

$\text{Var}(\beta_{OLS}) = s^2\mu / N*\text{Var}(X)$

A variância verdadeira quando há heterocedasticidade é:

$\text{Var}(\beta_{OLS}) = wi\ s^2\mu / N*\text{Var}(X)$

sendo wi a distância da observação Xi em relação à média de X.

Essa é a base da correção de Huber-White para grandes amostras.

O apelido de "*sandwich*" refere-se ao cálculo do RSE, que é o produto de três matrizes. A matriz do meio, o recheio, é formada pelo produto "*observation-leved likelihood* e *pseudolikelihood score vectors*". Essa matriz é, assim, pré e pós- multiplicada pela "*model based variance matrix*" (o pão).

Wooldridge (2003, pp. 249) recomenda, como forma de lidar com a heterocedasticidade, o uso de estatísticas do tipo "*heterokedasticity-robust*" para grandes amostras, quando não sabemos determinar a sua causa. A obtenção de "heterokedasticity-robust standard errors" após a regressão OLS pode ser feita, entre outros, pela aplicação do método de "Huber-White".

Já Long e Erwin (2000) concluíram que o uso do HCCM (*Heteroscedascity Consisten Covariance Matrix*) permite eliminar os efeitos adversos da heterocedasticidade quando nada se conhece sobre sua forma. Alternativas ao uso do HCCM (denominado de HC0) são os três estimadores (HC1 , HC2 e HC3). O método de "Huber–White", cujo comando no STATA é denominado *robust*, para correção da heterocedasticidade pode ser usado em três versões HC1 , HC2 e HC3. Se o número de observações (n) está entre 250 e 500, o melhor é usar HC3. Para n maior do que 500, todos apresentarão resultados parecidos. Se n é menor ou igual a 250, tanto HC2 ou HC3 são melhores que HC1. O STATA usa como padrão para heterocedasticidade o HC3.

Cai e Hayes (2008), por sua vez, propuseram mais um tipo, o HC4 e apresentam uma boa revisão teórica dos estimadores HCCM (HC0 a HC4). Uma importante propriedade dos RSE é que a forma da heterocedasticidade presente não precisa ser conhecida (CROUX; DAHENE; HOORELBEKE, 2003).

É um procedimento usual na literatura o uso de WLS seguido de Huber-White (*robust standard errors*). Pode ser constado em Ghertman, Obadia e Arregle (1997), Devereux (2004), Kucera e Sarna (2006), Kurosaki (2006), Towsend (2007).

Conforme Colin e Trivedi (2005, p.81), por outro lado, mesmo sendo possível obter uma "*feasible GLS*" com erros homocedásticos, os autores geralmente preferem, por conveniência ou por razão dos pequenos ganhos na eficiência, utilizarem os estimadores WLS ou mesmo OLS com "*robust standard errors*".

Foi estabelecida para a tese uma rotina para a obtenção de uma regressão que corrija a heterocedasticidade:

1- Se uma ou mais das regressões, com os quatro pesos propostos (invX=1/X²) apresentou um ***chi 2*** < 3,84 (5%), a solução é usar a WLS, que apresentou o menor resultado.

2- Se com os pesos propostos não se obteve uma regressão cujos resíduos fossem homocedásticos, a OLS com *robust standart errors* foi a utilizada.

3- Após a correção, um teste de especificação deverá validar a regressão. Se uma WLS corrigiu a heterocedasticidade, mas foi reprovada no teste de especificação, deve-se usar a OLS com *robust standart errors*, desde que validada pelo mesmo teste.

3.6.6. Multicolinearidade

Chatterjee e Hadi (2006, p.69) observaram que, se nenhuma das variáveis explicativas é considerada significativa pelo *teste-t*, ainda que a regressão apresente um poder de explicação significativo, isso pode ser um sintoma de que duas ou mais das variáveis explicativas têm alta correlação. Nesse caso, os coeficientes de uma regressão não serão corretamente interpretados. Existindo tais correlações, ocorre a multicolinearidade, que não é um problema de especificação e sim da base de dados (CHATTERJEE e HADI, 2006 p.220). Conforme Greene (2002, p.81), se há multicolinearidade, alguns problemas podem surgir: os coeficientes podem apresentar altos erros padrão e pouca significância, mesmo se o R^2 da regressão for alto. Os sinais esperados podem não ocorrer. O pesquisador deverá, por esse motivo, ser muito cauteloso nas considerações baseadas em análise de regressão na presença de multicolinearidade (CHATTERJEE e HADI, 2006 p. 222).

Investigar a multicolinearidade envolve o valor do R^2 que resulta da regressão de cada variável explicativa contra as outras. A relação entre as variáveis explicativas pode ser julgada pela VIF (*variance inflation factor*).

A VIF para uma variável explicativa X é calculada da seguinte forma:

a) Sendo R^2 o quadrado do coeficiente múltiplo de correlação, que é obtido quando a variável X é regredida contra as outras variáveis explicativas,

$VIF_j = X/ 1 - R^2_j$

onde j = 1.....p (p é o nº de variáveis explicativas)

Se X apresentou uma relação linear forte com as outras variáveis, então R^2_j apresentou um valor próximo a um e, portanto, o VIF seria muito grande. Valores de VIF maiores do que dez são indicativos da presença de multicolinearidade (CHATTERJEE e HADI 2006 p.236).

Uma regressão deve ser guiada pelo principio da parcimônia, (CHATTERJEE e HADI, 2006, p.69). Nesta tese foram poucas as variáveis utilizadas (VR, Andar, Idade, Posição, Área) e, portanto, seria pouco provável, com estas variáveis explicativas, a existência de multicolinearidade. Os testes confirmaram a suposição.

3.6.7. Autocorrelação

O erro de uma observação qualquer não deve ser influenciado pelo erro de outra. Se existir tal influência, há uma autocorrelação e, embora os estimadores ainda sejam lineares, sem viés e consistentes, não mais terão a mínima variância (GUJARATI, 2005, p.411).

A autocorrelação entre os resíduos pode ser detectada pelo método gráfico ou através do teste de Durbin-Watson. O exame visual com o gráfico da dependente contra resíduos padronizados pode identificar a existência de multicolinearidade, quando o teste de Durbin-Watson é inconclusivo.

A má especificação do modelo de regressão, em função de resíduos na forma do modelo ou por exclusão de variáveis independentes importantes para a análise, é uma das causas da autocorrelação. A autocorrelação pode ser verificada pela denominada estatística de Durbin-Watson, em que a hipótese básica é a existência de autocorrelação entre resíduos. O teste de Durbin-Watson compara os resíduos (e) de um período (t) com os resíduos do período anterior (t-1).

A estatística de Durbin-Watson (d) é representada por:

$$d = \sum_{i=2}^{n} (e_i - e_{i-1}) \; / \; \sum_{i=1}^{n} e^2_i$$

A hipótese nula propõe que os erros não são correlacionados e a hipótese alternativa propõe a correlação dos erros.

Sendo r a correlação dos resíduos, d = 2 (1-r). Se os erros não são relacionados, o valor de r é muito pequeno e d se aproxima de 2 (dois).

Com o valor apurado da estatística, a tabela de Savin e White (1977) pode ser usada para uma dada significância, um dado número de observações e de regressores.

A tabela fornece dois valores de críticos de d (d_u e d_l).

Se d está entre:

$d < d_l$ ou $4 - d < d_l$ temos autocorrelação (rejeitando H_0)

$d > d_u$ ou $4 - d > d_u$ não temos autocorrelação (falha em rejeitar H_0)

Nos demais casos, não há uma conclusão e é possível usar um apoio gráfico.

3.6.8. Erros de especificação

Conforme Pace, Barry e Sirmans (1998), o desejo por modelos parcimoniosos é natural. Mas, segundo os autores, os modelos que estudam preços no mercado imobiliário devem sempre levar em conta considerações sobre a localização.

Quanto à questão da omissão de variáveis relevantes, supõe-se que a variável de localização do modelo da tese (VR) pode ser vista como um conjunto de atributos:- localização, segurança, amenidades e proximidade ao comércio. Na determinação de valores tais atributos, provavelmente, foram levados em consideração.

Deve ser considerado como uma limitação do método o período decorrido desde a última atualização da PGV (quase dez anos). Na cidade do Rio de Janeiro, no prazo decorrido, não devem ter ocorrido, contudo, modificações significativas que possam ter prejudicado o modelo. Por exemplo, a estação do metrô de Copacabana foi inaugurada em julho de 1998 e seus efeitos sobre os preços dos apartamentos próximos iniciaram-se anos antes.

Os habitantes da cidade do Rio de Janeiro atribuem problemas de criminalidade à proximidade de comunidades de baixa renda desde a década de 1900 (MATTOS, 2004). As amenidades que poderiam ter sido criadas no período

são praças e intervenções como o metrô, mas essas características já existiam em 1998. Em outras cidades provavelmente, o período de atualização pode ser menor e o problema será diminuído.

Em qualquer situação, se houve alguma intervenção tais como um novo *shopping center* construído nas proximidades ou novas praças, a influência desses novos fatores, transformaram algumas observações em *outliers* que, provavelmente, foram eliminados.

Sendo o objetivo principal da tese a construção de um índice e não a avaliação de imóveis, considerou-se que o modelo não foi prejudicado.

Outra variável sempre presente na literatura (especialmente em estudos nos EUA), o número de quartos, que a base de dados do ITBI não fornece, é considerada implícita na área privativa. Um apartamento de 100m² muito provavelmente terá três quartos. Mas um casal sem filhos pode ter anexado um dos quartos à sala e na percepção de um futuro comprador, o apartamento seria apenas considerado um bom apartamento de dois quartos. Na formação do preço da unidade, o que realmente conta é a área privativa. É comum o aparecimento de anúncios em jornais com o texto que se segue, indicando que o número de quartos não pode ser levado em consideração:

- "*Vende-se apartamento com sala, 2 quartos (original 3 quartos).....*"

Considerou-se assim que, embora não disponíveis na base de dados do ITBI, as variáveis número de quartos e segurança estão representadas indiretamente na área privativa e no VR.

A variável "distância ao centro de comércio" já não é uma unanimidade, pelo contrário, estudos em cidades americanas (MATHEWS; TURNBULL, 2007 e CLAPP; RODRIGUEZ, 2001) têm concluído pela falta de significância desse atributo. Além disso, como os índices na tese são baseados em representantes de uma região, o comércio é local e a "distância ao CBD" não tem o mesmo sentido.

Mathews e Turnbull (2007), em estudo em um bairro da cidade de Seattle, encontraram evidências de desvalorização de unidades imobiliárias muito perto de comércio e com alguma valorização entre as distancias de 100 a 400 metros, na composição dos efeitos negativos e positivos gerados pelo comércio.

Já Clapp e Rodriguez (2001), em estudo com a base de dados de uma municipalidade (Fairfax,Virginia) de 60.544 transações no período de 1975 à 1992, encontraram evidências de descentralização a partir de 1990. Como os empregos passaram a se distribuir ao longo das estradas (valor comercial mais acessível), o valor dos terrenos indicaria uma desvalorização nas proximidades desses sub centros. Tais resultados estão em linha com estudos de que, quando os empregos estão em sub centros, ficaria anulada a teoria de que os preços caem com a maior distancia do CBD.

Presumiu-se, assim, que, quando o modelo não pode ser validado pelos procedimentos já descritos, é necessária uma investigação da existência de alguma variável irrelevante. Quase sempre se trata de *andar*, *idade* ou *posição*. As variáveis, *área* e *VR* são sempre significativas e não devem ser eliminadas.

A variável *andar* pode não ser importante em locais onde os prédios só podem ter até três pavimentos (ex.: Ilha do Governador). *Posição* pode não ser uma variável significativa em lugares onde estar voltado para o logradouro é irrelevante (conjuntos habitacionais).

A variável *idade* pode não ser significativa em regiões onde as construções de prédios sejam recentes (ex. Campo Grande, Bangu, Barra da Tijuca), ou onde a oferta seja muito pequena (ex. Ipanema, Leblon, Copacabana).

3.6.8.1. Variáveis omitidas

Conforme Hair et al (2006, p.147), um erro de especificação com a omissão de variáveis relevantes pode causar os seguintes problemas, de acordo com a existência de correlação com as variáveis independentes incluídas:

a) Variáveis omitidas não - correlacionadas com as variáveis incluídas.

- O efeito (se houver) é reduzir a precisão preditiva da análise

b) Variáveis omitidas são correlacionadas com as incluídas.

- Os efeitos das variáveis incluídas são tendenciosos. Quanto maior a correlação, maior a tendência

Entretanto, a inclusão de variáveis irrelevantes, embora não provoquem tendências nos resultados, podem não só tornar os testes das variáveis independentes menos precisos, mas também reduzir a significância estatística e prática da análise (HAIR et al, 2006 p.147), e a parcimônia do modelo, o que pode afetar a interpretação de resultados. Outro tipo de erro que deve ser levado em consideração é o erro de medida, referente ao grau de precisão, especialmente se a variável dependente for uma medida consistente do conceito em estudo.

3.6.8.2. Seleção de variáveis

Andersson (2000) pesquisou sobre testes de variáveis explanatórias, visando sua utilização ou rejeição. Propôs que a intuição do pesquisador é muito importante, tanto na seleção inicial das variáveis, quanto na determinação daquelas que serão retiradas do modelo. O autor utilizou a base de dados do mercado de apartamentos de Cingapura e propôs uma nova abordagem na seleção de variáveis independentes para as funções de preços hedônicos no mercado imobiliário. Concluiu, assim, que a adição de mais variáveis, com valores pequenos de teste t, acrescentou pouco valor explanatório.

Também Chatterjee e Hadi (2006 p.12) propuseram que a forma inicial do modelo deve ser estabelecida, inicialmente, por especialistas na área em estudo, baseada nos seus conhecimentos. O autor sugere ainda que o modelo hipotético deve ser confirmado ou rejeitado pela análise dos dados coletados.

3.6.8.3. Teste da especificação: *linktest*

Uma regressão múltipla apresenta uma especificação incorreta quando não suporta uma boa relação entre a variável dependente e as variáveis explicativas observadas (WOOLDRIDGE 2003, p.278). Uma ferramenta para detecção de problemas com a forma funcional proposta é o teste F. Um dos testes comumente usados para detecção de problemas de especificação é o RESET (*regression specification error test*) conforme Ramsey (1969). Mas Wooldridge (2003, p.283) afirmou ter demonstrado que o RESET falha em detectar omissão de variáveis quando existem expectativas de serem lineares algumas das variáveis

independentes incluídas no modelo. Wooldridge (2003, p.283) considera o RESET apenas um simples teste da forma funcional. Nenhum gráfico *gladder,* ou o teste *"ladder of powers",* apresentou indicações da necessidade do uso de alguma variável quadrática.

Já o *linktest,* que é um dos testes proposto pelo pacote estatístico utilizado (STATA), é baseado na idéia de que, se uma regressão está corretamente especificada, não será possível, exceto por sorte, achar qualquer outra variável independente que seja significativa. O *linktest* cria duas novas variáveis, uma delas denominada "*hat*" (variável de previsão) e a outra denominada "*hatsq*" (o quadrado da variável de previsão). O modelo é refeito com essas duas variáveis independentes. A variável "*hat*" deve ser significativa, pois é o valor previsto. Já "*hatsq*" não deve ser significativa, pois, se o modelo está corretamente especificado, as previsões ao quadrado não devem ter poder de explicação significativo. Assim, o valor do *p-value* de *hatsq* deve ser maior do que 0,05 para que a regressão seja considerada corretamente especificada.

O *linktest* pode inclusive ser usado como teste para variáveis omitidas. (DESBORDES e VAUDAY, 2007). Como suporte poderia ser utilizado um gráfico (STATA, *rvpplot*) dos resíduos x valores da variável dependente, que, ao apresentar padrões, pode significar uma especificação ruim (BAUM, 2006 p.124)

Gujarati (2005 p. 462) estabelece um limite entre testes de significância, tendo em mente um modelo específico, testes e a construção de um modelo, iterativamente incluindo variáveis, testando novamente e assim por diante. Essa última prática, denominada mineração de dados, é condenada. O modelo deve ser construído e guiado pela teoria.

3.6.9.
Roteiro para validação das regressões.

a) Forma Funcional

Seja a regressão $\mathbf{Aval = \beta_0 + Idade\beta_1 + VR\beta_2 + posição\beta_3 + Andar\beta_4 + Área\beta_5}$

Para cada variável independente (X): *ladder of powers* e *gladder* (gráficos com as transformações)

Transformar as variáveis, conferir com *pnorm e qnorm* (no caso de alguma dúvida). Em caso de dúvida, usar a transformação logarítimica.

Checar a linearidade após a transformação, pelos gráficos da variável dependente x cada variável independente (*avplots*). Em dúvida, usar *acprplot X , lowess*.

Rodar a regressão para análise dos testes (F-test), p-value, R^2.

b) Pressupostos Para a Regressão Múltipla (depois de transformada)

Remoção Justificada

Normalidade dos resíduos

Heterocedasticidade

Multicolinearidade

Autocorrelação

Especificação

c) Remoção Justificada

- **RStudent** comando: predict r, rstudent . Listar quais resíduos estudentizados

excedem +3 ou –3 (pontos não usuais). Justificar a remoção.

- **Leverage** comando **:** predict lev, leverage. Os pontos com mais de (2k +2)/n ,

onde k = preditores n = observações , devem ser examinados. Justificar a remoção.

- **Severe Outliers**: é necessário eliminar os *severe outliers*. Comandos *predict y, resid* e *iqr y*. O comando *iqr y* fornece os limites para eliminação (*outfence*).

- **Distância de Cook** , comando predict d , cooksd. Listar pontos com d> 4/n

- **Dfbetas,** comando DFBETAs **.** Listar pontos com DFVInd NID Endereço if abs(DFVInd)> $2/\sqrt{n}$

A justificativa de uma eliminação é necessária pela NBR. Em Ipanema, Leblon e Copacabana, por exemplo, é justificável a eliminação das transações de apartamentos na praia ou na área de contato com comunidades, em andares altos (acima do 12º), onde uma vista pode fazer a diferença, e conforme sua área seja menor ou igual a 50m² e maior ou igual a 250m². A *idade* até 2 anos (inclusive)

também justifica uma eliminação, pois pode ser um apartamento nunca habitado que por causa disso foi valorizado.

d) Correção da Planilha

Corrigir a planilha excel, retirando os pontos com justificativa de remoção. Criar, ainda, uma coluna nomeada de DW, que vai indexar em ordem crescente de data da transação. A indexação anterior (NID), por data terá algumas observações removidas. Essa DW irá ser útil para o teste de Durbin-Watson.

e) Normalidade dos Resíduos

A nova regressão (após as observações removidas)

predict r , resid

p norm r ; q norm r

f) Homocedasticidade

Comando: *hettest*

Correção por WLS com pesos analíticos a serem testados ($1/X^2$). É recomendado testar com todos os quatro pesos. Para determinar quem contribui para a heterocedasticidade (chi2 maior que 3,84), explica-se com as características do bairro em questão. Se nenhuma correção apresentar um chi-quadrado menor de 3,84, considerar a causa da heterocedasticidade como de *natureza não conhecida* e nesse caso, manter a OLS , corrigindo para "*robust standard errors*"

Aplicar a correção Huber-White (*robust*). Visualizar, se necessário, com *rvfplot , yline(0).* Se o resultado do teste for significativo, verificar se ainda existem *severe outliers* ou variáveis irrelevantes, que devem ser removidas.

g) Multicolinearidade

Rodar a regressão com o comando VIF (Variance Inflation Factor), que fornece os valores que permitem confirmar ou excluir a possibilidade de multicolinearidade.

h) Autocorrelação

Usar a variável DW, que será o número de observações após a eliminação de pontos. O comando *tsset* vai tratar DW como uma variável de tempo, permitindo o uso do comando *dwstat,* que dá a estatística de Durbin-Watson.

Exame visual com o gráfico da dependente contra resíduos padronizados, se necessário, nos casos de dúvida.

predict r, resid

scatter r logAval

scatter r DW

obs: um apoio gráfico pode ser necessário em resultados de Durbin-Watson abaixo de 1,80.

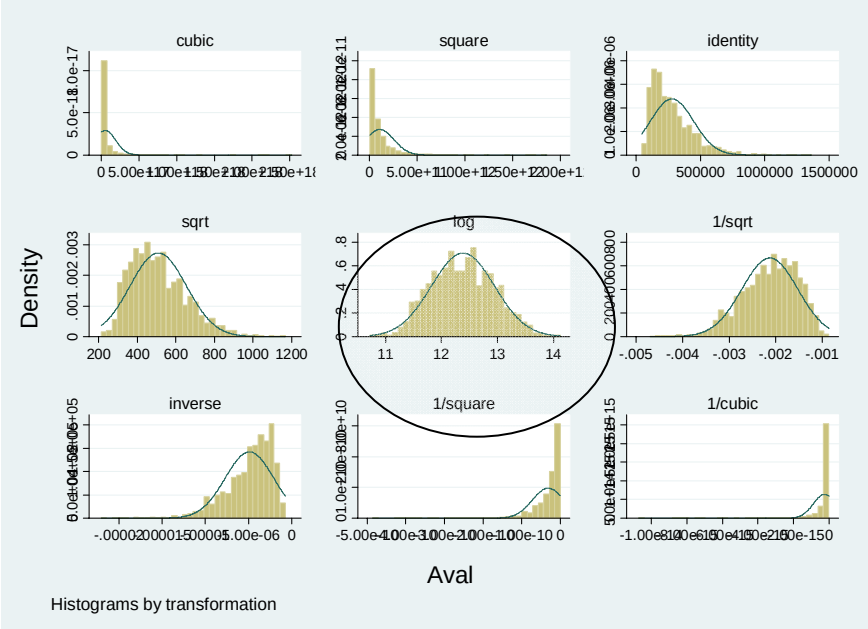
i) Especificação

Para confirmar a especificação usar o *linktest* (um *p-value* significativo para *hatsq* desaprova o modelo). Verificar também a existência de variáveis irrelevantes a serem excluídas. Embora a primeira abordagem da rotina de validação seja utilizar a correção da heterocedasticidade por WLS, adotou-se uma exceção nos casos em que tal abordagem não produziu, além da correção da heterocedasticidade, um *linktest* (teste da especificação do modelo) aceitável. Nesses casos, quando, antes do uso de WLS, o teste de especificação é aceitável, preferiu-se o método de Huber-White (RSE) ao WLS como correção. Vale ressaltar que os resultados do *linktest* não são afetados pela correção com "*robust standard errors*"

3.6.9.1. Validação de uma regressão (um exemplo, "*passo-a-passo*")

Exemplo : RH2 (Ipanema, Leblon, Copacabana no período 1999.4 a 2002.4)

a) Gráficos ***gladder*** para escolha da forma funcional



Transformation	formula	chi2(2)	P(chi2)
cubic	Aval^3	.	.
square	Aval^2	.	.
identity	Aval	.	**0.000**
square root	sqrt(Aval)	.	**0.000**
log	log(Aval)	**31.71**	**0.000**
1/(square root)	1/sqrt(Aval)	.	**0.000**
inverse	1/Aval	.	**0.000**
1/square	1/(Aval^2)	.	.
1/cubic	1/(Aval^3)	.	.

Figura 5- Teste *gladder* (Avaliação)

O gráfico p-norm mostra que os resíduos são normais:

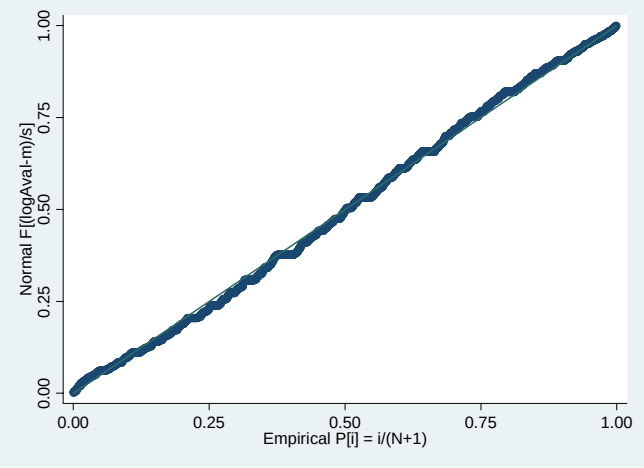
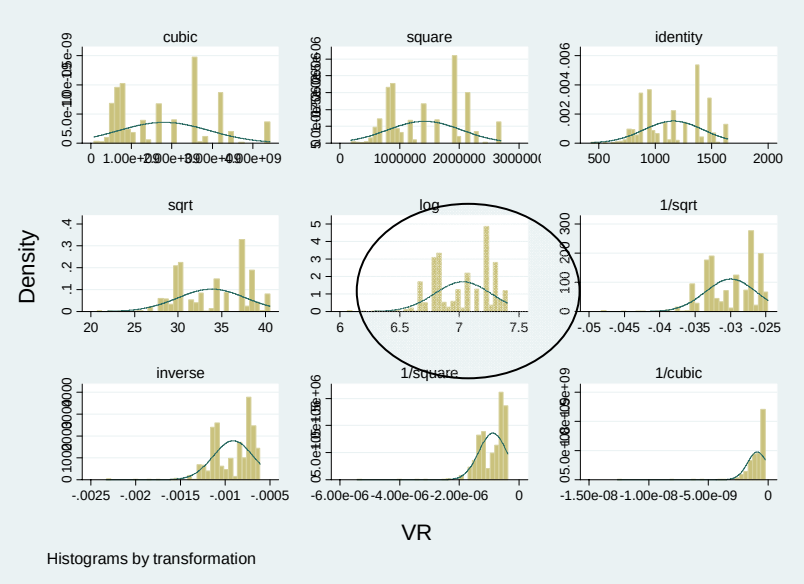


Figura 6- Gráfico *pnorm* (Avaliação)

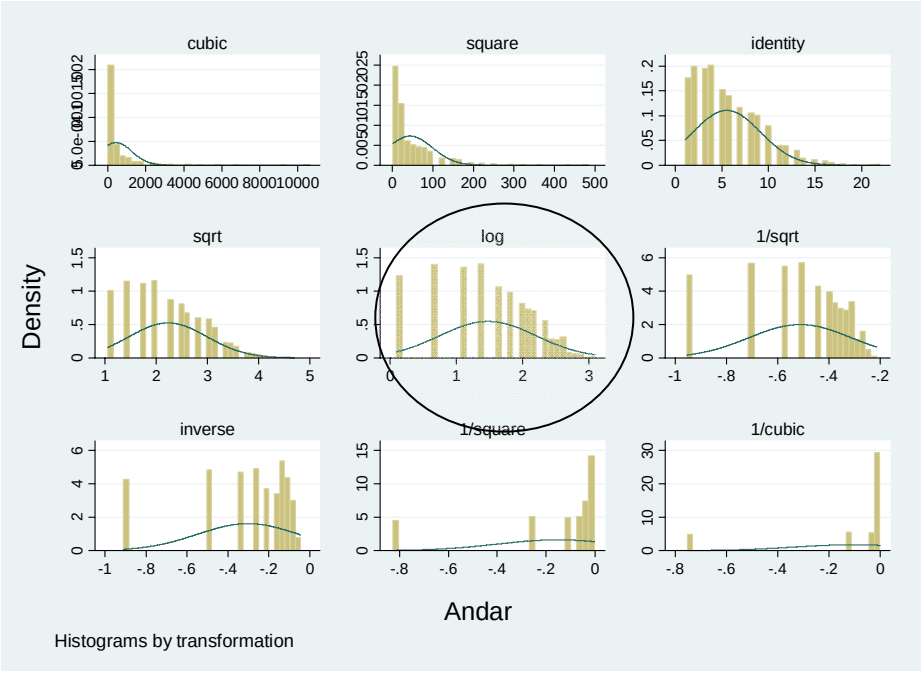
Adotada a forma **logAval**



Transformation	formula	chi2(2)	P(chi2)
cubic	VR^3	.	**0.000**
square	VR^2	.	.
identity	VR	.	.
square root	sqrt(VR)	.	**0.000**
log	log(VR)	.	**0.000**
1/(square root)	1/sqrt(VR)	.	**0.000**
inverse	1/VR	.	**0.000**
1/square	1/(VR^2)	.	.
1/cubic	1/(VR^3)	.	.

Figura 7- Teste *gladder* (VR)

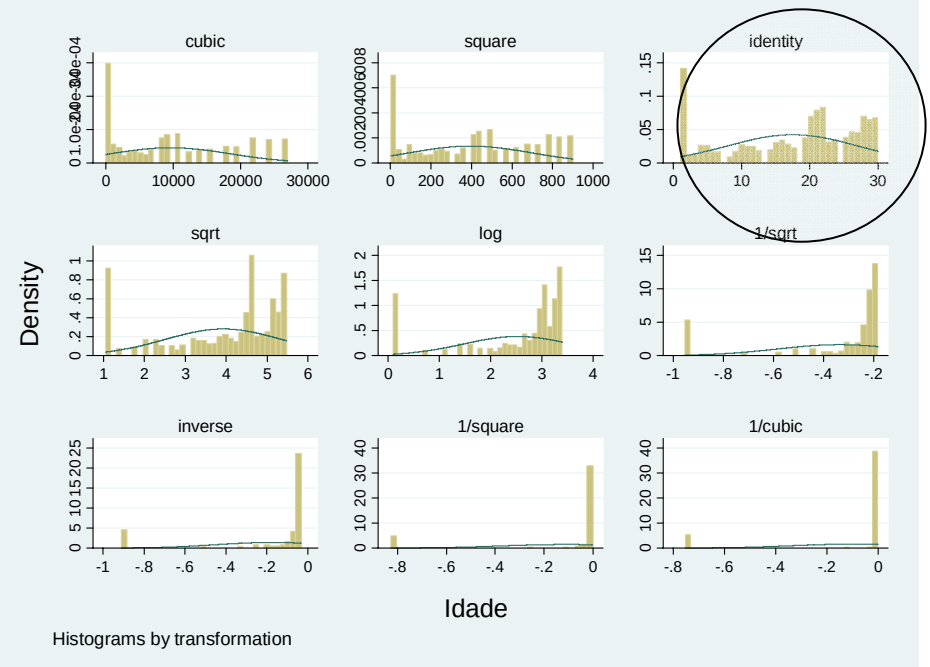
Em dúvida, utilizou-se o *ladder of powers*. Adotada a forma logVR



Transformation	formula	chi2(2)	P(chi2)
cubic	Andar^3	.	.
square	Andar^2	.	.
identity	Andar	.	**0.000**
square root	sqrt(Andar)	.	**0.000**
log	log(Andar)	.	**0.000**
1/(square root)	1/sqrt(Andar)	.	**0.000**
inverse	1/Andar	.	**0.000**
1/square	1/(Andar^2)	.	**0.000**
1/cubic	1/(Andar^3)	.	**0.000**

Figura 8- Teste *gladder* (Andar)

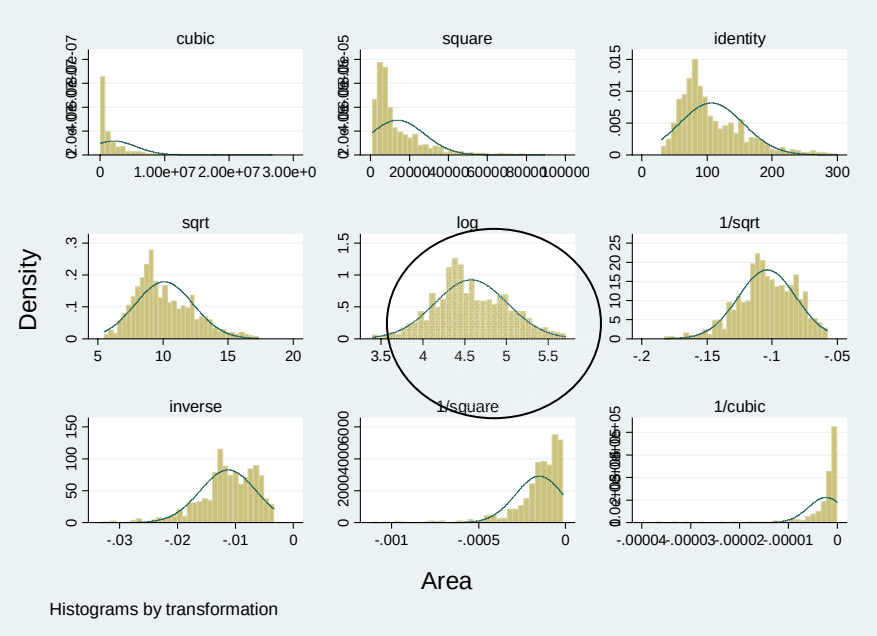
Adotada a forma logAndar



Transformation	formula	chi2(2)	P(chi2)
cubic	Idade^3	.	**0.000**
square	Idade^2	.	.
identity	Idade	.	**0.000**
square root	sqrt(Idade)	.	**0.000**
log	log(Idade)	.	**0.000**
1/(square root)	1/sqrt(Idade)	.	**0.000**
inverse	1/Idade	.	**0.000**
1/square	1/(Idade^2)	.	**0.000**
1/cubic	1/(Idade^3)	.	**0.000**

Figura 9- Teste *gladder* (Idade)

Adotada a própria identidade, Idade.



Transformation	formula	chi2(2)	P(chi2)
cubic	Area__^3	.	.
square	Area__^2	.	**0.000**
identity	Area__	.	**0.000**
square root	sqrt(Area__)	.	**0.000**
log	log(Area__)	**29.44**	**0.000**
1/(square root)	1/sqrt(Area__)	**50.83**	**0.000**
inverse	1/Area__	.	**0.000**
1/square	1/(Area__^2)	.	.
1/cubic	1/(Area__^3)	.	.

Figura 10- Teste *gladder* (Área)

Adotada a forma logArea

Uma primeira regressão foi obtida para exame e remoção dos pontos atípicos.

b) Os resíduos foram examinados quanto aos seguintes processos: resíduos estudentizados, *leverage*, *severe outliers*, distância de Cook, dfbetas. Foram eliminados os pontos atípicos, justificados conforme os critérios da pesquisa.

Uma vez removidos os pontos com justificativa, a regressão final foi obtida.

A regressão conforme o STATA 10 foi:

Source	SS	df	MS	Number of obs =	2720
				F(17, 2702) =	1350.34
Model	145.425314	17	8.55443024	Prob > F =	0.0000
Residual	17.1172087	2702	.006335014	R-squared =	0.8947
				Adj R-squared =	0.8940
Total	162.542523	2719	.059780258	Root MSE =	.07959

logAval	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
logVR	.8507989	.0169521	50.19	0.000	.8175585	.8840393
logAndar	.0210651	.0050335	4.18	0.000	.0111952	.030935
Posição	.0222737	.003421	6.51	0.000	.0155656	.0289818
Idade	-.0028821	.0001684	-17.11	0.000	-.0032124	-.0025518
logArea	.9287358	.0092694	100.19	0.000	.91056	.9469116
var20	.0256822	.0082368	3.12	0.002	.0095311	.0418333
var21	.0275303	.007843	3.51	0.000	.0121514	.0429092
var22	.0389672	.0078049	4.99	0.000	.023663	.0542714
var23	.045482	.0079313	5.73	0.000	.0299299	.0610341
var24	.0690334	.0081698	8.45	0.000	.0530137	.0850531
var25	.0774606	.008177	9.47	0.000	.0614268	.0934945
var26	.0782859	.00816	9.59	0.000	.0622853	.0942865
var27	.0904565	.0080469	11.24	0.000	.0746778	.1062353
var28	.1104456	.0083911	13.16	0.000	.093992	.1268992
var29	.1241583	.0084574	14.68	0.000	.1075747	.1407419
var30	.1267977	.0075574	16.78	0.000	.1119789	.1416165
var31	.1836143	.0081123	22.63	0.000	.1677073	.1995214
_cons	.8793185	.0496936	17.69	0.000	.7818771	.9767599

Figura 11 – Regressão (STATA)

Todas as variáveis são significativas e todos os sinais são os esperados.

```
. hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of logAval

         chi2( 1)      =    20.01
         Prob > chi2   =   0.0000

. linktest
```

Source	SS	df	MS	Number of obs =	2720
				F(2, 2717) =	11552.99
Model	145.440387	2	72.7201935	Prob > F =	0.0000
Residual	17.1021359	2717	.006294492	R-squared =	0.8948
				Adj R-squared =	0.8947
Total	162.542523	2719	.059780258	Root MSE =	.07934

logAval	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
_hat	1.393443	.254347	5.48	0.000	.8947103	1.892176
_hatsq	-.0364685	.0235677	-1.55	0.122	-.0826808	.0097439
_cons	-1.059216	.6854312	-1.55	0.122	-2.403235	.2848036

Figura 12- Testes *hettest e linktest*

O teste *hettest* indica a ocorrência de heterocedasticidade, apresentando um *chi2* de 20,01, acima do desejável (menor do que 3,84). Foi necessária uma correção da heterocedasticidade. Conforme a rotina, utilizou-se uma regressão WLS.

Foram calculados os seguintes pesos (STATA, comando *gen*):

gen **invIdade** = (1/Idade)^2

gen **invlogAndar** = (1/logAndar)^2

gen **invlogArea** = (1/logArea)^2

gen **invlogVR** = (1/logVR)^2

Nenhuma das quatro regressões (WLS) com os pesos acima obteve um chi2 menor de 3,84. Deve-se, por este motivo, corrigir com o método de Huber-White.

Aplicando o método de Huber-White, obteve-se a regressão a seguir com *robust standard errors* (R.S.E). Aplicado o *linktest,* foi constatada a correta especificação.

O R^2 ajustado (0,8947), o teste F, os testes t, os VIF e o teste de Durbin-Watson foram todos satisfatórios. A regressão foi considerada válida e os coeficientes das variáveis dicotômicas, com o primeiro período igualado a zero, foram utilizados para a construção do índice regional.

logAval	Coef.	Robust HC3 Std. Err.	t	P>|t|	[95% Conf.	Interval]
logVR	.8507989	.0183924	46.26	0.000	.8147344	.8868635
logAndar	.0210651	.0049817	4.23	0.000	.0112968	.0308333
Posição	.0222737	.0035414	6.29	0.000	.0153295	.0292179
Idade	-.0028821	.0001561	-18.46	0.000	-.0031882	-.0025761
logArea	.9287358	.0089244	104.07	0.000	.9112364	.9462351
var20	.0256822	.0060804	4.22	0.000	.0137595	.0376049
var21	.0275303	.0058937	4.67	0.000	.0159736	.039087
var22	.0389672	.0065904	5.91	0.000	.0260444	.05189
var23	.045482	.0068697	6.62	0.000	.0320117	.0589523
var24	.0690334	.0068038	10.15	0.000	.0556921	.0823747
var25	.0774606	.0074635	10.38	0.000	.0628259	.0920954
var26	.0782859	.0074411	10.52	0.000	.0636951	.0928767
var27	.0904565	.0076038	11.90	0.000	.0755468	.1053663
var28	.1104456	.0073432	15.04	0.000	.0960467	.1248445
var29	.1241583	.0081449	15.24	0.000	.1081873	.1401292
var30	.1267977	.0068761	18.44	0.000	.1133148	.1402806
var31	.1836143	.0083544	21.98	0.000	.1672327	.199996
_cons	.8793185	.0539721	16.29	0.000	.7734876	.9851494

```
. hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of logAval

         chi2( 1)      =    20.01
         Prob > chi2   =   0.0000
```

```
. linktest
```

Source	SS	df	MS		
Model	145.440387	2	72.7201935	Number of obs =	2720
Residual	17.1021359	2717	.006294492	F(2, 2717) =	11552.99
				Prob > F =	0.0000
Total	162.542523	2719	.059780258	R-squared =	0.8948
				Adj R-squared =	0.8947
				Root MSE =	.07934

logAval	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
_hat	1.393443	.254347	5.48	0.000	.8947103	1.892176
_hatsq	-.0364685	.0235677	-1.55	0.122	-.0826808	.0097439
_cons	-1.059216	.6854312	-1.55	0.122	-2.403235	.2848036

```
. vif
```

Variable	VIF	1/VIF
var30	2.31	0.433398
var22	2.10	0.475243
var21	2.08	0.479688
var23	2.03	0.491929
var27	2.00	0.500488
var26	1.97	0.508773
var31	1.95	0.512441
var25	1.93	0.518494
var24	1.93	0.519412
var20	1.89	0.528352
var28	1.85	0.540860
var29	1.82	0.549688
logArea	1.30	0.771312
logVR	1.29	0.776377
Posição	1.10	0.909828
logAndar	1.09	0.918443
Idade	1.09	0.920443
Mean VIF	1.75	

```
. tsset DW
        time variable:   DW, 1 to 2720
                delta:   1 unit

. dwstat

Durbin-Watson d-statistic(  18,  2720) =  1.903727
```

Figura 13- Correção pelo RSE, linktest, VIF e teste de Durbin-Watson

3.6.10.
Agregação dos índices das RH - obtendo o IMPA

A agregação dos índices das regiões homogêneas foi feita pela média aritmética ponderada, segundo um critério de atribuição de pesos pelo volume negociado em reais no período, para todas as transações de todos os bairros representados em uma RH. Assim, cada região homogênea é responsável por uma contribuição ao índice geral. O método de agregação é similar ao utilizado na construção do INPC, onde são apurados preços em onze regiões metropolitanas do Brasil. A agregação dos onze índices regionais do INPC é feita por média aritmética ponderada, com pesos calculados a partir da população residente em cada estado da região considerada (IBGE, 2006).

A Tabela 6 mostra, em termos percentuais, a participação de cada índice de cada RH em cada período, conforme o volume de reais das transações.

RH	1999-2002	Participa	2002-2005	Participa	2005-2007	Participa
1	5,167	11,4059	1,6234	4,1476	0,2164	0,9571
2	6,607	14,5856	0,9429	2,4090	0,2551	1,1278
3	5,776	12,7512	6,5666	16,7771	4,0972	18,1172
4	3,695	8,1560	6,5259	16,6732	4,2808	18,9290
5	3,924	8,6629	7,3512	18,7817	5,4551	24,1214
6	0,209	0,4612	0,8566	2,1885	0,8095	3,5795
7	1,780	3,9292	2,5945	6,6286	1,4606	6,4585
8	13,209	29,1589	7,1942	18,3805	3,3579	14,8478
9	4,933	10,8891	5,4850	14,0138	2,6830	11,8617
		100,0000		100,0000		100,0000

Tabela 6- Participação de cada RH no índice
Fonte: Elaborado pelo autor

A Tabela 6 mostra a contribuição do índice de cada RH para a construção do IMPA:

	Índice	Participa	Índice	Participa	Índice	Participa	Índice	Participa	Índice	Participa	Índice	Participa	Índice	Participa	Índice	Participa	Índice	Participa	IMPA9902
Trimestre	RH1	%	RH2	%	RH3	%	RH4	%	RH5	%	RH6	%	RH7	%	RH8	%	RH9	%	
1999.4	100,0000	11,4059	100,0000	14,5856	100,0000	12,7512	100,0000	8,1560	100,0000	8,6629	100,0000	0,4612	100,0000	3,9292	100,0000	29,1589	100,0000	10,8891	100,000
2000.1	98,7839	11,4059	106,0919	14,5856	103,3495	12,7512	105,9195	8,1560	98,7747	8,6629	106,2253	0,4612	101,4643	3,9292	103,0367	29,1589	101,7267	10,8891	102,713
2000.2	98,8418	11,4059	106,5443	14,5856	105,3709	12,7512	111,5663	8,1560	97,4801	8,6629	105,0171	0,4612	106,0026	3,9292	105,5838	29,1589	101,2503	10,8891	104,256
2000.3	101,0506	11,4059	109,3874	14,5856	106,3314	12,7512	105,4391	8,1560	101,3209	8,6629	103,9856	0,4612	99,8369	3,9292	103,8610	29,1589	97,5215	10,8891	103,722
2000.4	103,3028	11,4059	111,0355	14,5856	108,0198	12,7512	107,3415	8,1560	100,1535	8,6629	103,3635	0,4612	106,4130	3,9292	106,2724	29,1589	100,4935	10,8891	105,771
2001.1	101,4637	11,4059	117,2286	14,5856	109,5627	12,7512	119,1678	8,1560	99,7026	8,6629	107,5371	0,4612	105,7162	3,9292	110,7193	29,1589	104,9576	10,8891	109,362
2001.2	106,0982	11,4059	119,5255	14,5856	113,2428	12,7512	113,8014	8,1560	98,0526	8,6629	110,2535	0,4612	104,8924	3,9292	111,1167	29,1589	105,6639	10,8891	110,287
2001.3	107,2841	11,4059	119,7529	14,5856	111,5494	12,7512	113,6012	8,1560	102,6688	8,6629	110,2048	0,4612	104,0369	3,9292	109,6887	29,1589	106,1364	10,8891	110,224
2001.4	107,0899	11,4059	123,1563	14,5856	115,1873	12,7512	106,7250	8,1560	104,1786	8,6629	114,2081	0,4612	107,4144	3,9292	111,4529	29,1589	106,1046	10,8891	111,394
2002.1	109,1242	11,4059	128,9572	14,5856	119,7587	12,7512	119,8511	8,1560	113,1246	8,6629	113,8048	0,4612	112,2281	3,9292	116,0422	29,1589	109,1585	10,8891	116,759
2002.2	112,4002	11,4059	133,0939	14,5856	120,0297	12,7512	117,0673	8,1560	112,5745	8,6629	113,6506	0,4612	116,8787	3,9292	117,0035	29,1589	112,1951	10,8891	118,289
2002.3	113,3638	11,4059	133,9053	14,5856	122,0812	12,7512	118,9999	8,1560	113,9828	8,6629	116,1415	0,4612	115,6656	3,9292	118,1179	29,1589	113,2703	10,8891	119,464
2002.4	113,4331	11,4059	152,6210	14,5856	126,1979	12,7512	123,6803	8,1560	113,6766	8,6629	117,1297	0,4612	120,8326	3,9292	118,6975	29,1589	112,5160	10,8891	123,376
Trimestre																			**IMPA0205**
2002.4	100,0000	4,1475	100,0000	2,4090	100,0000	16,7770	100,0000	16,6731	100,0000	18,7817	100,0000	2,1885	100,0000	6,6289	100,0000	18,3805	100,0000	14,0138	100,000
2003.1	103,3836	4,1475	126,7334	2,4090	123,5103	16,7770	122,5968	16,6731	120,6832	18,7817	109,6273	2,1885	121,2854	6,6289	107,4719	18,3805	108,0414	14,0138	116,503
2003.2	103,3794	4,1475	130,2230	2,4090	122,1417	16,7770	120,7752	16,6731	120,6933	18,7817	108,8001	2,1885	122,7480	6,6289	110,4524	18,3805	111,1372	14,0138	117,116
2003.3	103,8341	4,1475	144,8718	2,4090	112,6254	16,7770	119,5300	16,6731	121,4501	18,7817	107,5830	2,1885	124,7639	6,6289	120,0794	18,3805	123,3902	14,0138	119,419
2003.4	106,0235	4,1475	142,6196	2,4090	122,6906	16,7770	122,1640	16,6731	121,1163	18,7817	106,7150	2,1885	120,7306	6,6289	119,8268	18,3805	123,4039	14,0138	121,190
2004.1	111,4036	4,1475	152,2162	2,4090	125,0732	16,7770	132,7752	16,6731	127,6045	18,7817	114,7932	2,1885	130,0506	6,6289	127,9056	18,3805	138,1598	14,0138	129,379
2004.2	111,6153	4,1475	156,5190	2,4090	125,5363	16,7770	133,7654	16,6731	131,1281	18,7817	113,7882	2,1885	132,2302	6,6289	127,8058	18,3805	137,7872	14,0138	130,448
2004.3	111,3576	4,1475	153,7745	2,4090	117,8556	16,7770	133,1794	16,6731	130,3970	18,7817	115,2649	2,1885	127,3964	6,6289	128,0524	18,3805	136,9183	14,0138	128,483
2004.4	111,9757	4,1475	149,4988	2,4090	117,2770	16,7770	132,6593	16,6731	130,7033	18,7817	116,1408	2,1885	129,8608	6,6289	126,4274	18,3805	137,7226	14,0138	128,276
2005.1	117,7499	4,1475	159,8840	2,4090	123,3711	16,7770	135,6691	16,6731	135,6484	18,7817	121,2700	2,1885	132,1716	6,6289	131,9012	18,3805	145,9231	14,0138	133,640
2005.2	116,4930	4,1475	160,3211	2,4090	128,0395	16,7770	135,8206	16,6731	135,8339	18,7817	122,3154	2,1885	133,8528	6,6289	135,1463	18,3805	150,1352	14,0138	135,762
2005.3	114,1020	4,1475	164,7111	2,4090	122,4168	16,7770	135,6980	16,6731	136,1580	18,7817	121,3107	2,1885	133,2390	6,6289	134,2852	18,3805	147,773	14,0138	134,314
2005.4	119,0019	4,1475	161,1913	2,4090	110,7629	16,7770	131,2428	16,6731	130,5637	18,7817	123,0388	2,1885	127,3376	6,6289	135,9159	18,3805	148,4577	14,0138	130,726
Trimestre																			**IMPA0507**
2005.4	100,0000	0,957	100,0000	1,1278	100,0000	18,1171	100,0000	18,929	100,0000	24,1216	100,0000	3,5795	100,0000	6,4585	100,0000	14,8478	100,0000	11,8617	100,000
2006.1	107,4420	0,957	107,0173	1,1278	101,6616	18,1171	104,7750	18,929	103,8517	24,1216	103,9821	3,5795	98,7527	6,4585	103,1332	14,8478	105,3277	11,8617	103,443
2006.2	107,9577	0,957	104,1390	1,1278	102,4187	18,1171	105,2957	18,929	103,4296	24,1216	104,7764	3,5795	101,7968	6,4585	98,2412	14,8478	104,242	11,8617	102,920
2006.3	106,2292	0,957	105,3092	1,1278	102,3484	18,1171	109,6337	18,929	95,9751	24,1216	105,3552	3,5795	102,4819	6,4585	97,1509	14,8478	104,3977	11,8617	101,848
2006.4	107,5212	0,957	103,6889	1,1278	102,3904	18,1171	106,5310	18,929	101,2316	24,1216	108,0289	3,5795	99,8469	6,4585	96,4981	14,8478	105,6103	11,8617	102,503
2007.1	109,7679	0,957	103,2362	1,1278	105,5125	18,1171	107,2396	18,929	101,0045	24,1216	103,5219	3,5795	102,9097	6,4585	98,4585	14,8478	107,0289	11,8617	103,660
2007.2	109,9237	0,957	109,1379	1,1278	105,3588	18,1171	97,4899	18,929	107,4282	24,1216	100,4979	3,5795	99,0729	6,4585	98,6151	14,8478	108,5588	11,8617	103,253

Tabela 7- Contribuição trimestral das RH ao índice
Fonte: Elaborado pelo autor

A Figura 14 representa as etapas de obtenção do IMPA

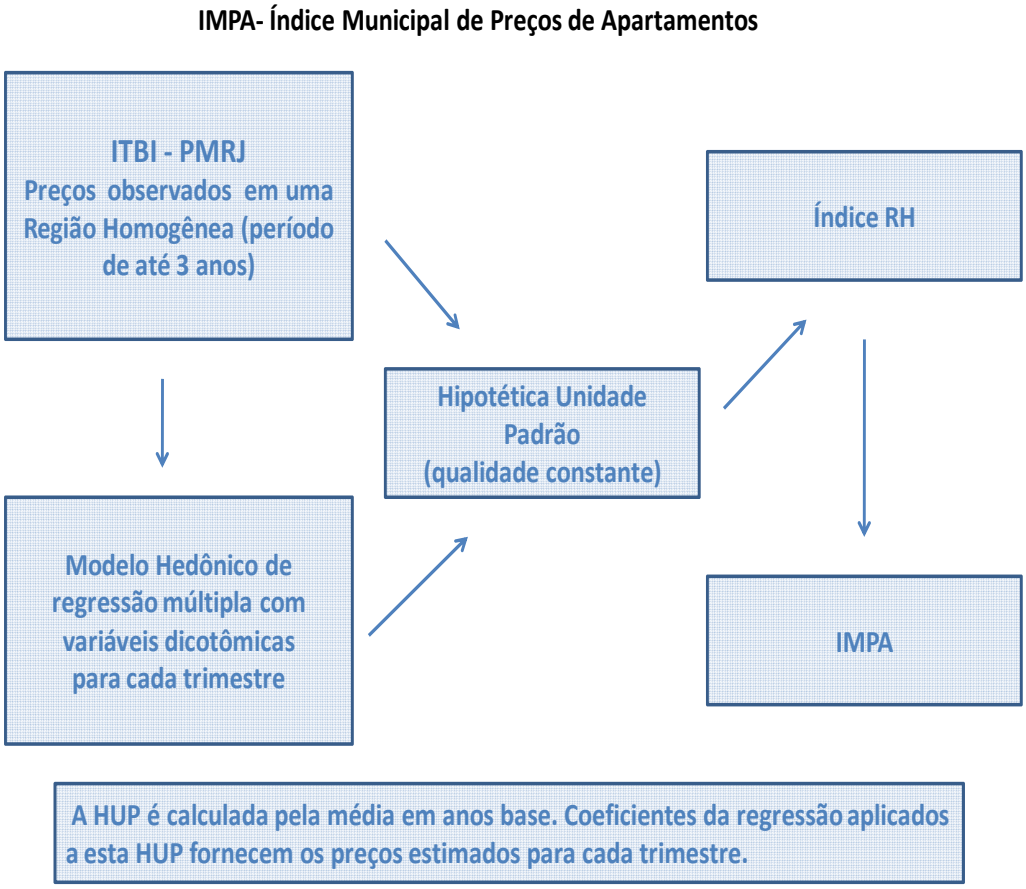


Figura 14- Esquema da obtenção do Impa
Fonte: elaborado pelo autor

3.7. Limitações do método

O objetivo principal foi construir um índice de preços para uma grande cidade. Se também fosse um dos objetivos a avaliação das unidades, além das variáveis já citadas, outras variáveis seriam necessárias para que o modelo obtido pudesse ter essa utilidade. As variáveis seriam aquelas clássicas dos modelos hedônicos: vista, distância ao centro de comércio, distância às amenidades (praia, parques, praças), meio ambiente, segurança, meios de transporte e acessos. O estado da arte seria plotar cada endereço no programa *Google Earth* e, em cada região, verificar a distância a transportes de massa, a centros de comércio

(shoppings), a amenidades (praia, parques). Tais variáveis poderiam contribuir para a construção de um modelo para avaliação de preços. Esta tese limita-se, entretanto, à construção de um índice de preços.

Nessas condições, a proposta foi utilizar apenas aquelas variáveis existentes na base de dados do ITBI e a de localização (VR) da PGV do IPTU. A variável de localização parte do pressuposto lógico de que as ruas que pagam um IPTU maior são aquelas mais valorizadas e mais bem localizadas nos seus bairros.

A planta genérica de valores (PGV) não foi digitalizada nem consta da base do DOM (Diário Oficial do Município) por ter um número excessivo de páginas (386 p.). A PGV, com as páginas digitalizadas a partir do DOM, foi adquirida em uma empresa de serviços digitais (um arquivo por página) e, posteriormente, agrupada em um só arquivo para facilitar a consulta com a utilização de um programa de computador gratuito [4]. Como cada logradouro tem de ser localizado em um documento digitalizado a partir de um *scanner* (com má qualidade de exibição) e digitado posteriormente em uma planilha Excel, não se pode considerar uma precisão absoluta. Além disso, algumas poucas ruas não constam da lei original.

Na ocorrência do logradouro não constar da PGV, arbitrou-se um valor baseado em uma rua próxima e equivalente, depois de visualizada no programa *Google Earth* na forma híbrida (ruas e imagem de satélite).

Não se pode garantir que, a partir de 1997, não tivessem acontecido modificações importantes na urbanização, degradação na vizinhança e outros eventos que tenham influência direta na valorização dos logradouros atingidos após alguma intervenção.

Dentro das limitações apresentadas, os resultados comprovaram que a utilização da base de dados do ITBI e da VR para cada logradouro permitiu a construção do modelo mesmo nos casos em que uma região homogênea foi representada por mais de um bairro. A representação de uma região por mais de um bairro foi adotada quando passou a ser necessária a utilização de um número maior de transações ou ainda quando apenas um bairro não representou, a critério do pesquisador, a totalidade da região.

4- *PDF Split-Merger*, disponível em http://www.verypdf.com

3.7.1.
Credibilidade

A credibilidade das avaliações feitas pelo setor responsável do ITBI pôde ser constatada nas seis reuniões com o gerente do setor, onde se verificou uma extrema preocupação no estabelecimento de preços de mercado para todas as transações em dúvida.

O setor do ITBI responsável pelo acompanhamento dos preços praticados no mercado imobiliário segue os mesmos princípios geralmente adotados em avaliações e as fontes são participantes diretos do mercado: corretoras, construtores e imobiliárias. Um amplo acompanhamento dos jornais com um grande número de anúncios imobiliários é feito semanalmente. Também são acompanhados os lançamentos imobiliários que fornecem preços que podem ser aproveitados como subsídios para uma avaliação.

Mesmo assim, um preço mais alto do que o usual pode ser considerado como preço de mercado. A prefeitura usa o preço declarado, como base de cálculo, mesmo se esse for superior a sua avaliação. Não se considerou que este procedimento possa causar qualquer distorção. Seria uma preocupação, quanto a uma possível lavagem de dinheiro, se não houvesse uma taxação de imposto de renda sobre o lucro imobiliário de 15% (quinze por cento) além dos 2% (dois por cento) do ITBI sobre a base de cálculo.

Como os critérios da PMRJ já são aplicados há mais de vinte anos dentro das técnicas de avaliação de unidades imobiliárias, supõe-se que os preços constantes da base de dados do ITBI expressam a realidade do mercado imobiliário da cidade do Rio de Janeiro para apartamentos prontos.

3.7.2.
Vagas de automóveis e Idade

O número de vagas de cada unidade não é um dado disponível na ficha do ITBI, a não ser que a vaga seja privativa, ou seja, tenha fração ideal e indicação de posição alfanumérica. A maior parte das vagas é de uso comum, de modo que cada unidade tem apenas o direito ao uso de suas vagas, com a localização geralmente designada por sorteio para um determinado período.

A questão principal foi evitar que existissem transações de unidades sem nenhuma vaga de automóvel, o que, em certos bairros, como Copacabana causa uma desvalorização. A regulamentação que trata da obrigatoriedade da existência vagas de garagem apareceu na lei nº. 894 de 22/08/1957, com modificações pelo decreto E nº 3800 de 20/04/1970 e 322 de 03/03/1976. As mudanças obrigaram a vinculação do número de vagas ao número de quartos conforme a região. Supõe-se que as eliminações das transações pelo critério de área, idade e sextos também retiraram qualquer unidade sem vaga de automóvel, bem como aquelas com vagas adicionais.

O início dos anos 1970 foi também a época do chamado milagre econômico brasileiro, com grande expansão das vendas de automóveis. Os lançamentos imobiliários já ofereciam vagas de automóveis adicionais como ação de marketing.

Ao mesmo tempo, outra preocupação da pesquisa foi limitar a idade do imóvel. Não há necessidade de uma licença de obras na reforma de apartamentos, mas geralmente os preços praticados levam em conta as obras que o comprador será obrigado a fazer para habitar a unidade. Desse modo, arbitrou-se em trinta anos a idade máxima das transações para uniformização do critério de idade. Tal procedimento não prejudicou a questão de vagas, uma vez que imóveis de trinta anos em fins de 1999 teriam sido projetados a partir de 1967, dez anos após a lei da obrigatoriedade de vagas.

3.7.3.
A escolha das regiões homogêneas

Laferrère (2003) propôs que o modelo francês reside na divisão em regiões consideradas como áreas de valorização homogênea. Para o pesquisador que tem experiência no mercado imobiliário, a escolha de regiões com valorização homogênea não é uma tarefa difícil. Na cidade do Rio de Janeiro, pode-se utilizar o conceito de Regiões Administrativas (RA), que são um grupamento de bairros por proximidade, em limites definidos pela prefeitura para efeitos de alocação de recursos e de administração.

O Rio de Janeiro, entretanto, tem cento e setenta e sete bairros. Para a pesquisa, optamos por considerar os bairros de maior número de transações como

representantes de uma região homogênea (RH). As RH foram estabelecidas pela localização e pelo conhecimento do mercado imobiliário que o pesquisador possui. Em algumas prefeituras, onde a PGV estiver desatualizada, poderá ser necessária uma consultoria para definição dessas áreas homogêneas.

Foi feito uma divisão da base de dados do ITBI para nove regiões consideradas homogêneas. A Barra da Tijuca, pela sua extensão geográfica e pela diversidade de critérios de parcelamento da terra, foi dividida em duas.

Para cada uma das RH, usou-se o valor monetário de todas as transações de todos seus bairros, e não apenas dos representantes, para a apuração do peso no índice municipal.

3.7.4. Anos-Base

A adoção de uma base de referência com apenas um ano de transações e com revisão a cada três anos, deveu-se ao fato de que o mercado imobiliário no Brasil ainda não atendeu à demanda por apartamentos, sua população é relativamente jovem e ainda cresce. Considerou-se, assim, que os atributos médios das transações ocorridas no primeiro ano de cada período representam razoavelmente uma hipotética unidade padrão.

3.7.5. Inflação

Nos países onde existe uma alta inflação mensal, os preços de um trimestre para outro podem aumentar consideravelmente, levando a dificuldades de interpretação (DIEWERT, 2007). Entendeu-se que a inflação existente no Brasil após 01/10/1999, primeiro trimestre considerado, não causou nenhuma distorção nos preços. Por este motivo, os preços não foram deflacionados, pois não há consenso sobre qual o índice de inflação que deveria ser considerado. Será mais fácil comparar a evolução dos preços com os índices inflacionários escolhidos conforme a necessidade de cada estudo.

3.7.6.
Sazonalidade

Em lançamentos imobiliários, os incorporadores evitam alguns períodos que consideram de difícil atração de compradores, entre 15 de dezembro e 31 de janeiro. Já no mercado de unidades prontas, a comercialização é mais pulverizada. Não foi observada qualquer sazonalidade no mercado de apartamentos prontos.

3.7.7.
Zoneamento

A pesquisa utilizou unidades residenciais de até 30 anos de idade. As cidades brasileiras têm autonomia para estabelecer normas do seu planejamento urbano. Isso significa determinar o aproveitamento máximo de cada lote de terreno disponível para a construção de novos apartamentos. Por vezes, a legislação é modificada, normalmente para diminuir o aproveitamento dos terrenos, tendo em vista a saturação de infra-estrutura dos bairros com maior densidade habitacional.

Na cidade do Rio de Janeiro os bairros já saturados, como por exemplo, Copacabana, Ipanema e Leblon, apresentam prédios com alturas diferentes, refletindo mudanças na legislação.

Quando os bairros estão sem terrenos disponíveis para construção, provavelmente podem ocorrer preços que não levem em consideração a idade, posição ou andar, mas privilegiando como atributos preferenciais a localização e a área. Nesses bairros, em tais condições, os sinais dos coeficientes da idade e andar, que se espera serem positivos, podem eventualmente ser negativos, sem que isso indique a ocorrência de algum problema com os pressupostos de regressão.

Deve ser esperada, entretanto, a ocorrência de pontos atípicos nessas localidades. Por exemplo, em um local de pouca oferta, um prêmio de escassez pode ser adicionado ao preço. Em lugares onde, normalmente, o preço do metro quadrado da área excedente a um determinado padrão de apartamento tenderia a cair pode ocorrer juntamente o contrário. O fenômeno é devido a pouca oferta de unidades com uma grande área, por isso o preço do metro quadrado excedente pode subir.

Todos esses pontos atípicos provavelmente foram eliminados pelos critérios adotados na pesquisa.

3.7.8. Autocorrelação espacial

Gonzalez e Formoso (2000) alertaram para os modelos de regressão que contêm análise de séries de tempo e que podem apresentar relações seriais entre os erros.

Ocorre, contudo, que o modelo não apresenta medidas repetidas sobre o mesmo objeto (preço do imóvel). Desse modo, considerou-se improvável a existência de vendas repetidas a cada trimestre de uma mesma unidade imobiliária.

A correlação espacial no mercado imobiliário é causada pela falta de explicação correta no modelo de regressão das variáveis de preços no espaço (Gonzalez e Formoso, 2000). No modelo, a variável VR, que atribui um valor a cada endereço, resolve tal questão ao introduzir um atributo para cada imóvel em uma pequena região, uma quadra, uma rua.

3.8. *Weighted least squares* (WLS) e *robust standard errors* (RSE)

Quando a presença da heterocedasticidade é confirmada por um teste, por exemplo, o *hettest*, pode-se usar o método do RSE (*robust standard errors*) após o uso da regressão OLS (WILCOX, 1999, p.229; RODRIGUEZ-OREGGIA, RODRIGUES-ROSE,2004; BONNEAU,2007), ou mesmo após uma regressão WLS (WHITE, 1980; GHERTMAN, OBADIA, ARREGLE, 1997; DEVEREUX, 2003) e também após uma regressão do tipo "*pooled cross section*" (LEVITT, 2002; RICHARDSON,LANIS, 2007).

Ghertman, Obadia e Arregle (1997), na presença de heterocedasticidade, propõem o uso de "*weighted least squares*" e, posteriormente, o "*robust standard errors*" para correção da heterocedasticidade, quando a causa é desconhecida. Wilcox (1999, p.229) indica o uso do RSE pelo Stata para correção (robust) da heterocedasticidade comprovada na OLS .

Levitt (2002), em estudo da relação crime e força policial, utilizou "*pooled cross-section*" para períodos anuais de 122 cidades americanas com mais de 100.000 habitantes. Usou, ainda, WLS com pesos pela população de cada cidade e *robust standard errors*. Já Devereux (2003), em estudo das relações entre mudanças no salário e mudanças na oferta de trabalho, usou WLS com "*robust standard errors* ".

Rodriguez-Oreggia e Rodrigues-Rose (2004) usaram o teste de Cook-Weisberg (1983) e, se existisse heterocedasticidade, corrigiam com o método de Huber-White. Richardson e Lanis (2007) usaram OLS *pooled cross-sectional* com *robust standard errors* (*Huber-White Sandwich).*

Bonneau (2007), em estudo cujos dados indicam várias observações de um mesmo estado, mas em anos diferentes, supôs que as observações entre estudos podiam não ser independentes. Para isso, usou "*Huber-White Sandwich*" para a correção dos erros após a OLS.

Não existe, portanto, uma unanimidade sobre a ordem de utilização e cabe, pois, ao pesquisador estabelecer sua rotina para correções.