



IBAPE NACIONAL
Instituto Brasileiro de Avaliações
e Perícias de Engenharia



IBAPE BAHIA
Instituto Brasileiro de
Avaliações e Perícias de
Engenharia da Bahia

Acuracia das Predicoes de Avaliação em Massa de Terrenos Urbanos por Modelos Espaciais

**Sandro Ricardo Vasconcelos Bandeira;
Antonio Augusto Ferreira De Oliveira;
João Freire Prado; Heverton Alves de
souza**



O conteúdo dos trabalhos técnicos apresentados no COBREAP é de inteira responsabilidade de seus autores.

ACURÁCIA DAS PREDIÇÕES DE AVALIAÇÃO EM MASSA DE TERRENOS URBANOS POR MODELOS ESPACIAIS

RESUMO

O presente trabalho se propõe a apresentar uma avaliação em massa de terrenos do Município de Fortaleza com o comparativo de 2 (duas) abordagens para mensurar a dependência espacial dos preços de imóveis observados: a regressão espacial, através de modelos de defasagem espacial e erro espacial, e a geoestatística com a krigagem ordinária dos resíduos da regressão clássica pelos mínimos quadrados ordinários (MQO) e posterior reincorporação no modelo inicial. Ademais, será apresentada uma alternativa de utilização da regressão robusta dos erros MQO dada a presença de heterocedasticidade e autocorrelação serial (HAC) pelo procedimento de Newey-West. Ao final, medir-se-á a performance de cada uma das abordagens sobre uma amostra de teste não participante da modelagem inicial escolhida aleatoriamente dos dados iniciais.

PALAVRAS-CHAVE: *Avaliação em Massa de Terrenos; Mínimos Quadrados Ordinários (MQO); Inferência Robusta (HAC); Regressão Espacial; Krigagem Ordinária.*

1.INTRODUÇÃO

O modelo de preços hedônicos tenta explicar os preços de imóveis de determinado mercado em função de um conjunto de atributos representativos “por parte de quem adquire o bem, que decide desembolsar uma quantia maior por ele” (Fávero, 2007). Essa abordagem conceitual é a fundamentação econômica para a utilização de modelos de regressão, conforme definido no item 3.1.34 da ABNT NBR 14653-1:2019. Segundo aquele autor, quanto mais afetividade existe nos atributos de um bem, sob a ótica dos compradores, maior será o seu “valor hedônico”.

É cediça a lição que os três atributos mais importantes na composição de um bem imobiliário são: “localização, localização e localização”. A localização está presente em duas das três categorias de atributos que formam a estrutura de preços hedônicos imobiliários: qualidade da vizinhança e acessibilidade. O outro componente se refere aos atributos “estruturais”, relacionados as características intrínsecas do imóvel como área, frente etc. (DUBIN, 1992 p. 439).

A norma brasileira de avaliações de bens não define a avaliação em massa. No contexto desse trabalho, a definimos como a criação de um macromodelo para uma grande quantidade de imóveis de determinada área em dado instante de tempo. Esses macromodelos têm fundamental importância para o estudo de plantas de valores genéricos, onde se faz necessário a determinação dos valores venais de uma municipalidade para lançamento dos tributos patrimoniais, tais como o imposto territorial e predial urbano (IPTU) e imposto de transmissão onerosa da propriedade *inter vivos* e cessão de direitos relativos (ITBI), bem como na cobrança de contribuição de melhoria. No caso particular de avaliações em massa de terrenos, objeto de estudo desse artigo, grande utilidade é a correta determinação de seus valores venais para o estabelecimento de políticas públicas relacionadas ao ordenamento territorial com aplicação da contrapartida dos instrumentos urbanísticos de operações urbanas consorciadas (OUC) e outorgas onerosas do direito de construir (OODC). Em ambos instrumentos, há sempre a necessidade de

se estabelecer um preço público sobre o “solo criado” pela flexibilização dos índices de aproveitamentos básicos.

A busca de um macromodelo, ou seja, um modelo único para todos os imóveis de determinada tipologia construtiva em um município de grande porte, tem grandes desafios a serem superados. A experiência nos mostra que a seleção de um conjunto de variáveis explicativas para a explicação dos preços de mercado observados na abordagem hedônica conduz a violação de vários pressupostos da regressão linear clássica pelos mínimos quadrados ordinários (MQO). Seja pela presença de variância não constante dos resíduos (heterocedasticidade), seja pela autocorrelação serial e espacial dos resíduos. Além disso, procura-se que aquele conjunto de variáveis possa explicar de fato toda a heterogeneidade observada, o que leva ao aumento do seu número, violando o princípio da parcimônia, e possíveis problemas de multicolinearidade. Também é comum a utilização de várias variáveis “proxies espaciais”, o que, por si só, já traz um erro de medida devido aos efeitos de transbordamento espacial das fronteiras de onde foram extraídas (DUBIN, 1992 p. 435).

O anexo C da ABNT NBR 14653-2:2011 refere-se aos termos “avaliações em massa” e “planta de valores genéricos”, recomendando o uso do tratamento de dados pela regressão espacial em tais casos, dada a esperada constatação da existência de autocorrelação espacial entre os dados observados.

O presente trabalho tem por objetivo apresentar uma avaliação em massa de terrenos do Município de Fortaleza com o comparativo de 2 (duas) abordagens para mensurar a dependência espacial dos preços de imóveis observados: **1)** regressão espacial utilizando os modelos de defasagem espacial e erro espacial e **2)** geoestatística com o ajuste de um variograma teórico sobre os resíduos da regressão clássica com posterior interpolação dos mesmos pela krigagem ordinária (KO) e reincorporação no modelo inicial para melhorar as estimativas finais. Além disso, será apresentada uma alternativa de utilização da regressão robusta a presença de heterocedasticidade e autocorrelação serial dos erros (HAC). Ao final, medir-se-á a performance de cada uma das abordagens sobre uma amostra de teste não participante da modelagem inicial.

2. METODOLOGIA

2.1. ÁREA DE ESTUDO, DESCRIÇÃO DOS DADOS e SOFTWARES UTILIZADOS

A área de estudo corresponde ao Município de Fortaleza-CE, com área geográfica de 314,93 km². A amostra foi composta de 1.334 dados coletados de janeiro a agosto de 2019 das seguintes fontes de informação (abordagem de 2 dicotômicas agrupadas): a) dados de mercado; b) dados de avaliação de ITBI e c) dados de transação efetiva.

Inicialmente a amostra foi dividida aleatoriamente entre treinamento e teste na proporção de 75% (1.000 dados) e 25% (334 dados), respectivamente. A razão dessa divisão foi medir o desempenho de cada uma das abordagens espaciais na amostra de teste, para aferir a melhor acurácia dentre elas. Depois da análise exploratória dos dados de treinamento, eliminou-se 26 dados espúrios, restando 974 dados para a modelagem. A amostra de teste permaneceu inalterada com 334 dados.

A descrição das 15 (quinze) variáveis utilizadas no modelo está no ANEXO 1. As distribuições espaciais dos dados de amostra de treinamento e teste estão nos mapas do ANEXO 2 e ANEXO 3, respectivamente, onde podemos observar uma boa representatividade espacial delas.

Os seguintes softwares foram utilizados: R com a biblioteca spdep (BIVAND, 2013), Python com a biblioteca PySal 2.0.0 (REY, et al., 2007), eViews (para a correção dos erros MQO à heterocedasticidade e autocorrelação, segundo metodologia de Newey-West) e os pacotes do Geostat desenvolvido pelo prof. Dr. Sidney Rosa Vieira, para cálculos de krigagem e elaboração do semivariograma.

2.2. MODELO DE REGRESSÃO ROBUSTA A PRESENÇA DE HETEROCEDASTICIDADE E AUTOCORRELAÇÃO SERIAL DOS ERROS (HAC)

Neste trabalho foi utilizado o modelo de regressão robusto a presença de heterocedasticidade e autocorrelação serial dos erros (HAC) a ser definido na sequência.

O modelo de regressão linear clássica pode ser definido como:

Equação 1

$$y = X\beta + \varepsilon$$

que na forma matricial é:

Equação 2

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}_{(n \times 1)}, \quad X = \begin{bmatrix} 1 & X_{12} & \cdots & X_{1k} \\ 1 & X_{22} & \cdots & X_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n2} & \cdots & X_{nk} \end{bmatrix}_{(n \times k+1)}, \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}_{(k+1 \times 1)} \quad e \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}_{(n \times 1)}$$

, onde y é o vetor das variáveis dependentes, X é a matriz das variáveis independentes (regressores exógenos não estocásticos), β é o vetor de coeficientes da regressão (ou parâmetros) a serem estimados e ε sendo o vetor de resíduos (ou erros) estocásticos.

Quando os parâmetros do modelo são obtidos pelo MQO, admite-se que os erros são independentes e homocedásticos, com variância constante dada por σ^2 :

Equação 3

$$E[\varepsilon \varepsilon^T | X] = \sigma^2 I$$

Ademais, os erros não dependem das variáveis explicativas. Esse pressuposto também é conhecido como condição de exogeneidade (ARBIA, 2014 p. 2). A contrário senso, quando uma variável independente x_j é correlacionada com o erro, dizemos que ela é uma variável explicativa endógena (WOLLDRIDGE, 2016 p. 76):

Equação 4

$$E[\varepsilon | \mathbf{X}] = 0$$

Sob as hipóteses de Gauss-Markov, o estimador de β pelos mínimos quadrados ordinários é o melhor estimador linear não tendencioso (MELNT, ou BLUE - *Best Linear Unbiased Estimator*) (ALMEIDA, 2012 p. 16).

A multicolinearidade pode ser testada com a matriz de correlação das variáveis quando tomadas duas a duas (com atenção especial a resultados superiores a 0,80) (ABNT, 2011 p. 36) ou pelo cálculo, para cada regressor, do fator de inflação da variância (VIF)¹², definido por:

Equação 5

$$VIF_j = \frac{1}{1 - R_j^2}$$

, onde R_j^2 é o coeficiente de determinação quando se faz a regressão do regressor X_j contra todos os demais regressores. Caso este não esteja correlacionado com os demais, o valor de R_j^2 será 0 (zero) e seu VIF será 1. Por outro lado, quando o regressor X_j se encontrar correlacionado com os demais, seu R_j^2 será alto, diminuindo o denominador da Equação 5 e elevando o seu VIF. Um VIF a partir de 2 (dois) é considerado moderado e a partir de 10, forte (DOANE, 2014 p. 572). Altos valores de VIF podem alargar os intervalos de confiança dos coeficientes da regressão, tornando a estatística t menos confiável. Convém por oportuno e relevante ressaltar, que um VIF de 10 de um regressor não recomenda sua imediata retirada do modelo, sob pena de má especificação do mesmo (Idem p. 572).

Para testar a presença de heterocedasticidade, utilizou-se os testes de Breusch-Pagan³ e Koenker-Bassett⁴. Ambos rejeitaram a hipótese nula de homocedasticidade.

Devida a não homocedasticidade, foram mantidas as estimativas obtidas pelo MQO, mas os erros foram corrigidos pelo procedimento de Newey-West, o qual garante que os mesmos sejam consistentes tanto quanto à heterocedasticidade, quanto à autocorrelação serial (HAC) (GUJARATI, et al., 2011 p. 448). Esse procedimento está presente na maioria dos softwares de econometria (inclusive eviews) e tem como característica a não alteração dos valores dos coeficientes do MQO, mas o cálculo robusto dos seus intervalos de confiança.

¹ Preferimos usar o acrônimo em inglês VIF, variance inflation factor, a FIV (fator de inflação de variância).

² Pode-se também usar o teste de White para heterocedasticidade.

³ O teste de Breusch-Pagan é sensível à condição de normalidade dos resíduos (GREENE, 2012 p. 316), o que nem sempre é observado nas avaliações em massa com grande número de dados na amostra.

⁴ Esse teste realiza uma regressão com a variável dependente sendo o quadrado dos resíduos e a variável dependente sendo os quadrados dos valores estimados de y : $\varepsilon^2 = \beta_0 + \beta_1 \cdot (\hat{y})^2 + u$. A hipótese nula desse teste é o coeficiente β_1 ser zero (condição de homocedasticidade).

Desta forma, depois da correção dos erros padrões dos coeficientes β , caso o nível de significância máximo encontrado seja superior a 10, 20 ou 30%, o regressor pode ser eliminado do modelo a fim de atender o requisitos do item 6 (grau de fundamentação) da Tabela 1 do subitem 9.2.1 da ABNT NBR 14653-2:2011. Isso de fato ocorreu com apenas uma variável (para se atingir o grau de fundamentação II): índice de aproveitamento básico da zona onde estava situado o terreno, de acordo com o plano diretor do Município de Fortaleza.

2.3. ABORDAGENS PARA MENSURAR A DEPENDÊNCIA ESPACIAL

Segundo Almeida (2012, p. 74), existem 2 abordagens principais para mensurar a dependência espacial: a utilização direta da geoestatística (uso de variograma) e a adoção de matrizes de ponderação espacial (**W**) para a regressão espacial que correspondem ao campo da econometria espacial.

A adoção de matrizes para a ponderação espacial é uma tentativa de incorporar o “arranjo espacial das interações resultantes do fenômeno a ser estudado” (Idem, p. 75). Várias são as possibilidades de construção desta matriz, e, no âmbito desse trabalho, adotamos a matriz de distâncias inversas, o que, de certa forma, representa melhor a “Primeira Lei da Geografia de Waldo Tobler”: “todas as coisas estão relacionadas com todas as outras, mas coisas próximas estão mais relacionadas do que coisas distantes”. Ressalte-se, entretanto, a possibilidade de se trabalhar com matriz de contiguidades, principalmente no caso de a variável dependente estar representada por polígonos. Para a avaliação em massa de imóveis, é comum a representação do evento observado por um ponto de coordenada x e y (de preferência no sistema referencial de coordenadas planas UTM, *datum* SIRGAS 2000) situado na frente do imóvel, ou no centroide do lote a que ele pertence.

Desta feita, a matriz de vizinhança é uma matriz quadrada, onde cada elemento w_{ij} representa o inverso da distância do ponto i ao ponto j, com diagonal principal w_{ii} igual a 0 (zero), dada que a distância de um ponto a ele mesmo é nula.

Equação 6

$$\mathbf{W} = \begin{bmatrix} 0 & w_{12} & \cdots & w_{1n} \\ w_{21} & 0 & \cdots & w_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{n1} & w_{n2} & \cdots & 0 \end{bmatrix}_{(n \times n)}$$

Ademais, ponderou-se a linha dessa matriz, cuja soma é 1 (um) e, de tal forma, que o elemento w_{ij} é dado pela expressão:

Equação 7

$$w_{ij}^S = \frac{\left(\frac{1}{d_{ij}}\right)}{\sum_{j=1}^n \left(\frac{1}{d_{ij}}\right)}$$

, onde d_{ij} é a distância entre os elementos i e j.

Dantas (2003, p. 88) definiu o “raio de influência de contágio espacial” entre as observações em estudo a partir do alcance do variograma experimental ajustado aos dados. A ideia é que, se há estacionariedade do variograma a partir da distância do alcance, o poder de influência dos vizinhos (autocorrelação espacial) é desprezível.

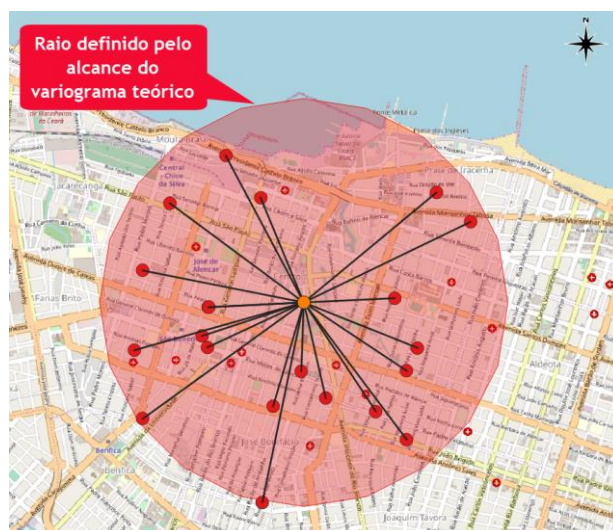


Figura 1 – Esquema de definição da matriz de vizinhança pelo “raio de influência de contágio espacial” obtido pelo alcance do variograma segundo metodologia de Dantas (2003, p. 88). Fonte: elaboração própria.

Nesse trabalho, não se usou essa abordagem. Considerou-se todos os vizinhos, sem raio de influência predeterminado por qualquer análise anterior. Entretanto, recomenda-se fortemente o uso de tal técnica, pois, no caso de avaliação em massa os dados de amostra podem ser bastante numerosos ocasionando um esforço computacional muito custoso e demorado. A adoção de um raio de contágio acarreta em usar matrizes esparsas, atribuindo o valor zero (contribuição desprezível) na distância entre os vizinhos fora do raio, o que torna os cálculos muito mais eficientes computacionalmente.

O diagnóstico de dependência espacial pode ser obtido, inicialmente, pelo Índice de Moran I, que representa um índice global de dependência espacial. Ele tem como hipótese nula a independência espacial (nesse caso assume o valor zero). Valores positivos do índice indicam correlação direta e valores negativos, correlação inversa. Esse teste é realizado sobre os resíduos da regressão clássica pelo MQO⁵. Entretanto, ele não indica qual modelo espacial é o mais adequado (ANSELIN, et al., 2014 p. 122). Formalmente a estatística desse teste é dada por:

Equação 8

$$I = \frac{n}{\sum_{i=1}^n (y_i - \bar{y})^2} \frac{\sum_{i=1}^n \sum_{j=1}^n w_{ij} (y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}^n \sum_{j=1}^n w_{ij}}$$

, onde n é o número de observações, y é a variável estudada, $\bar{y}$ é a sua média e w_{ij} é o elemento na matriz de vizinhança correspondente ao par de elementos i e j. Deve-se observar que se a matriz for normalizada por linha, o termo $\sum_{i=1}^n \sum_{j=1}^n w_{ij}$ é igual a 1. Portanto, o Índice de Moran I fica igual a:

Equação 9

$$I = n \frac{\sum_{i=1}^n \sum_{j=1}^n w_{ij} (y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

⁵ Nada impede que se calcule do índice de Moran I sobre a variável dependente ou mesmo variáveis independentes para uma análise preliminar e exploratória das dependências espaciais.

Os testes de Multiplicadores de Lagrange, LM_ρ e LM_λ , defasagem espacial e erro espacial, respectivamente, têm hipótese nula de independência espacial (ANSELIN, et al., 2014 pp. 104-107). Para a escolha do melhor modelo a ser empregado, adotou-se o proposto por Almeida (2012) e Dantas (2003): a) realiza-se a estimação do modelo MQO; b) verifica-se a presença de autocorrelação espacial; c) caso confirmada, realiza-se os testes LM_ρ e LM_λ (caso contrário, adota-se o modelo MQO); d) o teste que apresentar menor Akaike (critério de informação de Akaike, AIC), será o escolhido⁶ (ABNT NBR 14653-2:2011, item C.3.2.2, p.43).

2.4 REGRESSÃO ESPACIAL UTILIZANDO OS MODELOS DE DEFASAGEM ESPACIAL E ERRO ESPACIAL

O modelo de defasagem espacial ou *Spatial Autoregressive Model* (SAR) é dado pela equação abaixo:

Equação 10

$$Y = \rho W_y + X \cdot \beta + \varepsilon$$

, onde W é a matriz de vizinhança, W_y é um vetor n por 1, resultado da multiplicação da matriz de vizinhança com o vetor dos valores das variáveis dependentes, X é a matriz das variáveis independentes, ρ e β são os parâmetros a serem estimados e ε o vetor de resíduos estocásticos.

Conforme supracitado, a matriz de vizinhança tem diagonal principal zero e foi normalizada na linha. Destarte, a linha i de W_y representa a ponderação dos preços observados dos vizinhos sobre seu próprio valor. O vetor W_y também é conhecido como operador de defasagem espacial ou coeficiente autorregressivo espacial.

De plano, observa-se que temos uma dependência de Y_i (preço) com o preço de seus vizinhos (Y_j), caracterizando a autocorrelação espacial. Fazendo um paralelo com as séries temporais, se na Equação 10, não tivermos as variáveis exógenas representadas pela matriz X , teríamos o modelo autorregressivo, onde a variável dependente Y_i (valor observado na posição espacial i) se relaciona com os valores observados de seus vizinhos (nas séries temporais seria a autocorrelação no tempo t com os períodos anteriores⁷). Por conta disso, esse modelo é conhecido como autorregressivo espacial (SAR).

O modelo de erro espacial ou *Spatial Error Model* (SEM) é expresso formalmente por:

Equação 11

$$Y = X \cdot \beta + u$$

, onde

Equação 12

$$u = \lambda W_u + \varepsilon$$

, onde W é a matriz de vizinhança, W_u é um vetor n por 1, resultado da multiplicação da matriz de vizinhança com o vetor dos resíduos, X é a matriz das variáveis

⁶ Ver também FLUXOGRAMA - REGRESSÃO ESPACIAL no ANEXO .

⁷ Diferentemente da autocorrelação temporal, onde a direção da causalidade ocorre apenas em um sentido (passado influencia o presente e futuro), a autocorrelação espacial pode ocorrer em várias direções do espaço tridimensional (LONGLEY, et al., 2013 pp. 100-101).

independentes, λ e β são os parâmetros a serem estimados e ε o vetor de resíduos estocásticos.

As Equação 11 e Equação 12 podem ser reescritas da seguinte forma após algumas transformações algébricas (I_n é a matriz identidade de ordem n):

Equação 13

$$Y = X \cdot \beta + (I_n - \lambda \cdot W)^{-1} \cdot \varepsilon$$

Diferentemente do modelo anterior, a interação espacial ocorre nos resíduos devido a efeitos não modelados e que não estão distribuídos aleatoriamente no espaço (ALMEIDA, 2012 p. 161). Outra hipótese de autocorrelação espacial nos erros está no fato de utilização de variáveis *proxies* resultantes da divisão artificial das unidades geográficas na área em estudo (DANTAS, 2003 pp. 36-37). Por exemplo, a utilização da variável renda do responsável por setor censitário (dado do IBGE), ou a utilização do índice de desenvolvimento humano por bairro, sem uma devida suavização destas em uma superfície, podem ocasionar erros de medida pelo efeito de transbordamento (a variável se propaga para além da sua fronteira).

Observa-se ainda pela Equação 13 que os termos de dependência espacial estão no último termo da equação, dando a possibilidade de se interpretar cada coeficiente β da maneira “tradicional”, ou seja, como o efeito marginal do respectivo regressor sobre a variável dependente (ALMEIDA, 2012 p. 163).

2.5 GEOESTATÍSTICA – AJUSTE DE SEMIVARIOGRAMA E KRIGAGEM ORDINÁRIA

Quando uma determinada propriedade varia de um local para outro com algum grau de organização ou continuidade, expresso pela dependência espacial, a estatística clássica deve ser abandonada e dar lugar a uma estatística relativamente nova: a geoestatística. (VIEIRA, 2000)

Para verificar a dependência espacial entre os dados utiliza-se o variograma (ou semivariograma). Este pode ser definido como:

[...] sendo a variância da diferença entre uma variável observada em duas regiões diferentes, separadas por uma determinada distância. É suposto que o variograma seja função crescente dessa distância separadora. Existem várias funções crescentes em que o variograma pode se basear, tais como as funções exponenciais, gaussiana, esférica, entre outras. (FOTHERINGHAM et al., 2000; JOHNSTON et al., 2011 apud ALMEIDA, 2012).

Confirmada a dependência espacial, pode-se estimar os demais valores da variável de interesse não amostrados dentro do campo de estudo, sem tendenciosidade e com variância mínima, utilizando-se do método denominado krigagem.

Como é muito comum nos dados imobiliários a presença de efeitos de autocorrelação espacial, faz-se necessário realizar o diagnóstico da dependência espacial através da construção do semivariograma que será utilizado no processo de krigagem (MATHERON, 1965).

A krigagem tem por fundamento a Teoria da Variável Regionalizada (TVR), que identifica que a distribuição espacial de uma variável é composta pela soma de três componentes: uma componente estrutural, tendo média constante ou tendência; uma componente aleatória espacialmente correlacionada, também conhecida de

variação regionalizada e uma componente aleatória não correlacionada espacialmente chamado erro residual.

Sendo assim, o valor observado em uma posição x , representado por uma função aleatória $Z(x)$, fica definida por:

Equação 14

$$Z(x) = m(x) + e'(x) + e''$$

, onde $m(x)$ é uma função determinística descrita pela componente estrutural de Z em x , $e'(x)$ é a variação regionalizada e e'' um resíduo, ou seja, o ruído gaussiano espacialmente independente, com média zero e variância σ^2 .

O semivariograma é, por definição, dado pela equação abaixo:

Equação 15

$$\gamma(h) = \frac{1}{2} E\{[Z(x_i + h) - Z(x_i)]^2\}$$

, onde h é representa um vetor entre dois pontos no espaço. O semivariograma pode ser estimado por:

Equação 16

$$\gamma(h) = \frac{1}{2n} \sum_{i=1}^n [Z(x_i + h) - Z(x_i)]^2$$

À medida que h aumenta, $\gamma(h)$ também aumenta até um valor máximo no qual se estabiliza. Esse valor no qual $\gamma(h)$ se estabiliza chama-se patamar (ou *sill* em inglês), e é aproximadamente igual à variância dos dados $VAR(z)$ (VIEIRA, 2000).

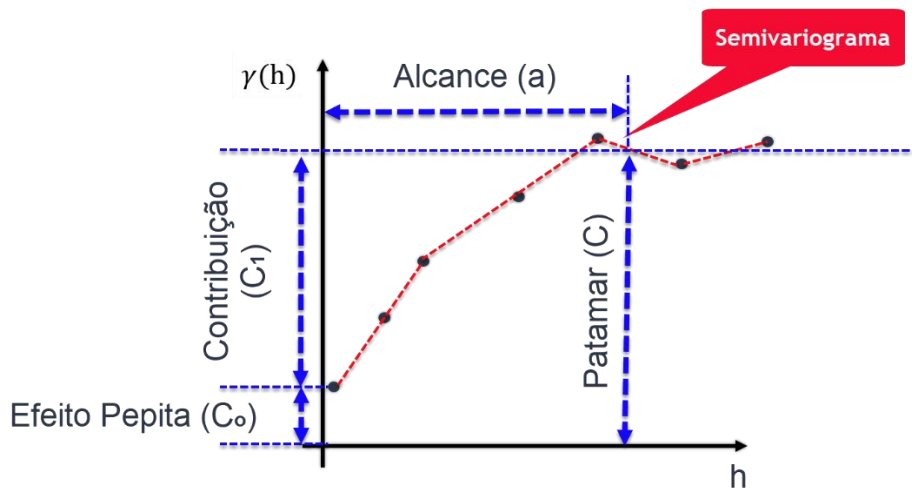


Figura 2 - Exemplo de semivariograma. Fonte: elaboração própria.

A distância na qual $\gamma(h)$ atinge o patamar é chamada de alcance (ou *range* em inglês) e é representada por a na Figura 2, determinando assim a distância limite da dependência espacial. Medições localizadas a distâncias maiores que o alcance a possuem distribuição espacial aleatória e por isso são independentes entre si.

À medida que h tende para 0 (zero), $\gamma(h)$ se aproxima de um valor positivo chamado efeito pepita (*nugget effect*), representado por C_0 . O valor de C_0 revela a descontinuidade do semivariograma para distâncias menores que a menor distância entre as amostras (VIEIRA, 2000). Parte dessa descontinuidade pode ser também devida a erros de medição (Delhomme, 1976 apud VIEIRA, 2000).

Tendo o semivariograma indicado dependência espacial entre as amostras, pode-se interpolar valores em qualquer posição do campo de estudo, sem tendência e com variância mínima. Este método de interpolação chama-se krigagem, nome dado por Matheron (1963) em homenagem ao matemático sul-africano Krige.

2.6 MEDIDAS DE DESEMPENHO

Como medidas de desempenho e acurácia para comparação entre os diversos métodos, as seguintes métricas foram calculadas:

- **Nível de avaliação** (*sales ratio*) mediano (SR_m):

Equação 17

$$SR_m = \text{mediana de } \frac{\text{valor projetado}}{\text{valor observado}}$$

- **Coeficiente de dispersão (COD):**

Equação 18

$$COD = \frac{100}{SR_m} \times \left(\frac{\sum_{i=1}^n |SR_i - SR_m|}{n} \right)$$

onde SR_i é o nível de avaliação de cada terreno individualmente considerado e n é o número total de dados da amostra.

- **Média percentual absoluta do erro (MAPE):**

Equação 19

$$MAPE = \frac{100}{n} \times \sum_{i=1}^n \left| \frac{\text{valor observado}_i - \text{valor projetado}_i}{\text{valor observado}_i} \right|$$

- **Raiz quadrada da média dos erros ao quadrado (RMSE):**

Equação 20

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (\text{valor observado}_i - \text{valor projetado}_i)^2}{n}}$$

- **R^2 :**

Equação 21

$$R^2 = 1 - \frac{\sum_{i=1}^n (\text{valor observado}_i - \text{valor projetado}_i)^2}{\sum_{i=1}^n (\text{valor observado}_i - \text{valor observado}_{\text{médio}})^2}$$

Convém ressaltar que o nível de avaliação e o coeficiente de dispersão são medidas consagradas internacionalmente para aferir performance de avaliações em massa (IAAO, 2010).

Por fim, toda a sequência metodológica utilizada nesse trabalho se encontra no ANEXO 3.

RESULTADOS

REGRESSÃO LINEAR CLÁSSICA (MQO) E PROCEDIMENTO DE NEWEY-WEST (HAC)

A primeira análise relevante realizada foi a escolha e utilização de todas as variáveis consideradas importantes para uma avaliação em massa. Entretanto, os testes de Breusch-Pagan e Koenker-Basset rejeitaram a hipótese nula de homocedasticidade a um nível de significância de 1%. Quando se aplicou a correção dos resíduos MQO para serem robustos à heterocedasticidade e autocorrelação serial, a variável “INDICEAPROVEITAMENTOBASICO” não mais se mostrou significativa a 20%, o que acarretaria o grau de fundamentação I da ABNT NBR 14653-2:2011 (Tabela 1, segunda coluna). Dessa maneira, por opção de uma modelagem que atingisse no mínimo o grau de fundamentação II, retirou-se essa variável e estimou-se novamente o modelo MQO. Novamente, os testes de Breusch-Pagan e Koenker-Basset rejeitaram a hipótese nula de homocedasticidade a um nível de significância de 1%. Entretanto, desta vez, com a correção HAC, todas as variáveis se mostraram significantes a um nível de 5% (Tabela 1, terceira coluna).

Tabela 1 - Resultado da correção dos resíduos MQO pelo procedimento Newey-West em duas estimativas (com e sem a variável “INDICEAPROVEITAMENTOBASICO”. Fonte: elaboração própria.

Nome da Variável	1ª estimativa Coeficiente	2ª estimativa Coeficiente
CONSTANTE	4,559426 (0,0000)	4,634085 (0,0000)
LOTEAMENTO_CONDOMINIO	0,613098 (0,0000)	0,606696 (0,0000)
AVENIDA	0,234526 (0,0000)	0,230246 (0,0000)
ESQUINA	0,074586 (0,0273)	0,071466 (0,0209)
LN_RENDA	0,123396 (0,0000)	0,113735 (0,0000)
LN_AREATERRENO	-0,032769 (0,0392)	-0,035019 (0,0063)
LN_PERCENTUALAREAPRESERVACAO	-0,087525 (0,0010)	-0,092156 (0,0000)
INFRAESTRUTURA_AJUSTADA	0,042417 (0,0111)	0,044015 (0,0039)
INDICEAPROVEITAMENTOBASICO	0,039094 (0,2295)	Variável excluída
DENSIDADECOMERCIALIZACAOTRECHOLOGRADO	0,120755 (0,0316)	0,124030 (0,0094)
LN_VALOR_M2_TERRENO_FACE_QUADRA_IPTU_2014	0,428668 (0,0000)	0,436488 (0,0000)
DISTANCIA_VIA_PRINCIPAL	-0,000281 (0,0000)	-0,000291 (0,0000)
LN_INTERACAO_INCORPORACAO_200	0,010554 (0,0069)	0,010940 (0,0022)
LN_IDH_EDUC	1,653683 (0,0000)	1,821720 (0,0000)

TRANSACAO	0,295852 (0,0000)	0,291739 (0,0000)
OFERTA	0,324088 (0,0000)	0,324137 (0,0000)
Teste Breusch-Pagan	126,349 (0,0000)	126,160 (0,0000)
Teste Koenker-Basset	109,602 (0,0000)	108,750 (0,0000)

Notas: a) *p-value* entre parênteses; b) o prefixo “LN_” antes da variável indica a transformação com logaritmo natural e c) a variável dependente foi transformada com o logaritmo natural.

DIAGNÓSTICO DA MULTICOLINEARIDADE

A matriz de correlação das variáveis independentes se encontra na Figura 3, onde não se observa nenhum valor superior a 0,8.

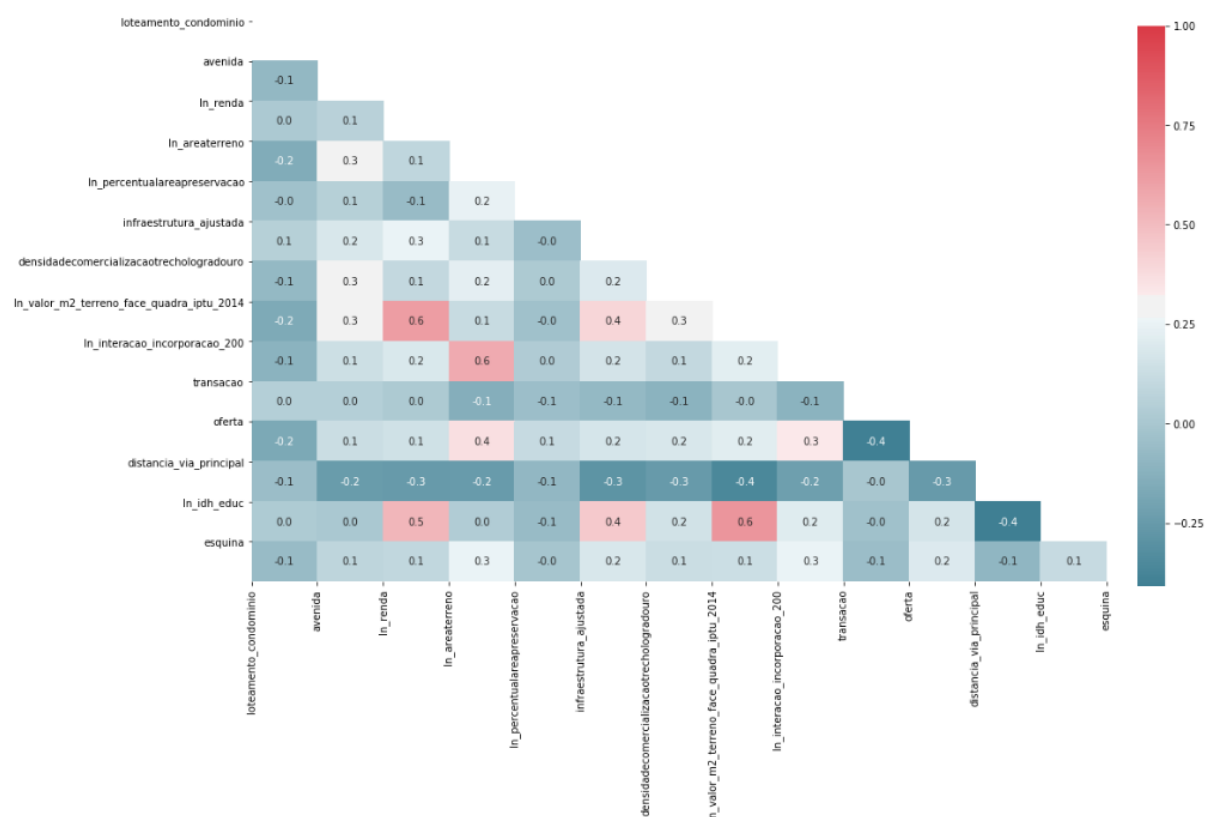


Figura 3 - Matriz de correlação entre as variáveis independentes. Fonte: elaboração própria.

O cálculo dos VIF está disposto na tabela abaixo. Observa-se que, apesar de algumas variáveis apresentarem forte VIF, elas não podem ser retiradas do modelo sob pena de má especificação (ausência de variáveis relevantes). Tome-se como exemplo as variáveis área do terreno e valor do m² de face de quadra de 2014 (esta última apresentou forte significância e alto valor no seu coeficiente).

Tabela 2 - Cálculo e interpretação dos VIF de cada regressor (DOANE, 2014 p. 572). Fonte: elaboração própria.

	Variavel	Fator de Inflação da Variância (VIF)	Interpretação
1	ln_areaterreno	33,20	Forte
2	ln_percentualareapreservacao	30,70	Forte
3	ln_valor_m2_terreno_face_quadra_ipatu_2014	29,50	Forte
4	ln_idh_educ	21,00	Forte
5	infraestrutura_ajustada	15,70	Forte
6	oferta	5,60	Moderada
7	ln_renda	4,90	Moderada
8	distancia_via_principal	2,40	Moderada
9	densidadecomercializacaotrechologradouro	2,10	Moderada
10	avenida	1,60	-
11	ln_interacao_incorporacao_200	1,60	-
12	esquina	1,50	-
13	transacao	1,30	-
14	loteamento_condominio	1,10	-

DIAGNÓSTICO DA DEPENDÊNCIA ESPACIAL (MORAN I e MULPLICADORES DE LAGRANGE)

O primeiro teste realizado de dependência espacial foi o Moran I sobre os resíduos MQO, que, conforme supracitado, tem hipótese nula de não autocorrelação espacial. O índice de Moran I dado pela Equação 9. Importante ressaltar que este índice representa o coeficiente angular da reta de regressão do gráfico de dispersão entre os resíduos MQO e os seus valores defasados espacialmente (também chamado de diagrama de espalhamento de Moran). Esse gráfico tem a interpretação que se segue. Os quadrantes AA (alto-alto) e BB (baixo-baixo) indicam pontos com valores semelhantes, de associação positiva (AA) ou negativa (BB). Já os quadrantes BA (baixo-alto) e AB (alto-baixo) indicam associação negativa (ou contrária), no sentido de que uma localização possui vizinhos com valores distintos.

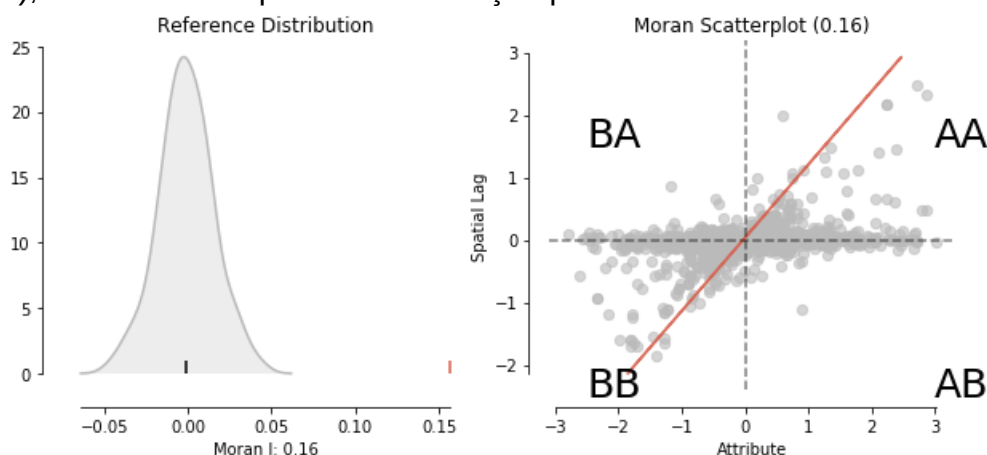


Figura 4 - Resultados do diagnóstico de autocorrelação espacial dos resíduos da regressão linear clássica (MQO) pelo índice de Moran I. Fonte: elaboração própria.

A figura acima ainda indica o valor do índice de Moran I (0,16) e como ele é significativo, pois está bem à extrema direita da distribuição esperada de referência (parte esquerda da Figura 4). O p -value foi menor que $2,2\text{e-}16$, enquanto o valor esperado seria de $-0,0049865542$.

Conforme citado alhures, o índice de Moran I só indica a presença de autocorrelação espacial, mas não indica qual dos modelos espaciais deve ser utilizado. Para isso, realizou-se o teste de Multiplicadores de Lagrange (LM_p e LM_λ) (ANSELIN, et al., 2014 pp. 104-107), cujos resultados se encontram na Tabela 3. Como o modelo de erro espacial apresentou menor Akaike, este foi o modelo escolhido para a projeção final sobre os dados da amostra de teste.

Tabela 3 – Resultado dos testes de multiplicadores de Lagrange para seleção do modelo mais adequado. Fonte: elaboração própria.

	LM_p (defasagem espacial)	LM_λ (erro espacial)
Valor	47,888	80,509
p -value	$4,513\text{e-}12$	$< 2,2\text{e-}16$
Akaike	948,124	908,3743

AJUSTE DO SEMIVARIOGRAMA E KRIGAGEM ORDINÁRIA

No caso em estudo o melhor modelo teórico encontrado foi o exponencial com $C_0=0,00914909$, $C_1=0,13871208$ e $a=394,429489$, cuja equação é a que se segue. A Figura 5 apresenta o seu gráfico.

Equação 22

$$\gamma(h) = C_0 + C_1 \times \left(1 - e^{-3\left(\frac{h}{a}\right)} \right)$$

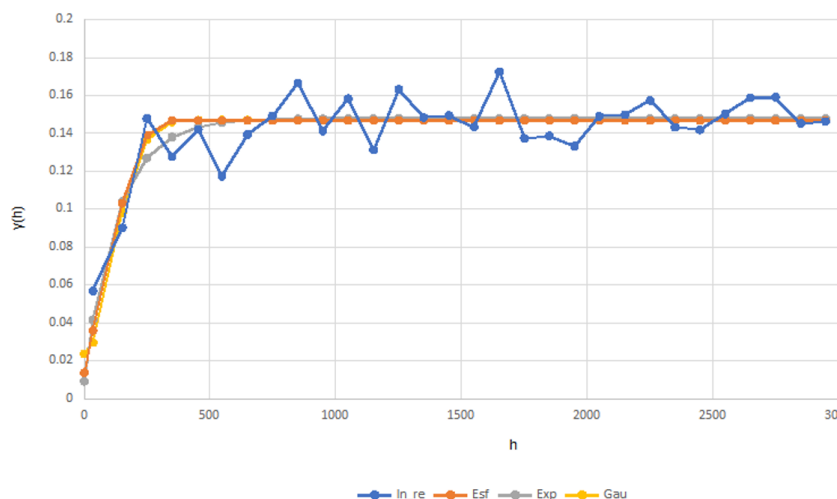


Figura 5 - Semivariograma experimental e teórico dos resíduos MQO. Fonte: elaboração própria.

O efeito pepita indica que apenas $6,18\%^8$ da variabilidade total dos resíduos se deve à componente aleatória, e os $93,82\%$ restantes são explicados pela componente de autocorrelação espacial existente nos resíduos.

⁸ Dado por $((C_0 / (C_0 + C_1)) * 100)$.

O alcance de 394,43m indica que a variância cresce até essa distância entre os terrenos, havendo uma autocorrelação espacial até esta. Distâncias maiores que o alcance tem distribuição espacial aleatória.

Utilizou-se a krigagem ordinária, ou seja, não precisamos ter conhecimento prévio da média. Supondo-se que estimar valores z^* , para qualquer local x_0 , onde não se tem valores medidos, é dado por:

Equação 23

$$Z_{x_0}^* = \sum_{i=1}^n \lambda_i Z(x_i)$$

com $\sum_{i=1}^n \lambda_i = 1$

onde, n é o número de valores medidos $z(x_i)$ e λ_i são pesos associados a cada valor medido $z(x_i)$.

Com os parâmetros do modelo teórico do variograma, efetuou-se o procedimento de auto validação comumente chamado de “Jack-knifing”. Esse procedimento envolve a estimativa de cada ponto medido “fazendo de conta” que ele não existe, durante a sua estimativa. O “Jack-knifing” indicou como sendo 12 vizinhos o que resulta em melhores resultados.

Posteriormente, efetuou-se a krigagem ordinária para toda a área de estudo, considerando uma malha de 50m por 50m e 12 vizinhos em todas as direções. Considerou-se que o sistema é isotrópico.

Para cada dado de teste, através de suas coordenadas, encontrou-se o resíduo krigeado. Esse resíduo foi reintroduzido no modelo, ou seja:

$$Valor_{ajustado} = Valor_{projetado\ MQO} + Resíduo_{krigeado}$$

AFERIÇÃO DE PERFORMANCE SOBRE A AMOSTRA DE TESTE

O resultado do comparativo entre os modelos MQO, de erro espacial (SER) e com a incorporação dos resíduos (MQO) interpolados por krigagem se encontra na Tabela 4. Através dela se observa que o modelo SER apresentou melhor coeficiente de determinação (R^2), menor raiz do erro quadrático médio (RMSE) e menor coeficiente de dispersão (COD). Já o modelo geoestatístico apresentou melhores resultados na média percentual absoluta do erro (MAPE) e menor nível de avaliação. O modelo clássico MQO perdeu em todos os itens de performance. Nenhum modelo atingiu ao critério do IAAO de COD máximo de 20%. Todos os modelos tiveram nível de avaliação superior a 1 (um), o que indica um aumento mediano na projeção em relação ao observado.

Tabela 4 - Resultados comparativos das medidas de performance. Fonte: elaboração própria.

Medida de Performance	MQO Treino	SER Treino	MQO Teste	SER Teste	Krigagem Ordinária (Teste)
R^2	0,65	0,70	0,61	0,66	0,65
RMSE (R\$/m ²)	491,33	455,10	535,96	499,13	509,33
MAPE (%)	32,04	29,04	41,96	39,04	38,56
Nível de avaliação	1,01	1,01	1,03	1,04	1,01
COD	31,80	28,85	40,52	37,45	37,99

CONCLUSÕES E RECOMENDAÇÕES

O presente trabalho desenvolveu um macromodelo de avaliação em massa para se estimar os valores de mercado dos terrenos do Município de Fortaleza.

Analisando os diversos resultados extraídos das respectivas modelagens, verificou-se que os modelos SER e Krigagem Ordinária (KO) possuem um poder de explicação quase que equivalentes. Em relação ao MAPE, a KO teve o melhor resultado. Observa-se que o nível de avaliação mede a proximidade dos valores estimados com o valor observado no mercado, e que o modelo KO apresentou o nível mais próximo de 1, ou seja, o valor estimado está bem próximo do valor observado, indicando assim uma melhor precisão no processo de estimativa da avaliação em massa.

Os resultados entre os modelos SER e KO, no quesito do COD, apresentaram-se muito próximos, ou seja, se o modelo escolhido for o KO, haverá um pequeno aumento na dispersão dos valores estimados.

Considerando o poder de processamento envolvido entre os dois melhores modelos, e por ser a KO um modelo de fácil implementação, bastando apenas uma boa modelagem no processo de regressão linear múltipla e o georreferenciamento dos dados, deve-se analisar a possibilidade de seu uso, sempre que possível.

Verificou-se também que a utilização de todos os dados para a construção da matriz de vizinhança acarretou em um desnecessário tempo de cálculo e grande esforço computacional, mesmo utilizando-se de bibliotecas já disponíveis nas linguagens R e Python.

Recomenda-se a construção de um semivariograma teórico (esférico, exponencial, linear ou gaussiano) para a determinação do raio de contágio espacial (amplitude encontrada), simplificando em muito a construção daquela matriz.

Sugere-se também a divisão aleatória da amostra em dados de treinamento e teste, para se aferir a consistência dos modelos encontrados, através dos parâmetros definidos pelo IAAO e os demais apresentados neste trabalho.

ANEXO 1

VARIÁVEIS

Quadro 1 - Descrição das Variáveis Utilizadas. Fonte: elaboração própria.

Nome da Variável	Descrição
Avenida (avenida)	Variável que assume valor 1(um) se o dado está situado em avenida ou rodovia e zero em caso contrário.
Área do Terreno (areaterreno)	Variável que indica sua área territorial (em m²).
Esquina (esquina)	Variável que indica se o mesmo se encontra ou não em esquina.
Loteamento em Condomínio (loteamento_condominio)	Variável que indica se o terreno se encontra em condomínio fechado.
Renda (renda)	Variável de macrolocalização, representada pela renda média do chefe da família, em salários mínimos, ajustada a uma superfície de tendência construída pelo processo de <i>Krigeagem</i> , tomando-se como base os dados de renda média do responsável no setor censitário, divulgada pelo censo do IBGE (2010).
Oferta (oferta)	Variável representando os dados coletados em OFERTA. Agrupada com TRANSAÇÃO. Quando OFERTA e TRANSAÇÃO são iguais a 0 (ZERO) simultaneamente, indica que o dado é uma avaliação de ITBI.
Transação (transacao)	Variável representando as TRANSAÇÕES do mercado imobiliário ou valor declarado pelo contribuinte nas declarações de ITBI quando essa declaração esteja dentro do limite de mais ou menos 5% do avaliado pela SEFIN. Agrupada com OFERTA. Quando OFERTA e TRANSAÇÃO são iguais a 0 (ZERO) simultaneamente, indica que o dado é uma avaliação de ITBI.
Valor Básico de Terrenos 2014 (valor_m2_terreno_face_quadra_ipu_2014)	Variável indicando o valor unitário (R\$/m²) base do terreno para o lançamento do IPTU, referente ao ano 2014.
Densidade Comercialização no Trecho de Logradouro (densidadecomercializacaotrechologradouro)	Variável de densidade de comercialização no trecho de logradouro onde está situado o imóvel. Representa o percentual de imóveis comerciais em relação ao total de imóveis no trecho de logradouro. Em casos de zero absoluto, para não inviabilizar sua transformação logarítmica, considerou-se 0,01.
Área de Preservação (%) (percentualareapreservacao)	Variável representando a área de preservação (ZPA1) que atinge o imóvel, segundo o plano diretor do Município de Fortaleza (PDPFor) (em m²). Em casos de zero absoluto, para não inviabilizar sua transformação logarítmica, considerou-se 0,01.
Índice de Aproveitamento Máximo Equivalente (indiceaproveitamentomaximo_equivalente)	Variável que representa o índice de aproveitamento máximo equivalente (no caso de mais de uma zona cortando o lote, determina-se a ponderação das áreas) onde está situado o imóvel segundo plano diretor. Em casos de zero absoluto, para não inviabilizar sua transformação logarítmica, considerou-se 0,01.

Nome da Variável	Descrição
Interação Incorporação 200 (interacao_incorporacao_200)	Variável interação: $\text{area_terreno} * \text{ia_maximo_eq} * \text{densidadeverticalizacaokernel_200} * \text{lote_incorporacao}$
Distância à Via Principal (distancia_via_principal)	Variável que mede a mínima distância entre lote e a via mais próxima mais importante. As vias mais importantes do Município foram determinadas estatisticamente e são em número de 260.
Índice de Desenvolvimento Humano Educacional CENSO 2010 (idh_educ)	Variável do índice de desenvolvimento humano do ano de 2010 referente ao indicador <u>educação</u> no <u>setor censitário</u> onde se situa o imóvel.
Valor Unitário (valorunitario)	Variável explicada ou dependente, que pode ser fornecida como o valor total estimado ou sua razão pela área territorial.

ANEXO 2

MAPAS DA DISTRIBUIÇÃO ESPACIAL DA AMOSTRA DE TREINAMENTO E TESTE

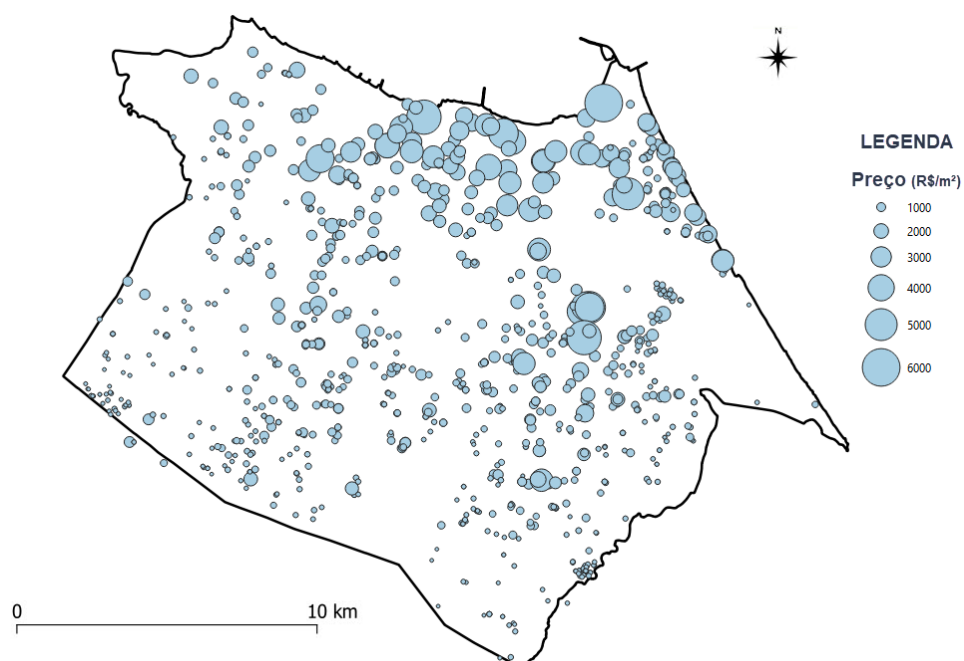


Figura 6 - Distribuição espacial dos 974 dados da amostra de treinamento. Fonte: elaboração própria.

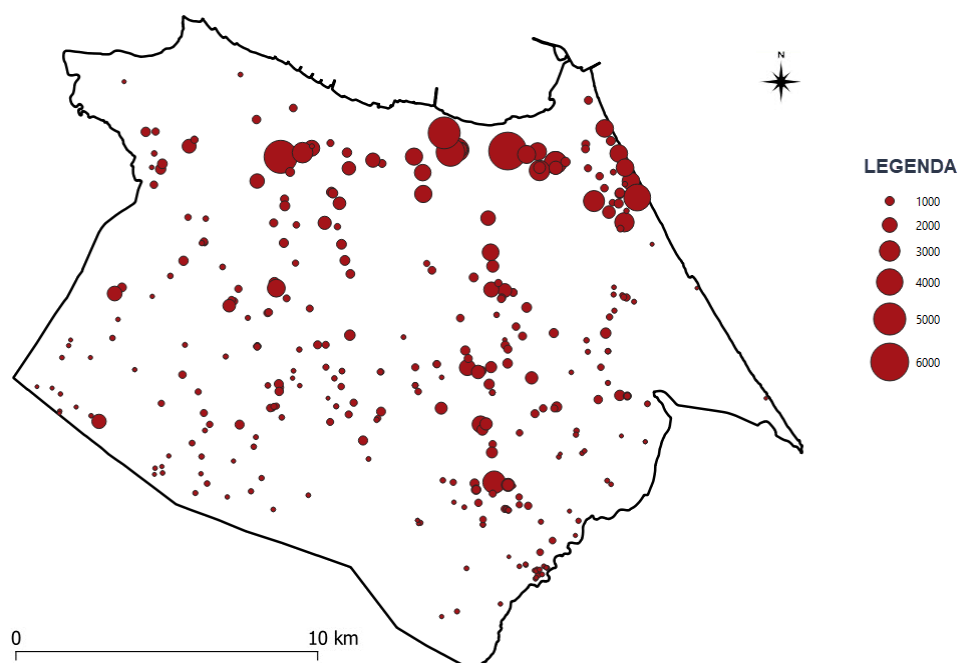


Figura 7 - Distribuição espacial dos 334 dados da amostra de teste. Fonte: elaboração própria

ANEXO 3

FLUXOGRAMA - REGRESSÃO ESPACIAL

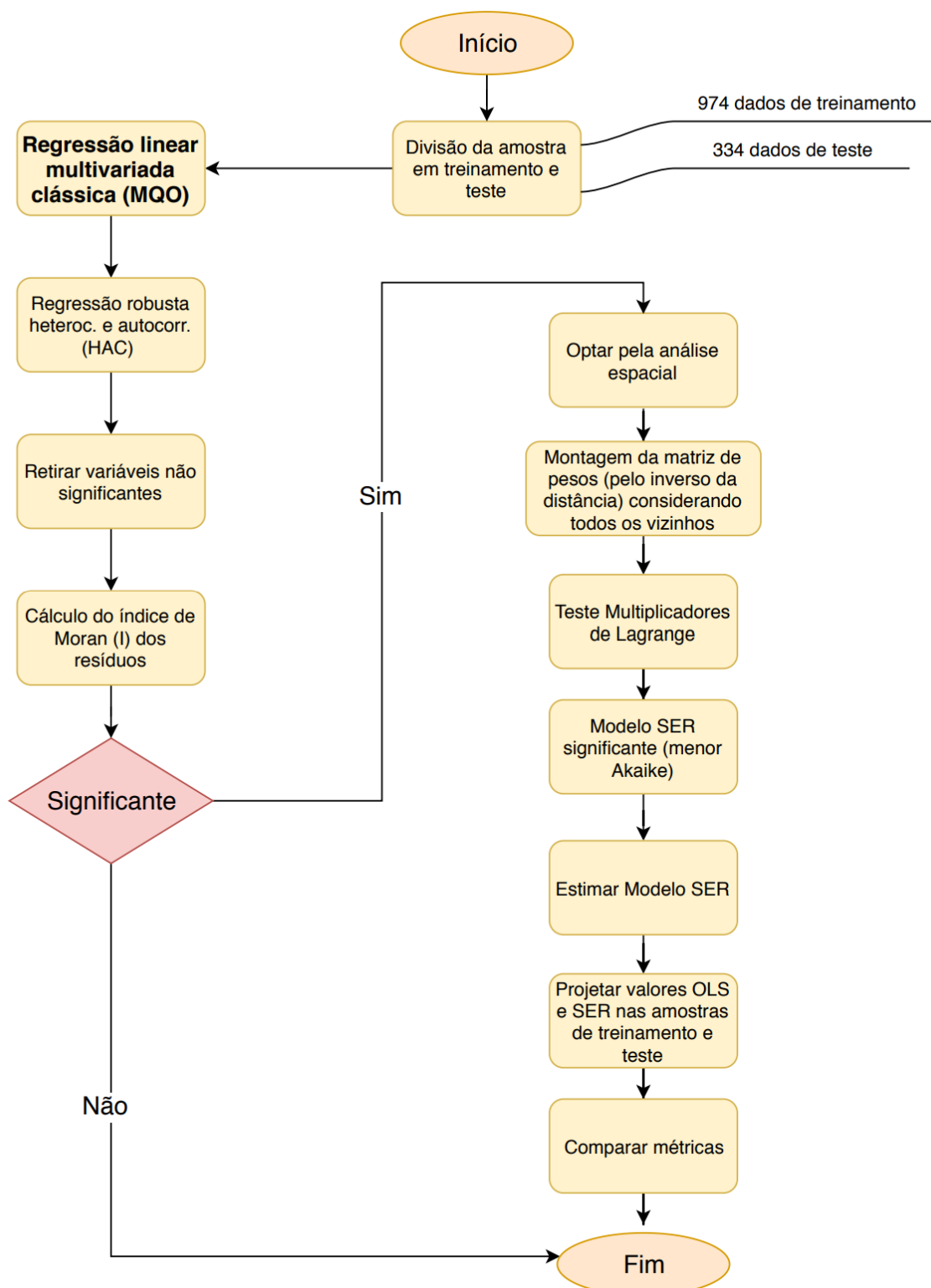


Figura 8 - Fluxograma da regressão espacial. Fonte: elaboração própria adaptado de (ANSELIN, et al., 2014 p. 110) e (ALMEIDA, 2012 p. 231).

ANEXO 4

MAPA DOS RESÍDUOS DA REGRESSÃO CLÁSSICA (MQO)

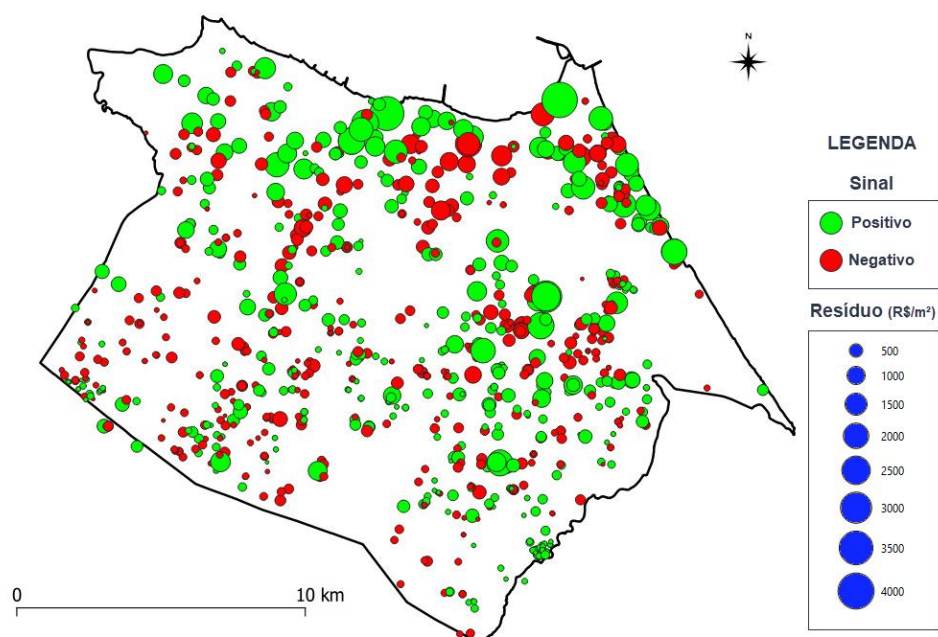


Figura 3 - Resíduos da Regressão Linear Clássica (MQO) – Amostra de treinamento. Fonte: elaboração própria.

BIBLIOGRAFIA

- ABNT. **NBR 14653-1**: avaliação de bens. Rio de Janeiro: [s.n.], 2019.
- _____. 2011. **NBR 14653-2**: avaliação de bens parte 2: imóveis urbanos. Rio de Janeiro: [s.n.], 2011.
- ALMEIDA, Eduardo. **Econometria Espacial Aplicada**. Campinas: Alínea, 2012.
- ANSELIN, Luc e REY, J. Sergio. **Modern Spatial Econometrics in Practice**. Chicago: GeoDa Press LLC, 2014.
- ARBIA, Giusepe. **A Primer for Spatial Econometrics With Applications in R**. London: Palgrave Macmillan, 2014.
- BIVAND, Edzer Pebesma, GOMEZ-RUBIO, Virgilio. **Applied spatial data analysis with R**, 2. ed. New York: Springer, 2013.
- DANTAS, Rubens A. **Modelos espaciais aplicados ao mercado habitacional: um estudo de caso para a cidade do Recife**. 2003. 114f. Tese (Doutorado em Economia) - Universidade Federal de Pernambuco. Recife.
- DOANE, David P. SEWARD, Lori E. **Estatística Aplicada à Administração e Economia**. 4. ed. Porto Alegre: AMGH, 2014.
- DUBIN, Robin A. Spatial autocorrelation and neighborhood quality. **Regional Science and Urban Economics**, 1992, v. 22, p. 433-452.
- FÁVERO, Luiz P. L. **Mercado Imobiliário - Técnica de Precificação e Comercialização**. São Paulo: Saint Paul Editora Ltda., 2007.
- GREENE, William H. **Econometric Analysis**. Boston: Pearson, 2012. 7th.
- GUJARATI, Damodar N. e PORTER, Dawn C. 2011. **Econometria Básica**. 5. ed. Porto Alegre: AMGH Editora Ltda., 2011.
- IAAO, International Association of Assessing Officers. **Standards on Ratio Studies**. Kansas City: [s.n.], 2010.
- LONGLEY, Paul A., et al. **Sistemas e Ciência da Informação Geográfica**. 3. ed. Porto Alegre: Bookman, 2013.
- MATHERON, G. **Les Variables Regionalisées et Leur Estimation**. Paris: Masson, 1965.
- REY, S.J. e ANSELIN, L. PySAL: A Python Library of Spatial Analytical Methods. [S.l.]: **Review of Regional Studies**, 2007. v. 37, p. 5-27.
- VIEIRA, Sidney R. Geoestatística em estudos de variabilidade espacial do solo. In: NOVAIS, R. R. et al (Orgs.). **Tópicos em ciência do solo**. [S.l.]: Sociedade Brasileira de Ciência do Solo, 2000. p. 1-53.
- WOLLDRIIDGE, Jeffrey M. **Introductory Econometrics: A Modern Approach**. 6e. Mason, Ohio: CENGAGE Learning, 2016.