Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Estatística - EST

Análise dos dados do AirBnb usando regressão quantílica

Álvaro Jeronimo da Silva Kothe - 17/0004694

Brasília
2021

Álvaro Jeronimo da Silva Kothe

Análise dos dados do AirBnb usando regressão quantílica

Relatório apresentado para o Departamento de Estatística, Instituto de Ciências Exatas, Universidade de Brasília, como parte dos requisitos necessários para o grau de Bacharel em Estatística.

Orientadora: Prof.ª **Dra. Thais Carvalho Valadares Rodrigues**

Brasília
2021

Resumo

O modelo de regressão quantílica é capaz de modelar o quantil condicional da variável resposta, além de possuir diversas propriedades de invariância e garantir um conhecimento maior da variável resposta que não seria fornecido pela média. Além disso, a regressão quantílica é um ótimo método sobre dados assimétricos por ser mais robusta. Estendendo esse método para dados espaciais tem-se a regressão quantílica espacial Bayesiana. Neste trabalho será utilizado o modelo proposto por Reich et al. (2011), que realiza estimativas em duas etapas para dados correlacionados espacialmente.

O intuito desse trabalho é estudar as hospedagens do *Airbnb* do Rio de Janeiro, em que o preço da diária do aluguel se apresentou assimétrico e correlacionado espacialmente para algumas medidas resumo. Foram aplicados métodos de regressão quantílica apresentados acima a fim de verificar os principais aspectos de uma acomodação e do seu hospedeiro que influenciam no seu preço.

Com o modelo espacial foi possível observar que existem aspectos da acomodação que impactam de forma semelhante entre os bairros, como por exemplo, o número de camas na acomodação. No geral, a região sudeste, além de ser a que possui o maior número de anúncios, é a que apresenta os aluguéis mais caros. Em contrapartida, a região nordeste foi a que se apresentou mais econômica.

Foi realizada uma análise mais detalhada para a mediana do preço do aluguel, e observou-se que duas das variáveis que mais impactam no preço mediano são o número máximo de hóspedes e o número de quartos. Além disso, algumas relações interessantes são a presença de ar condicionado, que está associada a um aumento no valor do aluguel de 13,6% a 61,7% dependendo da região de interesse, e se o anfitrião é um *superhost*, que está associado a uma redução no valor do aluguel em 18,6%, ou um aumento em 17,8%, dependendo da região de interesse.

Palavras-chaves: Regressão quantílica, Regressão quantílica espacial Bayesiana, Preço de aluguel, Airbnb, Rio de Janeiro

Lista de Tabelas

Lista de Figuras

Sumário

1 Introdução

É fundamental quando for viajar para uma outra cidade ou outro país ter um alojamento. Quando se busca alojamentos na internet, é possível encontrar uma grande variedade de preços e tipos de acomodações, e no momento de realizar a escolha, as pessoas em geral consideram o preço, quantas pessoas acomoda, proximidade com pontos turísticos, ambientes de lazer na acomodação, etc. Dado esse grande número de opções disponíveis é interessante escolher sempre as propriedades com o melhor custo-benefício.

Como há diversas propriedades anunciadas na internet, fica fácil ver diversas propriedades com o seu preço e as experiências que têm a oferecer sem mesmo precisar falar previamente com o dono. Como o preço da diária pode acabar sendo muito baixo ou muito alto, é interessante construir um modelo capaz de explicar o que afeta e como afeta os preços, desde os valores mais baixos até os mais altos.

Nesse contexto, um modelo que consegue explorar os principais fatores que diferenciam entre uma propriedade barata e uma cara é o modelo de regressão quantílica. Considerando ainda que a variância e a assimetria da variável preço podem variar dependendo da localidade da propriedade, a regressão quantílica espacial mostra-se mais adequada do que a regressão convencional para a média, a qual supõe normalidade e homoscedasticidade para a variável resposta. Recentemente diversos métodos foram desenvolvidos para modelagem espacial não gaussiana (Griffin e Steel, 2006; Dunson e Park, 2008; Gelfand et al., 2005; Reich et al., 2007) Estes métodos tratam a distribuição condicional da resposta dada a localização espacial e as covariáveis como uma quantidade desconhecida a ser estimada a partir dos dados. Neste trabalho, será feito algo parecido para modelar a densidade condicional do preço de aluguel. Para tanto, será utilizado o modelo de regressão quantílica espacial simplificado proposto por Reich et al. (2011), que será brevemente descrito na seção de metodologia.

Existem na literatura alguns modelos mais flexíveis para ajuste dos dados, por exemplo, modelos de redes neurais (Rather et al., 2015; Ordieres et al., 2005; Odom e Sharda, 1990). No entanto, estas abordagens apresentam desvantagens na parte de interpretação dos parâmetros, não possibilitando mensurar o efeito de cada covariável na variável resposta. Além disso, fica comprometida também a inferência sobre os estimadores. Com a regressão quantílica linear é possível, por exemplo, avaliar se a interação entre o número de camas e o número de quartos causa uma alteração significativa no preço, e em quanto é alterado esse valor se uma dessas covariáveis sofrer mudança, ainda sendo avaliados desde os valores mais baixos até os mais altos, ou seja, em diferentes quantis da variável resposta. Nesse trabalho será utilizado o Modelo de Regressão Espacial Quantílica Bayesiana proposto por Reich et al. (2011), e os resultados serão comparados com os modelos de regressão quantílica e regressão de mínimos quadrados ordinários.

2 Análise espacial

Uma parte da análise exploratória é identificar a estrutura de correlação espacial que melhor descreve os dados (Druck et al., 2004). Indicadores de autocorrelação são estatísticas construídas com o objetivo de caracterizar a dependência espacial dos dados. Esses indicadores podem ser separados em dois tipos: globais e locais. O primeiro descreve a autocorrelação de todo o objeto em estudo e o segundo descreve diferentes regimes de correlação espacial em diferentes sub-regiões.

Neste capítulo será apresentada a matriz de proximidade espacial e os indicadores locais e globais de autocorrelação espacial.

2.1 Matriz de proximidade espacial

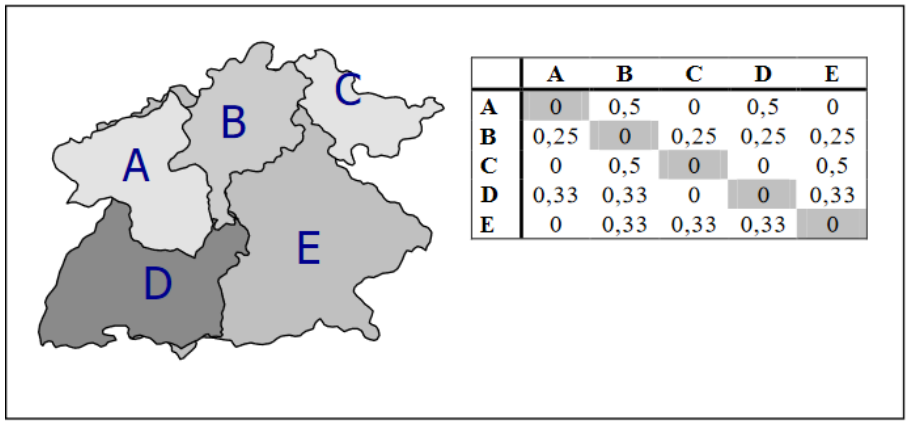
A matriz de proximidade espacial é utilizada para descrever a variabilidade espacial de dados de área. Dado um conjunto de n áreas $\{A_1, \ldots, A_n\}$, a matriz $\boldsymbol{W^{(1)}}$ de dimensão $(n \times n)$ é construída de forma que cada elemento w_{ij} representa uma medida de proximidade entre A_i e A_j. Esta medida de proximidade pode ser calculada a partir de um dos seguintes critérios (Druck et al., 2004):

- $w_{ij} = 1$, se o centróide de A_i está a uma determinada distância de A_j, caso contrário $w_{ij} = 0$;
- $w_{ij} = 1$, se A_i compartilha um lado comum com A_j, caso contrário $w_{ij} = 0$;
- $w_{ij} = l_{ij}/l_i$, onde l_{ij} é o comprimento da fronteria entre A_i e A_j e l_i é o perímetro de A_i;

Apesar desses serem os critérios mais populares, outros também podem ser adotados dependendo do objetivo do estudo.

Geralmente, a matriz $\boldsymbol{W}$ é padronizada com o intuito de simplificar o cálculo de índices de autocorrelação espacial, então cada elemento w_{ij} é dividido pela soma dos pesos da linha de $\boldsymbol{W}$ correspondente. Na Figura 1 é apresentado um exemplo de uma matriz de proximidade espacial binária padronizada pelas linhas.

A matriz de proximidade espacial pode ser generalizada para vizinhos de maior ordem (vizinhos dos vizinhos), em que, de forma análoga aos vizinhos de primeira ordem, pode-se construir as matrizes $\boldsymbol{W^{(2)}}, \ldots, \boldsymbol{W^{(n)}}$.



	A	B	C	D	E
A	0	0,5	0	0,5	0
B	0,25	0	0,25	0,25	0,25
C	0	0,5	0	0	0,5
D	0,33	0,33	0	0	0,33
E	0	0,33	0,33	0,33	0

Figura 1 – Matriz de proximidade espacial de primeira ordem, normalizada pelas linhas FONTE: Druck et al. (2004).

2.2 Índice global de Moran I

Uma parte fundamental na análise exploratória espacial é a caracterização da correlação espacial global. Nesse contexto, o índice mais utilizado é o índice global de Moran I (Moran, 1950), que é expresso por:

$$I = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(z_i - \bar{z})(z_j - \bar{z})}{\sum_{i=1}^{n}(z_i - \bar{z})^2}, \tag{2.1}$$

em que n é o número de áreas, z_i o valor considerado na área i, $\bar{z}$ é o valor médio na região de estudo e w_{ij} são os elementos da matriz de proximidade espacial padronizada. Caso haja interesse na autocorrelação entre vizinhos de ordem superior, basta substituir w_{ij} na Equação (2.1) pelos elementos $w_{ij}^{(k)}$ da matriz de ordem maior $\boldsymbol{W}^{(k)}$.

O índice de Moran I pode ser usado para testar a hipótese nula (H_0) de independência espacial. O índice pertence ao intervalo $[-1, 1]$, tal que 0 significa independência espacial, valores positivos indicam correlação direta e negativos correlação inversa. Sob H_0, a distribuição assintótica do índice de Moran I é Normal. Sua média e variância são (Cliff e Ord, 1981):

$$E(I) = -\frac{1}{n-1}, \tag{2.2}$$

$$Var(I) = \frac{n^2 S_1 - nS_2 + 3S^2}{(n+1)(n-1)S^2} - \left(-\frac{1}{n-1}\right)^2, \tag{2.3}$$

onde:

$$S = \sum_{i=1}^{n} w_{ij},$$

$$S_1 = 0,5\sum_{i=1}^{n}(w_{ij} + w_{ji})^2,$$

$$S_2 = \sum_{i=1}^{n}\left(\sum_{j=1}^{n} w_{ij} + \sum_{j=1}^{n} w_{ji}\right)^2.$$

Portanto, tem-se que:

$$\frac{I - E(I)}{\sqrt{Var(I)}} \sim N(0,1), \quad \text{quando } n \to \infty.$$

Outra possibilidade, sem pressupostos em relação à distribuição e sendo ainda a abordagem mais comum, é um teste de pseudo-significância (Druck et al., 2004), em que, sob H_0, constrói-se a distribuição empírica do estimador gerando-se m permutações aleatórias dos valores dos atributos nas áreas da região de estudo e calcula-se o valor do índice para cada arranjo espacial obtido. Ordenando os índices de forma decrescente, considere que o posto do índice observado na amostra original seja p, então, o p-valor do teste é obtido pela razão $p/(m+1)$ (Hope, 1968).

2.3 Diagrama de espalhamento de Moran

Segundo Druck et al. (2004), o diagrama de espalhamento de Moran é uma maneira de visualizar a dependência espacial. Construído com base nos valores normalizados (centralizados em zero e divididos pelo seu desvio), permite analisar o comportamento da variabilidade espacial. Seu objetivo é comparar os valores normalizados numa área com a média dos seus vizinhos em um gráfico bidimensional, em que z é o valor normalizado e wz é a média dos vizinhos, ilustrado na Figura 2, ele é dividido em quatro quadrantes:

- Q1 (valores positivos, médias positivas) e Q2 (valores negativos, médias negativas): indicam pontos de correlação espacial positiva, em que a localização possui vizinhos com valores semelhantes.
- Q3 (valores positivos, médias negativas) e Q4 (valores negativos, médias positivas): indicam pontos de associação espacial negativa, em que a localização possui vizinhos com valores distintos.

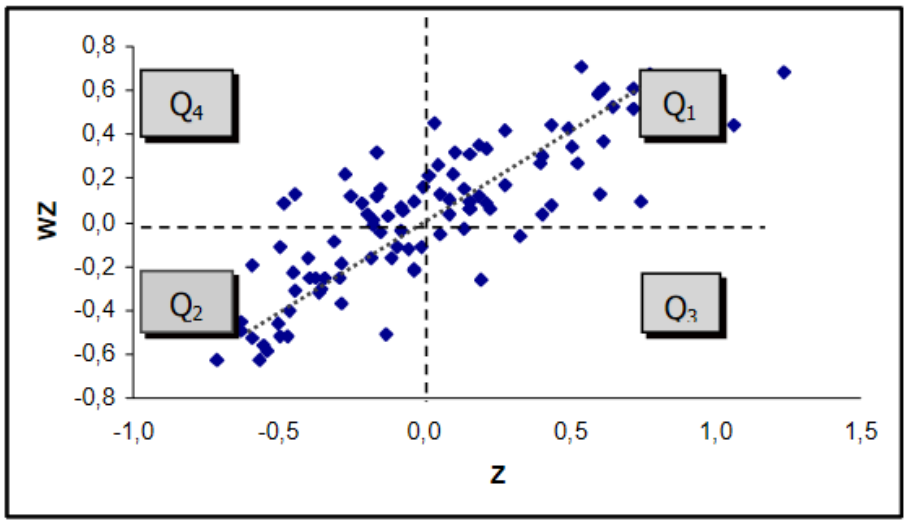


Figura 2 – Diagrama de espalhamento de Moran
FONTE: Druck et al. (2004).

Os quadrantes Q1, Q2, Q3 e Q4 são chamados de alto-alto, baixo-baixo, alto-baixo e baixo-alto, respectivamente. O índice de Moran I é equivalente ao coeficiente de regressão linear

dos pontos do diagrama de espalhamento de Moran, indicando a inclinação da reta de regressão de wz em z.

O mapa de espalhamento de Moran é a visualização georreferenciada do diagrama de dispersão de Moran. As áreas da região são pintadas de quatro cores, representando os quadrantes (Druck et al., 2004).

2.4 Índice local de Moran

Quando há um grande número de áreas, é provável que ocorram diferentes regimes de associação espacial e que apareçam máximos locais de autocorrelação espacial, em que a dependência espacial é mais aparente. Assim, os indicadores locais produzem um valor específico para cada área, permitindo a identificação de agrupamentos. O índice local de Moran pode ser expresso para cada sítio i a partir dos valores normalizados (Druck et al., 2004) como:

$$I_i = \frac{z_i \sum_{j=1}^{n} w_{ij} z_j}{\sum_{j=1}^{n} z_j^2}. \tag{2.4}$$

A significância estatística do índice local de Moran pode ser calculada com testes de pseudo-significância de forma análoga ao caso do índice global (Druck et al., 2004). A presença de áreas com índices locais significativos é um indício de não estacionariedade. Assim, é útil gerar um mapa indicando as regiões que apresentam correlação local significativamente diferente do resto dos dados, denominado mapa de indicadores locais (ou do inglês, *LISA map*).

A combinação do mapa de espalhamento de Moran com os indicadores locais dá origem ao mapa de Moran. Seu intuito é indicar quais classificações do mapa de espalhamento de Moran são significativas de acordo com a significância dos índices locais (Druck et al., 2004).

3 Regressão Quantílica

Neste capítulo será abordado o processo de minimização para encontrar os parâmetros da regressão quantílica, suas propriedades e uma medida de qualidade de ajuste utilizada para regressão sobre quantis.

3.1 Quantil, ordem e otimização

Qualquer variável aleatória X pode ser caracterizada pela sua função de distribuição

$$F(x) = P(X \leq x), \tag{3.1}$$

em que, para qualquer $0 < \tau < 1$,

$$F^{-1}(\tau) = \inf\{x : F(x) \geq \tau\} = Q(\tau) \tag{3.2}$$

é chamado o τ-ésimo quantil de X. A mediana, $F^{-1}(0.5)$ é uma medida de posição da distribuição que divide os dados na metade, ou seja, 50% dos valores assumidos por X ocorrem abaixo da mediana. A vantagem da mediana em relação à média é que ela não é distorcida por valores extremos. Por exemplo, em estudos sobre ativos voláteis, como renda, a média pode ser distorcida por um pequeno número de valores extremamente altos ou baixos, enquanto a mediana não será influenciada por estes pontos discrepantes, desde que mais da metade dos dados não estejam contaminados por valores atípicos.

Pode-se encontrar o τ-ésimo quantil amostral, que normalmente é obtido ordenando as observações, resolvendo um simples problema de otimização (Koenker e Bassett, 1978). Um problema de teoria da decisão é encontrar uma estimativa pontual para uma variável aleatória X com função de distribuição F usando diferentes funções de perda. Considere a função de perda em trechos apresentada na Equação (3.3) e ilustrada na Figura 3

$$\rho_\tau(u) = u\left(\tau - I\left(u < 0\right)\right), \tag{3.3}$$

em que $I(\cdot)$ é a função indicadora e, para qualquer $\tau \in (0, 1)$, deseja-se encontrar $\hat{x}$ que minimiza a perda esperada

$$E\rho_\tau(X - \hat{x}) = (\tau - 1)\int_{-\infty}^{\hat{x}} (x - \hat{x})dF(x) + \tau\int_{\hat{x}}^{\infty} (x - \hat{x})dF(x). \tag{3.4}$$

Derivando a equação (3.4) em relação $\hat{x}$ e igualando a zero, temos

$$0 = (1 - \tau)\int_{-\infty}^{\hat{x}} dF(x) - \tau\int_{\hat{x}}^{\infty} dF(x) = F(\hat{x}) - \tau. \tag{3.5}$$

Como F é uma função monótona, qualquer elemento de $\{x : F(x) = \tau\}$ minimiza a perda esperada. Quando a solução é única, tem-se $\hat{x} = F^{-1}(\tau)$, caso contrário, obtém-se um intervalo

constituído pelo τ-ésimo quantil, e é selecionado como solução o menor elemento para satisfazer a convenção de que o quantil empírico deve ser contínuo pela esquerda (Koenker e Bassett, 1978).

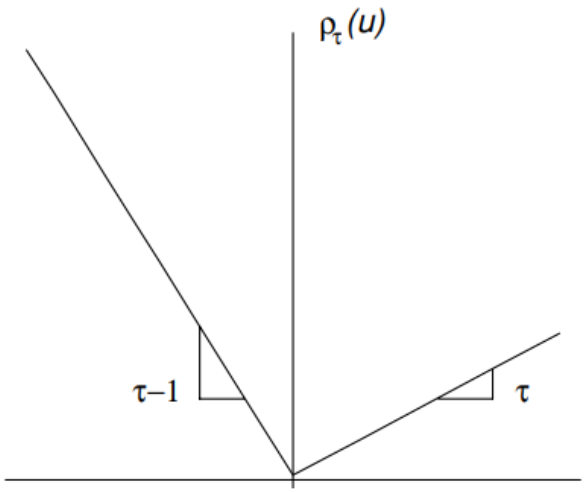


Figura 3 – Função de perda ρ
FONTE: Koenker (2005).

O problema de encontrar o τ-ésimo quantil amostral, que normalmente é visto como um problema de ordenar as observações, virou a solução de um problema de otimização. Assim, pode-se obter o τ-ésimo quantil amostral de Y resolvendo

$$\min_{\xi \in \mathbb{R}} \sum_{i=1}^{n} \rho_\tau(y_i - \xi). \tag{3.6}$$

3.2 Regressão Quantílica

De acordo com Hao (2007), a análise de regressão tem por objetivo expor a relação entre a média da variável resposta condicionada a variáveis explicativas. Modelos criados sobre a média condicional possuem diversas vantagens, sob situações ideais eles são capazes de descrever a relação entre as covariáveis e a variável resposta.

Entretanto, modelos realizados sob a média condicional possuem certas limitações. Primeiramente, eles não são capazes de descrever a variável resposta para valores extremos. Por exemplo, em estudos de renda familiar é de interesse estudar os principais fatores que diferenciam uma família pobre (cauda inferior) e uma rica (cauda superior), para que possam ser estudados e criar políticas que reduzam essa diferença.

Em segundo lugar, diversas suposições muitas vezes não são satisfeitas, principalmente a de homoscedasticidade. Variáveis respostas que possuem uma distribuição de cauda pesada podem acabar gerando resultados inapropriados, pois valores atípicos que se encontram na cauda da distribuição acabam influenciando bastante nos resultados da regressão para a média.

Na metade do século dezoito, foi criada uma alternativa à modelagem da média condicional: a regressão sobre a mediana, que substitui a estimação de mínimos quadrados pela distância absoluta mínima.

A modelagem da mediana condicional, assim como a da média, pode ser utilizada para obter uma estimativa central da variável resposta condicionada às variáveis explicativas, porém ela consegue ser mais consistente com valores discrepantes e ser até mais informativa para a interpretação dos parâmetros em casos que a distribuição da variável resposta é assimétrica.

A mediana, como descrito na seção 3.1, é um quantil específico que divide os dados pela metade. Koenker e Bassett (1978) introduziram a regressão quantílica, que modela o quantil condicional como função de preditores lineares. A regressão quantílica é uma extensão dos modelos de regressão lineares que descreve o quantil condicional.

Como a regressão quantílica se aplica para diversos quantis, consegue-se obter um conhecimento maior da distribuição da variável resposta em relação aos seus preditores, além de ser robusta a valores discrepantes.

O método baseia-se na minimização dos erros absolutos ponderados. Para entendê-lo, partiremos da comparação entre média e quantil, pois a regressão de mínimos quadrados ordinários forma a base para o desenvolvimento da regressão quantílica. Sabendo que a média resolve o problema (Kutner et al., 2005)

$$\min_{\mu\in\mathbb{R}} \sum_{i=1}^{n} (y_i - \mu)^2, \tag{3.7}$$

e expressando a média condicional de y dado $\boldsymbol{x}$ como $\mu(x) = \boldsymbol{x'\beta}$, então β pode ser estimado resolvendo

$$\min_{\beta\in\mathbb{R}^p} \sum_{i=1}^{n} (y_i - \boldsymbol{x_i'\beta})^2. \tag{3.8}$$

O estimador tem forma fechada e é dado por (Kutner et al., 2005)

$$\hat{\boldsymbol{\beta}} = \boldsymbol{(X'X)^{-1}(X'X)Y}. \tag{3.9}$$

De forma análoga, como o τ-ésimo quantil amostral, $q(\tau)$, é solução de

$$\min_{q\in\mathbb{R}} \sum_{i=1}^{n} \rho_\tau(y_i - q), \tag{3.10}$$

onde ρ_τ é a função de perda definida na Equação (3.3), e expressando o τ-ésimo quantil condicional como $Q_y(\tau|\boldsymbol{x}) = \boldsymbol{x'\beta}(\tau)$, as estimativas de $\boldsymbol{\beta}(\tau)$ podem ser obtidas resolvendo (Koenker, 2005)

$$(\hat{\beta}_1(\tau), \ldots, \hat{\beta}_p(\tau))' = \operatorname*{argmin}_{\beta\in\mathbb{R}^p} \sum_{i=1}^{n} \rho_\tau(y_i - \boldsymbol{x_i'\beta}). \tag{3.11}$$

O estimador não tem solução analítica, pois ele não é diferenciável em zero, o que exige a utilização de métodos numéricos para resolver o problema (3.11). Por isso, geralmente a regressão quantílica é reformulada em um problema de programação linear

$$\min_{(\beta,u,\nu)\in\mathbb{R}^p\times\mathbb{R}_+^{2n}} \{\tau 1_n' u + (1-\tau) 1_n' \nu | X\beta + u - \nu = y\}, \tag{3.12}$$

em que X é a matriz de delineamento com n linhas e p colunas (Koenker, 2005).

3.2.1 Propriedade de Equivariância

Quando há o interesse em realizar uma transformação na variável resposta, por exemplo, transformar a temperatura de graus Fahrenheit para Celsius, não é esperado que tais mudanças alterem significativamente as estimativas. Quando os dados são alterados de forma completamente previsível, como uma combinação linear, é esperado que as estimativas sejam alteradas de forma que o resultado da interpretação não seja alterado (Koenker, 2005). Várias dessas propriedades podem ser agrupadas como equivariância e são importantes no auxílio da interpretação dos resultados da regressão. Considere uma regressão no τ-ésimo quantil baseada nos valores observados (y, X), obtendo as estimativas $\hat{\beta}(\tau; y, X)$. Quatro propriedades de equivariância da regressão quantílica para $\hat{\beta}(\tau; y, X)$ são:

Teorema 3.2.1 (Koenker e Bassett (1978)) *Seja A uma matriz qualquer $p \times p$ inversível, $\gamma \in \mathbb{R}^p$, e $a > 0$. Então para qualquer $\tau \in [0, 1]$,*

(i) $\hat{\beta}(\tau; ay, X) = a\hat{\beta}(\tau; y, X)$;

(ii) $\hat{\beta}(\tau; -ay, X) = -a\hat{\beta}(1 - \tau; y, X)$;

(iii) $\hat{\beta}(\tau; y + X\gamma, X) = \hat{\beta}(\tau; y, X) + \gamma$;

(iv) $\hat{\beta}(\tau; y, XA) = A^{-1}\hat{\beta}(\tau; y, X)$;

As propriedades (i) e (ii) implicam em equivariância de escala, a propriedade (iii) é chamada de equivariância de locação e a propriedade (iv) é chamada de equivariância de reparametrização do delineamento.

Considerando que a matriz de delineamento X possua um intercepto, e retornando ao exemplo de transformação da escala de temperatura de Fahrenheit para Centígrados, a mudança na estimativa do intercepto seria simplesmente de $\hat{\beta}(\tau; y, X)$ para $5/9\hat{\beta}(\tau; y, X) - 32$. Estas propriedades de equivariância também estão nas estimativas de mínimos quadrados.

Outra propriedade de equivariância para os quantis, que é ainda mais forte do que as apresentadas no Teorema 3.2.1, é a de equivariância para transformações monótonas. Seja $h(\cdot)$ uma função não decrescente em $\mathbb{R}$. Então, para qualquer variável aleatória Y,

$$Q_{h(Y)}(\tau) = h(Q_Y(\tau)), \tag{3.13}$$

ou seja, os quantis de uma variável aleatória transformada $h(Y)$ são simplesmente os quantis transformados da variável original Y (Koenker, 2005). Porém, a esperança não compartilha essa propriedade para qualquer transformação não decrescente $h(\cdot)$:

$$Eh(Y) \neq h(E(Y)).$$

A propriedade apresentada na Equação (3.13) vem do fato que para qualquer função monótona h,

$$P(Y \leq y) = P(h(Y) \leq h(y)).$$

Esta propriedade possui várias implicações importantes. Diversas vezes, para satisfazer as suposições do modelo de mínimos quadrados, realiza-se uma transformação na variável resposta. Ao fazer a regressão utilizando a variável resposta transformada, fica comprometida a possibilidade de realizar interpretações ou predições para ela no formato original, pois a variável com a transformação inversa tem uma distribuição diferente da variável utilizada na construção do modelo. Agora, no caso da regressão quantílica, essas transformações são mais simples de interpretar por causa da propriedade de equivariância, o que permite utilizar a transformação inversa nas estimativas como se fossem uma estimativa da variável original condicionada nas variáveis explicativas para qualquer quantil.

3.2.2 Distribuição assintótica

De acordo com Koenker (2005, p. 72), a distribuição assintótica do τ-ésimo quantil amostral $\hat{\xi}_\tau$ de uma amostra aleatória $Y_1, \ldots, Y_n$ com função de distribuição F é

$$\sqrt{n}(\hat{\xi}_\tau - \xi_\tau) \sim N(0, \omega^2), \tag{3.14}$$

em que $\omega^2 = \tau(1-\tau)/f^2(\xi_\tau)$ e $f(\cdot)$ é a função de densidade de F.

Koenker (2005, p.72) estende a forma da distribuição conjunta de um único quantil para múltiplos quantis. Tomando $\hat{\zeta}_n = (\hat{\xi}_{\tau_1}, \ldots, \hat{\xi}_{\tau_m})$ e $\zeta_n = (\xi_{\tau_1}, \ldots, \xi_{\tau_m})$ o autor mostra que

$$\sqrt{n}(\hat{\zeta}_n - \zeta) \sim N(0, \Omega), \tag{3.15}$$

onde Ω é uma matriz $m \times m$ composta por

$$(w_{ij}) = \frac{\tau_i \wedge \tau_j - \tau_i\tau_j}{f(F^{-1}(\tau_i))f(F^{-1}(\tau_j))},$$

em que $\tau_i \wedge \tau_j$ é o mínimo entre τ_i e τ_j.

3.2.2.1 Distribuição assintótica das estimativas com erros independentes e identicamente distribuídos (i.i.d.)

Considere um modelo de regressão linear

$$y_i = x_i'\beta + \varepsilon_i,$$

com erros ε_i i.i.d. para $i = 1, \ldots, n$. Suponha que $\{\varepsilon_i\}$ tenham uma função de distribuição comum F e com densidade f, tal que $f(F^{-1}(\tau_i)) > 0$ para $i = 1, \ldots, m$, e $n^{-1}\sum_{i=1}^n x_i x_i' \equiv Q_n$ que converge para uma matriz positiva definida Q_0. Então, tem-se que a distribuição assintótica conjunta das mp estimativas da regressão quantílica $\hat{\zeta}_n = (\hat{\beta}_n(\tau_1)', \ldots, \hat{\beta}_n(\tau_m)')'$ tem a forma (Koenker, 2005)

$$\sqrt{n}(\hat{\zeta}_n - \zeta) \sim N(0, \Omega \otimes Q_0^{-1}), \tag{3.16}$$

em que $\otimes$ é o produto diádico, de tal forma que $\Omega \otimes Q_0^{-1} = \Omega(Q_0^{-1})^T$. Com este resultado é possível realizar inferência sobre as estimativas dos parâmetros utilizando o teste de Wald.

3.2.2.2 Distribuição assintótica das estimativas com erros não i.i.d.

Caso os erros não sejam i.i.d, a distribuição assintótica de $\hat{\beta}(\tau)$ fica mais complicada. A matriz de variância e covariância da estatística $\sqrt{n}(\hat{\beta}(\tau) - \beta(\tau))$ tem a forma (Koenker, 2005)

$$\sqrt{n}(\hat{\beta}(\tau) - \beta(\tau)) \sim N(0, \tau(1-\tau)H_n^{-1}J_nH_n^{-1}),$$

onde

$$J_n(\tau) = n^{-1}\sum_{i=1}^{n} x_i x_i'$$

e

$$H_n(\tau) = \lim_{n\to\infty} n^{-1}\sum_{i=1}^{n} x_i x_i' f_i(\xi_i(\tau)),$$

em que $f_i(\xi_i(\tau))$ é a densidade da variável resposta y_i avaliada no τ-ésimo quantil.

Ainda existe o interesse de se estimar a matriz de variância e covariância para vetores distintos das estimativas da regressão quantílica. Assim a estimativa da matriz de variância e covariância para um vetor de estimativas de regressão quantílica $(\hat{\beta}(\tau_1)', \ldots, \hat{\beta}(\tau_m)')'$ é dada por

$$\begin{aligned} &Cov(\sqrt{n}(\hat{\beta}(\tau_i) - \beta(\tau_i)), \sqrt{n}(\hat{\beta}(\tau_j) - \beta(\tau_j))) \\ &= [\tau_i \wedge \tau_j - \tau_i\tau_j]H_n(\tau_i)^{-1}J_nH_n(\tau_j)^{-1}, \end{aligned} \tag{3.17}$$

em que i e j assumem valores no conjunto $\{1, \ldots, m\}$, J_n e $H_n(\tau)$ foram definidos acima.

3.3 Qualidade de ajuste

Em modelos de mínimos quadrados ordinários, a qualidade de ajuste do modelo é geralmente medida pelo R-quadrado:

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}. \tag{3.18}$$

Essa medida varia de 0 a 1, em que quanto maior o seu valor melhor é o modelo ajustado com base nessa medida.

Uma medida análoga a R^2 para regressão quantílica é a medida $R(\tau)$. Essa medida, é baseada na função de perda (3.3), em que pode haver pesos diferentes caso $y_i > \hat{y}_i$ ou $y_i < \hat{y}_i$. A qualidade de ajuste é medida comparando o modelo de interesse com um modelo de apenas intercepto.

Para definir a medida $R(\tau)$, primeiro é necessário definir a medida $V^1(\tau)$, que é a soma ponderada dos erros do modelo de interesse para o τ-ésimo quantil, e $V^0(\tau)$ a soma ponderada dos erros do modelo que possui apenas o intercepto, ou seja,

$$V^1(\tau) = \sum_{i=1}^{n} d_\tau(y_i, \hat{y}_i) = \sum_{y_i \geq \hat{y}_i} \tau|y_i - \hat{y}_i| + \sum_{y_i < \hat{y}_i} (1-\tau)|y_i - \hat{y}_i| \tag{3.19}$$

e

$$V^0(\tau) = \sum_{i=1}^{n} d_\tau(y_i, \hat{Q}^{(\tau)}) = \sum_{y_i \geq \bar{y}} \tau |y_i - \hat{Q}^{(\tau)}| + \sum_{y_i < \bar{y}} (1-\tau)|y_i - \hat{Q}^{(\tau)}|, \tag{3.20}$$

em que $\hat{Q}^{(\tau)}$ é o valor estimado para o modelo de regressão quantílica que possui apenas o intercepto para o τ-ésimo quantil. A qualidade de ajuste para regressão quantílica é definida como

$$R(\tau) = 1 - \frac{V^1(\tau)}{V^0(\tau)}. \tag{3.21}$$

Como $V^1(\tau)$ e $V^0(\tau)$ são não-negativos, e $V^1(\tau)$ nunca é superior a $V^0(\tau)$, $R(\tau)$, assim como R^2 é no mínimo 0 e no máximo 1, em que quanto maior $R(\tau)$ melhor é o ajuste com base nessa medida.

4 Regressão Quantílica Espacial Bayesiana

Reich et al. (2011) propuseram uma regressão quantílica espacial utilizando a abordagem Bayesiana com o intuito de modelar o nível de ozônio na atmosfera aproveitando o efeito espacial existente nos dados. Este modelo será exposto a seguir (seções 4.1 e 4.2), assim como o modelo simplificado (seção 4.2.1), que também foi proposto no mesmo artigo. Para a modelagem dos preços do aluguel na cidade do RJ será adotado o método simplificado.

É assumido que cada nível do quantil pode ser expresso por uma combinação linear das covariáveis, conforme pode ser visto na Equação (4.1). Além disso, os autores modelam os parâmetros $\beta(\tau, s_i)$ do quantil com um número finito de funções bases com restrições em seus coeficientes para garantir que a função quantílica não se cruze em todos os valores das covariáveis. Mais detalhes serão apresentados na seção 4.1.

Sejam y_i e s_i o valor observado e a localização da i-ésima observação, respectivamente. Deseja-se estimar a densidade condicional de y_i como uma função de s_i e das covariáveis $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ip})'$, onde $X_{i1} = 1$ é o intercepto. Portanto, o modelo linear pode ser escrito como

$$Q_{Y_i}(\tau|\boldsymbol{X_i}, \boldsymbol{s_i}) = \boldsymbol{X_i'}\boldsymbol{\beta}(\tau, \boldsymbol{s_i}), \tag{4.1}$$

em que $\boldsymbol{\beta}(\tau, \boldsymbol{s_i}) = (\beta_1(\tau, \boldsymbol{s_i})), \ldots, \beta_p(\tau, \boldsymbol{s_i})'$ são os coeficientes pelo espaço geográfico para o τ-ésimo quantil.

4.1 Modelo inicial para o processo do quantil

Primeiramente será abordado o modelo que ignora a localização espacial e contém apenas o intercepto, com $\boldsymbol{X}_i = 1$. Neste caso, a função do quantil (4.1) é reduzida para apenas $Q_{Y_i}(\tau) = \beta(\tau)$. O processo $\beta(\tau)$ deve ser construído de forma que $Q_{Y_i}(\tau)$ seja não-decrescente em τ. Seja

$$\beta(\tau) = \sum_{m=1}^{M} B_m(\tau)\alpha_m, \tag{4.2}$$

onde M é o número de funções bases, $B_m(\tau)$ é uma função base conhecida de τ e α_m são coeficientes desconhecidos que determinam a forma da função quantílica. As funções base propostas por Reich et al. (2011) para B_m são os polinômios de *Bernstein*

$$B_m(\tau) = \binom{M}{m}\tau^m(1-\tau)^{M-m}. \tag{4.3}$$

Uma vantagem dessa função basal é que se $\alpha_m \geq \alpha_{m-1}, \forall m > 1$, então $\beta(\tau)$, e consequentemente $Q_{Y_i}(\tau)$, é uma função crescente de τ (Reich et al., 2011). Para garantir que a restrição de monotonicidade seja satisfeita, são geradas diversas sequências de restrições $\delta_m = \alpha_m - \alpha_{m-1} \geq 0$, para $m = 2, \ldots, M$. Essas restrições são suficientes para garantir que Q_Y seja uma função crescente.

Como as restrições de $\alpha = (\alpha_1, \ldots, \alpha_M)$ são expressas em termos da diferença para o termo seguinte, a função será reparametrizada para $\delta_1 = \alpha_1$ e $\delta_m = \alpha_m - \alpha_{m-1}$, para $m = 2, \ldots, M$. Assim os coeficientes originais da função basal podem ser expressos por $\alpha_m = \sum_{l=1}^{m} \delta_l$. Portanto, Reich et al. (2011) garantem a restrição do quantil introduzindo a variável latente irrestrita δ_m^*, tomando $\delta_1 = \delta_1^*$ e fazendo

$$\delta_m = \begin{cases} \delta_m^*, & \delta_m^* \geq 0 \\ 0, & \delta_m^* < 0 \end{cases} \tag{4.4}$$

para $m > 1$.

Segundo Reich et al. (2011), a variável latente δ_m^* possui prioris normais $\delta_m^* \sim N(\overline{\delta}_m(\Theta), \sigma^2)$, com hiperparâmetros Θ desconhecidos. É escolhido $\overline{\delta}_m(\Theta)$ para centralizar o processo do quantil em uma distribuição $f_0(y|\Theta)$, por exemplo, uma variável aleatória $N(\mu_0, \sigma_0^2)$ com $\Theta = (\mu_0, \sigma_0)$. Seja $q_0(\tau|\Theta)$ a função do quantil de $f_0(y|\Theta)$, $\overline{\delta}_m(\Theta)$ é escolhido de forma que

$$q_0(\tau|\Theta) \approx \sum_{m=1}^{M} B_m(\tau)\overline{\alpha}_m(\Theta), \tag{4.5}$$

onde $\overline{\alpha}_m(\Theta) = \sum_{l=1}^{m} \overline{\delta}_l(\Theta)$. As $\overline{\delta}_m(\Theta)$ são escolhidas por meio da regressão *ridge*:

$$(\overline{\delta}_1(\Theta), \ldots, \overline{\delta}_M(\Theta))' = \underset{d}{\operatorname{argmin}} \sum_{k=1}^{K} \left(q_0(\tau_k|\Theta) - \sum_{m=1}^{M} B_m(\tau_k) \left[\sum_{l=1}^{m} d_l \right] \right)^2 + \lambda \sum_{m=1}^{M} d_m^2, \tag{4.6}$$

onde $d_m \geq 0$ para $m > 1$ e $\{\delta_1, \ldots, \delta_K\} \in (0,1)^K$. Tomando a constante λ como zero tem-se o ajuste sem penalização, e tornando-a infinita obtém-se $\overline{\delta} = 0$ para todos os termos. Na prática, é escolhido $\lambda = 1$. Segundo Reich et al. (2011) isso permite que a curva do quantil se ajuste bem e os valores de $\overline{\delta}$ variem suavemente entre si. Quando $\sigma \to 0$ as funções dos quantis diminuem em direção ao quantil paramétrico $q_0(\tau|\Theta)$, e a verossimilhança é parecida com $f_0(y|\Theta)$.

4.2 Modelo espacial para o processo do quantil com covariáveis

Considerando agora as demais covariáveis, a função do quantil condicionado em $\boldsymbol{X_i}$ fica

$$Q_{y_i}(\tau|\boldsymbol{X_i}) = \boldsymbol{X_i'}\boldsymbol{\beta}(\tau) = \sum_{j=1}^{p} X_{ij}\beta_j(\tau). \tag{4.7}$$

Segundo Reich et al. (2011), as curvas dos quantis ainda são modeladas pela função base de polinômios de Bernstein

$$\beta_j(\tau) = \sum_{m=1}^{M} B_m(\tau)\alpha_{jm}, \tag{4.8}$$

em que α_{jm} são coeficientes desconhecidos e B_m foi definido na equação (4.3). O processo $\beta_j(\tau)$ deve ser construído de forma que $Q_{y_i}(\tau|\boldsymbol{X_i})$ seja não decrescente em função de τ para todo $\boldsymbol{X_i}$. O preditor linear com os termos bases é dado por

$$\boldsymbol{X_i'}\boldsymbol{\beta}(\tau) = \sum_{m=1}^{M} B_m(\tau)\theta_m(\boldsymbol{X_i}), \tag{4.9}$$

onde $\theta_m(X_i) = \sum_{j=1}^{p} X_{ij}\alpha_{jm}$. Portanto, se $\theta_m(X_i) \geq \theta_{m-1}(X_i), \forall m > 1$, então $\boldsymbol{X_i'}\boldsymbol{\beta}(\tau)$, e consequentemente $Q_{y_i}(\tau|\boldsymbol{X_i})$, é uma função crescente de τ.

Especificando a priori de α_{jm} para garantir a monotonicidade, e assumindo que $X_{i1} = 1$, que é o intercepto, as demais covariáveis são escalonadas de forma que $X_{ij} \in [0, 1]$ para $j > 1$. De forma similar, as restrições são escritas em termos da diferença com o termo adjacente, tomando $\delta_{j1} = \alpha_{j1}$ e $\delta_{jm} = \alpha_{jm} - \alpha_{jm-1}$ para $m = 2, \ldots, M$. Conforme demonstrado a seguir, a restrição é garantida introduzindo a variável latente irrestrita $\delta^*_{jm} \sim N(\overline{\delta}_{jm}(\Theta), \sigma^2_j)$ e tomando

$$\delta_{jm} = \begin{cases} \delta^*_{jm}, & \delta^*_{1m} + \sum_{j=2}^{p} I(\delta^*_{jm} < 0)\delta^*_{jm} \geq 0 \\ 0, & \text{caso contrário} \end{cases} \tag{4.10}$$

para todo $j = 1, \ldots, p$ e $m = 1, \ldots, M$. Como $X_{i1} = 1$ e $X_{ij} \in [0, 1]$ para $j = 2, \ldots, p$, então $X_{ij}\delta_{jm} \geq X_{ij}I(\delta_{jm} < 0)\delta_{jm} \geq I(\delta_{jm} < 0)\delta_{jm}$ para $j > 1$, e

$$\begin{aligned} \theta_m(\boldsymbol{X_i}) - \theta_{m-1}(\boldsymbol{X_i}) &= \sum_{j=1}^{p} X_{ij}\delta_{jm} \geq \delta_{1m} + \sum_{j=2}^{p} X_{ij}I(\delta_{jm} < 0)\delta_{jm} \\ &\geq \delta_{1m} + \sum_{j=2}^{p} I(\delta_{jm} < 0)\delta_{jm} \geq 0 \end{aligned} \tag{4.11}$$

para todo $\boldsymbol{X_i}$, o que fornece um processo do quantil válido. A curva do intercepto é centralizada em uma função de quantil paramétrica $q_0(\Theta)$, e os demais coeficientes ficam $\overline{\delta}_{jm}(\Theta) = 0$ para $j > 1$.

Para dados espaciais, o processo do quantil pode ser diferente para cada localização geográfica, então Reich et al. (2011) sugerem que

$$\beta_j(\tau, \boldsymbol{s}) = \sum_{m=1}^{M} B_m(\tau)\alpha_{jm}(\boldsymbol{s}), \tag{4.12}$$

onde $\alpha_{jm}(\boldsymbol{s})$ são coeficientes bases a serem estimados que variam no espaço geográfico. A monotonicidade é reforçada para cada localização espacial ao introduzir parâmetros latentes gaussianos $\delta^*_{jm}(\boldsymbol{s})$. Os parâmetros latentes se relacionam com os coeficientes da função base de forma que $\alpha_{jm} = \sum_{l=1}^{m} \delta_{jl}(\boldsymbol{s})$ e

$$\delta_{jm}(\boldsymbol{s}) = \begin{cases} \delta^*_{jm}(\boldsymbol{s}), & \delta^*_{1m}(\boldsymbol{s}) + \sum_{j=2}^{p} I(\delta^*_{jm}(\boldsymbol{s}) < 0)\delta^*_{jm}(\boldsymbol{s}) \geq 0 \\ 0, & \text{caso contrário} \end{cases} \tag{4.13}$$

para todo $j = 1, \ldots, p$ e $m = 1, \ldots, M$.

Para que a densidade condicional do quantil varie de forma suave no espaço geográfico, $\delta^*_{jm}(\boldsymbol{s})$ é modelado como processos espaciais. Os $\delta^*_{jm}(\boldsymbol{s})$ são processos espaciais Gaussianos

independentes entre j e m, com média $E(\delta^*_{jm}(\boldsymbol{s})) = \overline{\delta}_{jm}(\Theta)$ e matriz de variância e covariância exponencial $Cov(\delta^*_{jm}(\boldsymbol{s}), \delta^*_{jm}(\boldsymbol{s'})) = \sigma_j^2 \exp(-||s - s'||/\rho_j)$, em que σ_j^2 é a variância de $\delta^*_{jm}(\boldsymbol{s})$, ρ_j é o raio da correlação espacial e $||\cdot||$ é a função norma. Maiores detalhes para a estimação desse modelo podem ser encontrados em Reich et al. (2011).

4.2.1 Método aproximado

Os métodos descritos nas seções 4.1 e 4.2 se tornam inviáveis para banco de dados grandes, e possuem uma desvantagem que restringem a utilização de covariáveis com interação e termos quadráticos, o que levou os autores a considerarem um método aproximado que viabiliza a construção de um modelo ainda elaborado e a utilização de um banco de dados maior.

O método simplificado é composto por dois estágios. Primeiramente, é realizada a regressão quantílica para cada região do espaço geográfico para obter estimativas preliminares do processo do quantil e sua covariância assintótica. No segundo estágio é realizada uma suavização nessas estimativas utilizando o modelo Bayesiano espacial. Mais detalhes estão apresentados a seguir.

No primeiro estágio, as estimativas da regressão quantílica na região geográfica $\boldsymbol{s}$ são obtidas utilizando a equação (3.11), ou seja,

$$\begin{aligned}(\hat{\beta}_1(\tau_k, \boldsymbol{s}), \ldots, \hat{\beta}_p(\tau_k, \boldsymbol{s}))' \\ &= \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \sum_{s_i = s, y_i > \boldsymbol{X_i'}\beta} \tau_k |y_i - \boldsymbol{X_i'}\boldsymbol{\beta}| \\ &+ \sum_{s_i = s, y_i < \boldsymbol{X_i'}\beta} (1 - \tau_k)|y_i - \boldsymbol{X_i'}\boldsymbol{\beta}|\end{aligned} \tag{4.14}$$

A covariância assintótica é na forma

$$\begin{aligned}&Cov[\sqrt{n_s}(\hat{\beta}_1(\tau_k, \boldsymbol{s}), \ldots, \hat{\beta}_p(\tau_k, \boldsymbol{s})), \\ &\sqrt{n_s}(\hat{\beta}_1(\tau_l, \boldsymbol{s}), \ldots, \hat{\beta}_p(\tau_l, \boldsymbol{s}))] \\ &= H(\tau_k)^{-1} J(\tau_k, \tau_l) H(\tau_l)^{-1},\end{aligned} \tag{4.15}$$

em que n_s é o número de observações na região s, $H(\tau) = \lim_{n_s \to \infty} n_s^{-1} \sum_{i=1}^{n_s} \boldsymbol{X_i} \boldsymbol{X_i'} f_i(\boldsymbol{X_i'}\hat{\boldsymbol{\beta}}(\tau))$, $f_i(\boldsymbol{X_i'}\hat{\boldsymbol{\beta}}(\tau))$ é a densidade condicional de y_i avaliada em $\boldsymbol{X_i'}\hat{\boldsymbol{\beta}}(\tau)$ e $J(\tau_k, \tau_l) = [\tau_k \wedge \tau_l - \tau_k \tau_l] n_s^{-1} \sum_{i=1}^{n_s} \boldsymbol{X_i} \boldsymbol{X_i'}$, onde $\tau_k \wedge \tau_l$ é o mínimo entre τ_k e τ_l.

Essas estimativas não são suaves nem no espaço geográfico, nem no nível do quantil, e não garantem que as estimativas do quantil não se cruzem para todo $\boldsymbol{X}$. As estimativas obtidas na Equação (4.14) serão suavizadas em um segundo estágio utilizando o processo espacial descrito na Seção 4.2. Tomando $\hat{\boldsymbol{\beta}}(\boldsymbol{s_i}) = (\hat{\beta}_1(\tau_1, s_i), \ldots, \hat{\beta}_1(\tau_k, s_i), \hat{\beta}_2(\tau_1, s_i), \ldots, \hat{\beta}_p(\tau_k, s_i))'$ e $Cov(\hat{\boldsymbol{\beta}}(\boldsymbol{s_i})) = \boldsymbol{\Sigma_i}$, definida na equação 4.15, Reich et al. (2011) propõem no segundo estágio o ajuste do seguinte modelo

$$\hat{\boldsymbol{\beta}}(\boldsymbol{s_i}) \sim N(\boldsymbol{\beta}(\boldsymbol{s_i}), \boldsymbol{\Sigma_i}), \tag{4.16}$$

onde $\boldsymbol{\beta}(\boldsymbol{s_i}) = (\beta_1(\tau_1, s_i), \ldots, \beta_1(\tau_k, s_i), \beta_2(\tau_1, s_i), \ldots, \beta_p(\tau_k, s_i))'$ são funções dos polinômios de Bernstein definidos na equação (4.12).

Esta aproximação fornece uma redução do tempo computacional, pois a dimensão da variável resposta é reduzida do número de observações na região geográfica para o número de quantis na aproximação, e as posterioris que definem as estimativas β são conjugadas, o que permite que o algoritmo de Gibbs tenha uma convergência mais rápida. Para maiores detalhes vide Reich et al. (2011).

5 Materiais e métodos

5.1 Materiais

O banco de dados utilizado é sobre preços de aluguel do site Airbnb para a cidade do Rio de Janeiro. Os dados foram obtidos do site *insideairbnb*, que oferece informações de aluguel do *Airbnb* para diversas cidades e em diversos períodos, e a data de atualização do banco utilizado é 18 de Março de 2020.

Neste trabalho será utilizada a variável do banco chamada *listing*, que contém o preço base do aluguel. O preço base é aquele que não se refere a uma data específica, ou seja, é o preço que o usuário visualiza no site antes de selecionar uma data específica. Portanto, o interesse aqui é estudar o comportamento do preço das acomodações do Rio de Janeiro fora de datas especiais, como festividades e feriados. Algumas acomodações possuíam um preço base da diária acima de R$10.000,00. Para estas acomodações foi utilizado o preço do calendário, que é a informação do preço das diárias desde 18 de março de 2020 até 18 de março de 2021, considerando a mediana dos preços se esta for inferior ao corte de R$10.000,00. Caso contrário utilizou-se o menor preço disponível no calendário para a acomodação. Houve a alteração no preço de 101 acomodações com base nesses critérios.

O banco original possuía 104 covariáveis que forneciam diversas informações, por exemplo, o link do anúncio, a descrição da acomodação pelo hospedeiro, entre outros. Primeiramente, foram removidos os links, pois eles seriam irrelevantes no modelo, em seguida, foram removidas também as informações textuais fornecidas pelo hospedeiro, por exemplo, a descrição sobre o espaço, trânsito, entre outros, visto que era muito texto sendo fornecido, e a análise sobre essas variáveis não seria objetiva. Também havia informações que não estavam preenchidas na maior parte dos anúncios. Por exemplo, o tamanho da propriedade, apesar de ser extremamente relevante para explicar o preço do aluguel, foi fornecido para apenas 1,3% das acomodações. Logo, as variáveis ausentes em mais de 30% das observações foram removidas da análise. Em seguida, foram descartadas as variáveis dicotômicas que eram constantes dentro de algum bairro. Por exemplo, uma variável indicando se o lugar é compartilhado, que para alguns bairros era sempre falsa, o que iria gerar problemas para a regressão espacial. As variáveis categóricas com mais de duas categorias e com categorias que não ocorriam para algum bairro foram agrupadas, como é o caso do tipo de propriedade, que de 37 categorias foram reduzidas para três (apartamento, casa e outros). Por fim, foi realizada a regressão do primeiro estágio para a mediana para cada grupo de bairros separadamente (vide Capítulo 4), e foram removidas as covariáveis não significativas em pelo menos três sítios. Após essa limpeza e seleção o banco ficou com 89 covariáveis para explicar o preço, sendo que algumas delas serão agrupadas conforme detalhado na Tabela 1.

Algumas acomodações que ainda se referiam a eventos não cotidianos no Rio de Janeiro, como Olimpíadas, Copa do Mundo ou *Rock in Rio*, foram removidas do banco. Isso foi feito com base no título do anúncio do imóvel utilizando expressões regulares. Além disso, foram retiradas acomodações com o preço acima de R$100.000,00, pois as propriedades disponíveis a este preço não possuíam nenhuma característica que justificasse este preço de diária. Também foram removidas as propriedades que não estavam disponíveis para alugar durante o ano todo, imóveis cuja página não era atualizada há mais de três anos e propriedades em que o hospedeiro claramente informou incorretamente alguma informação, por exemplo uma acomodação com centenas de banheiros. O banco final ficou com 29.168 acomodações e com 24 colunas, entre elas, algumas foram criadas a partir de outros dados disponíveis no banco original, conforme descrito na Tabela 1.

Tabela 1 – Descrição das variáveis no banco.

Variável	Descrição
preco	Preço da diária da acomodação.
latitude	Latitude da acomodação.
longitude	Longitude da acomodação.
hosp_tempo	Tempo, em anos, que o hospedeiro está cadastrado no *Airbnb*.
hosp_rio	Variável dicotômica identificando se o hospedeiro está na cidade do Rio de Janeiro.
hosp_superhost	Variável dicotômica identificando se o hospedeiro é considerado *superhost* pelo *Airbnb*.
tipo_propriedade	Variável politômica identificando o tipo da propriedade, sendo elas Casa, Apartamento e Outros (Hiate, Tenda, Cabana, etc.).
max_hospedes	Limite de pessoas que a acomodação suporta, foi truncado em 25 pessoas para tratamento de *outlier*.
banheiros	Número de banheiros disponíveis na acomodação.
quartos	Número de quartos disponíveis na acomodação.
camas	Número de camas que existe na acomodação.
piscina	Variável dicotômica identificando se tem piscina na acomodação.
ar_condicionado	Variável dicotômica identificando se tem ar condicionado na acomodação.
dep_seg	Depósito exigido pelo hospedeiro no momento do aluguel, normalmente o valor é restituído após o término da estadia.
minimo_noites	Número mínimo de dias exigidos para aluguel.
meses_sem_att	Número de meses em que o hospedeiro não atualiza a página.
num_aval_ua	Número de comentários de hóspedes no último ano.
res_inst	Disponível para alugar diretamente pelo site.
canc_restrito	Variável dicotômica identificando se a política de cancelamento do hospedeiro é restrita.
hosp_nanun	Número de anúncios do hospedeiro no *Airbnb*.

Tabela 1 – Descrição das variáveis no banco (Continuação).

Variável	Descrição
disp_anual	Proporção de dias disponíveis para aluguel da acomodação no período de um ano.
distancia_costa	Distância da acomodação para a costa em quilômetros. Variável criada utilizando a latitude e longitude e um polígono de terras.
comod_TRI	Variável criada a partir de um modelo de Teoria de Resposta ao Item contendo 68 comodidades que não ficaram explícitas no modelo.
bairro	Variável identificando o bairro da acomodação. Alguns bairros foram agrupados por não atenderem a um mínimo de 100 acomodações por bairro.

5.1.1 Composição da variável bairro

Conforme indicado na Tabela 1, a variável bairro identifica o bairro da acomodação, mas alguns bairros foram agrupados por terem observações insuficientes. Primeiramente, os bairros com menos de 100 acomodações foram identificados por uma variável b_n ordenada de forma decrescente com base no número de acomodações do bairro. Em seguida, o algoritmo a seguir foi aplicado.

1. escolhe-se o primeiro bairro de b_n para agrupar;
2. calcula-se a distância do bairro selecionado para os demais bairros do banco;
3. o bairro selecionado é agrupado com o mais próximo, desde que a soma de acomodações seja pelo menos 100;
4. seleciona o próximo bairro de b_n;
5. repete os passos 2 - 4 até percorrer todos os bairros de b_n;

Com esse agrupamento proposto consegue-se garantir vários grupos para a regressão espacial, que não seria possível se fossem utilizadas as zonas ou regiões no agrupamento.

Os bairros agrupados estão ilustrados na Figura 4, sendo que em branco são os bairros que não fazem parte do banco. Além disso, todos os grupos estão detalhados na Tabela 9, no Apêndice B, com os seus respectivos números de acomodações.

Conforme ilustrado na Figura 4, 15 grupos foram formados após a aplicação do algoritmo proposto.

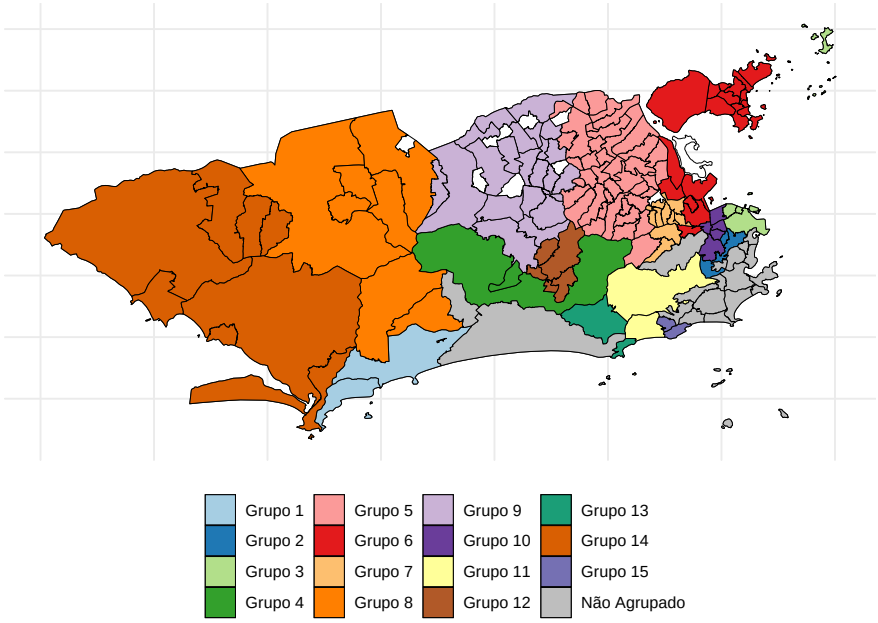


Figura 4 – Resultado do agrupamento de bairros.

5.1.2 Composição da variável comodidades

Modelos de Teoria de Resposta ao Item (TRI) são utilizados principalmente na área de avaliação educacional, na formulação de modelos para traços latentes dos respondentes de uma prova.

A TRI foi desenvolvida, principalmente, para suprir limitações que a teoria clássica do teste apresenta, como o fato do instrumento de medida ser dependente das características dos respondentes que se submetem ao questionário ou teste. A TRI possibilita modelar a probabilidade de um indivíduo dar uma certa resposta ao item como função dos parâmetros do item e da habilidade do respondente.

Essa teoria também pode ser aplicada para reduzir a dimensionalidade da matriz de delineamento em um problema de regressão. Na aplicação dessa monografia, o objetivo do modelo de TRI é indicar o grau de conforto que a acomodação possui (variável latente) tendo em vista as comodidades presentes na acomodação (covariáveis dicotômicas indicando a presença ou ausência de internet, Televisão, elevador, etc.). Neste caso, as covariáveis dicotômicas são equivalentes aos itens de uma prova, e a variável latente explicada pelo modelo de TRI é o grau de conforto da acomodação.

A variável comod_TRI, como descrito na Tabela 1, é o resultado de um modelo de Teoria de Resposta ao Item. O modelo engloba ao todo 68 itens, e na Tabela 8, no Apêndice A, são apresentadas as comodidades que estão implícitas em comod_TRI, com o quantitativo presente no banco de dados, o seu respectivo percentual e os parâmetros de TRI para as comodidades observadas.

5.1.2.1 Modelo de TRI para itens dicotômicos

De acordo com Andrade et al. (2000), o modelo logístico unidimensional de três parâmetros é o mais utilizado atualmente, sendo dado por:

$$P(U_{ij} = 1|\theta_j) = c_i + (1 - c_i)\frac{1}{1 + e^{-Da_i(\theta_j - b_i)}}, \tag{5.1}$$

com $i = 1, 2, \ldots, I$ e $j = 1, 2, \ldots, n$,

U_{ij}	é uma variável indicadora indicando o acerto do indivíduo j no item i.
θ_j	é o traço latente do j-ésimo indivíduo.
$P(U_{ij} = 1\|\theta_j)$	é a probabilidade de um indivíduo j com habilidade θ_j responder corretamente o item i.
a_i	é o parâmetro de discriminação do item i.
b_i	é o parâmetro de dificuldade do item i.
c_i	é o parâmetro de acerto ao acaso do item i.
D	é um fator de escala, normalmente igual a 1.

O item é considerado mais difícil quanto maior for o traço latente para que um sujeito consiga responder de forma positiva o item.

O parâmetro de discriminação representa o quanto ele consegue diferenciar indivíduos de diferentes traços latentes, itens com discriminação baixa faz com que indivíduos de baixa e alta proficiência tenham uma probabilidade próxima de responder positivamente o item.

Para o parâmetro de acerto ao acaso, ele representa a probabilidade de um respondente de baixa proficiência responder corretamente o item.

É possível obter o modelo logístico de dois parâmetros fixando em 0 o parâmetro c_i na Equação (5.1), e para se obter o modelo logístico de um parâmetro fixa-se em 1 o parâmetro a_i e em 0 o parâmetro c_i. Neste trabalho, o modelo utilizado foi o modelo logístico de dois parâmetros.

5.1.2.2 Métodos de Estimação em TRI

Os parâmetros de TRI podem ser estimados de duas formas, sendo elas, estimação conjunta e marginal. Neste trabalho será considerado o Método de Máxima Verossimilhança Marginal, utilizando o algoritmo EM, que propõe fazer a estimação em duas etapas: primeiro os parâmetros dos itens, e posteriormente, as proficiências.

O algoritmo EM é um processo iterativo, que cada iteração é composta por dois passos, primeiro calcula-se a Esperança (E), e em seguida Maximiza (M). No caso da TRI o objetivo é obter estimativas dos parâmetros dos itens $\zeta_i = (a_i, b_i \text{ e } c_i)$ na presença de variáveis não

observadas, sendo elas as proficiências dos respondentes $\boldsymbol{\theta}$. Tomando $f(\boldsymbol{u..}, \boldsymbol{\theta}|\boldsymbol{\zeta})$ a densidade conjunta dos dados completos, ou seja,

$$f(\boldsymbol{u..}, \boldsymbol{\theta}|\boldsymbol{\zeta}) = \prod_{j=1}^{n}\prod_{i=1}^{I} P_{ji}^{u_{ji}}(1-P_{ji})^{1-u_{ji}}, \tag{5.2}$$

em que P_{ji} é a probabilidade do j-ésimo indivíduo acertar o i-ésimo item (Equação (5.1)) e u_{ji} é a variável indicadora se o indivíduo j acertou de fato o item i. E considerando que $\hat{\boldsymbol{\zeta}}^{(k)}$ é a estimativa de $\boldsymbol{\zeta}$ na iteração k, então para se obter a estimativa de $\hat{\boldsymbol{\zeta}}^{(k+1)}$ pelo algoritmo EM, tem-se que:

- **Passo E:** Calcular $E[\log f(\boldsymbol{u..}, \boldsymbol{\theta}|\boldsymbol{\zeta})|\boldsymbol{u..}, \hat{\boldsymbol{\zeta}}^{(k)}]$
- **Passo M:** Obter $\hat{\boldsymbol{\zeta}}^{(k+1)}$ que maximiza a função do Passo E.

Para mais detalhes veja Andrade et al. (2000).

Com relação ao contexto dessa monografia, a variável *comod_TRI* descrita na seção 5.1 será a proficiência θ estimada da acomodação considerando as comodidades do imóvel.

5.2 Métodos

Para a realização deste trabalho, todos gráficos, tabelas, estimações dos parâmetros serão feitos através do *software* estatístico R (R Core Team, 2020). Destacando-se neste trabalho o pacote quantreg (Koenker, 2020) para as estimativas de regressão quantílica do primeiro estágio, e a função disponibilizada por Reich et al. (2011) para o processo de estimação e suavização espacial.

Para um estudo inicial dos dados será feita uma análise exploratória dos dados a partir de gráficos de dispersão e indicadores espaciais.

A fim de reduzir a dimensionalidade da matriz de delineamento, algumas variáveis dicotômicas serão agrupadas em um modelo de TRI. Algumas variáveis categóricas foram mantidas explícitas no modelo para fins de interpretação devido a sua maior relevância, são elas: se o hospedeiro mora no Rio de Janeiro (hosp_rio), se o hospedeiro é *superhost* (hosp_superhost), se existe piscina na acomodação (piscina), se a acomodação possui ar condicionado (ar_condicionado), se é possível alugar diretamente pelo site (res_inst) e se a política de cancelamento do hospedeiro é restrita (canc_restrito).

Todos os mapas foram feitos utilizando o arquivo de polígonos espaciais disponibilizados no pacote geobr (Pereira e Goncalves, 2021).

Nesta monografia, será utilizado o modelo simplificado da Regressão Quantílica espacial Bayesiana de Reich et al. (2011) para ajustar o preço de aluguel dos imóveis no Rio de Janeiro em função das covariáveis disponíveis no banco de dados.

Os modelos de regressão quantílica tradicional e de regressão por mínimos quadrados ordinários também serão utilizados para comparação do ajuste. Essa comparação será feita utilizando o erro quadrático médio e o erro médio absoluto, considerando apenas os valores preditos e o verdadeiro.

6 Resultados

6.1 Descrição dos dados

Para compreender melhor a natureza dos dados, de forma que a modelagem proposta seja viável, foi feita uma análise exploratória. Os resultados estão apresentados abaixo.

O histograma da variável resposta preço está apresentado na Figura 5, em que foram consideradas as acomodações com o preço até o quantil 99% (R$ 5.063,62), visto que o restante das acomodações tem o seu preço até dez vezes maior do que o preço do quantil de corte. Além disso, foi construído outro histograma com todas observações na Figura 6 (vide canto inferior direito).

A variável preço utilizada é o preço de listagem (preço base de aluguel) dos imóveis cadastrados no site *Airbnb* em março de 2020.

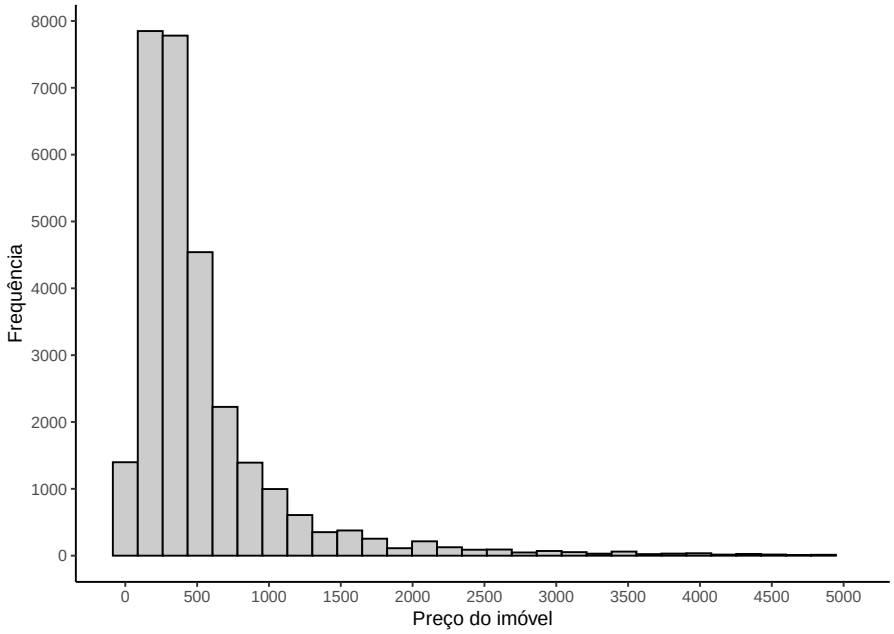


Figura 5 – Histograma do preço da acomodação até o quantil 99%.

Nota-se que os preços de diária mais comum estão entre R$ 200,00 e R$ 500,00. Além disso, existe uma assimetria com concentração à esquerda, que mostra que valores de aluguel elevados estão presentes. Ainda existe uma frequência alta de acomodações com o preço da diária acima de R$ 1.000,00, chegando até o valor de R$50.734,00, que constitui aproximadamente 12,42% das acomodações do Rio de Janeiro pelo *Airbnb*.

Nas Figuras de 6 a 9 são apresentadas matrizes de dispersão dos dados, sendo que na parte inferior esquerda da matriz está o gráfico de dispersão das covariáveis analisadas e na diagonal principal está o nome da covariável com o seu histograma. Por fim, a parte superior direita da matriz apresenta o coeficiente de correlação de Pearson das covariáveis em estudo. As

variáveis apresentadas nas Figuras 6 e 7 são: há quanto tempo o hospedeiro está cadastrado no *Airbnb* (hosp_tempo), o número máximo de hóspedes (max_hospedes), o número de banheiros (banheiros), o número de quartos (quartos), o número de camas (camas), o valor do depósito de segurança (dep_seg). Agora as covariáveis apresentadas nas Figuras 8 e 9 são: o número mínimo de noites de aluguel (minimo_noites), o tempo em que o hospedeiro não atualiza o anúncio (meses_sem_att), o número de avaliações no último ano (num_aval_ua), o número de anúncios do hospedeiro no *Airbnb*, o percentual de dias disponíveis para aluguel no ano (disp_anual), a distância até a costa (distancia_costa), o valor do modelo de teoria de resposta ao item (comod_TRI). Além disso, para as Figuras 6 e 8 é utilizada a variável preço do imóvel, enquanto que nas Figuras 7 e 9 é utilizado o logaritmo do preço.

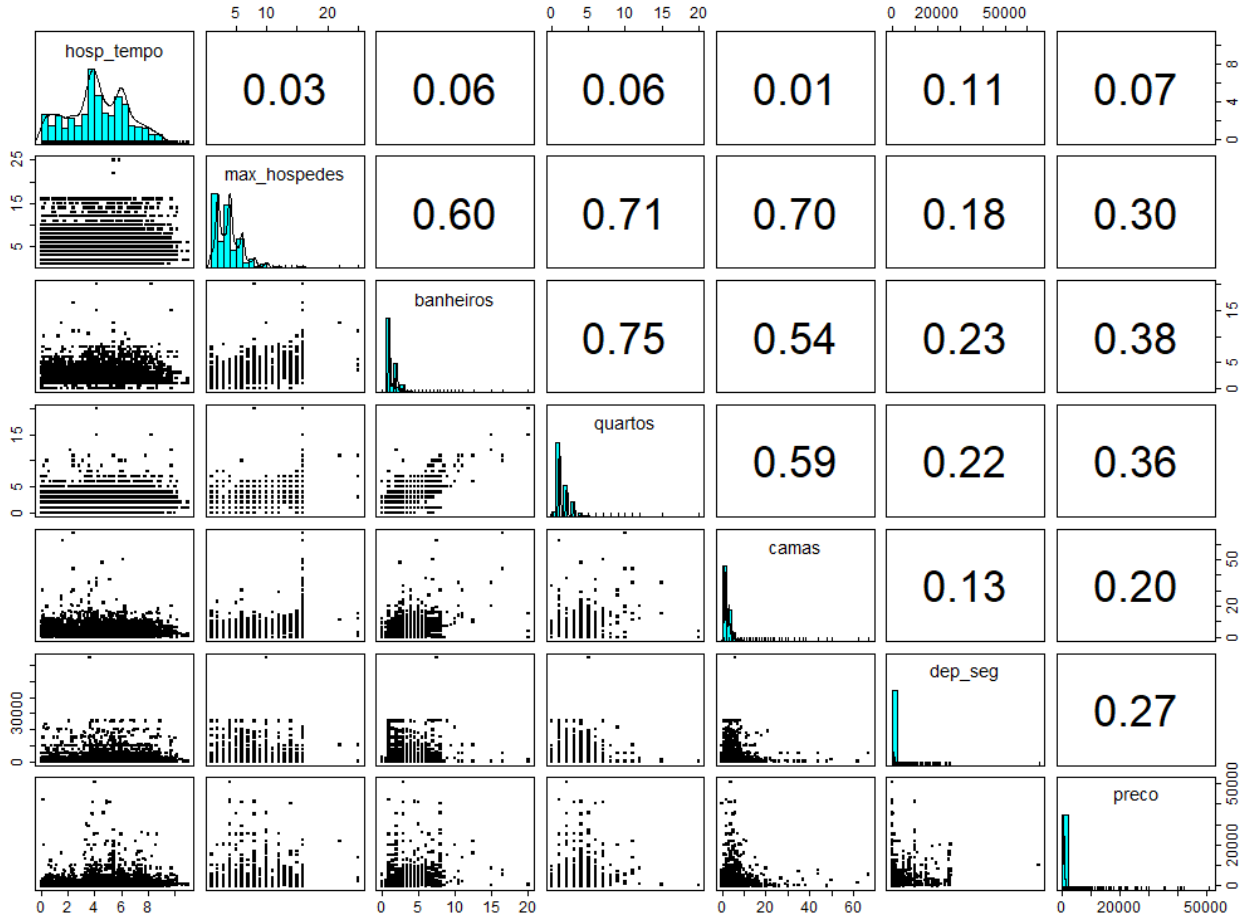


Figura 6 – Matriz de dispersão para os dados do *Airbnb* do Rio de Janeiro com a variável preço do aluguel.

Pelas Figuras 6 e 7 pode-se ver que a transformação logarítmica para o preço gerou uma maior correlação de Pearson para os dados, além disso todos os gráficos apresentaram uma correlação positiva. Pode-se ver que existe uma correlação elevada entre o número de banheiros, quartos, camas e o número máximo de hóspedes, que se relacionam diretamente com o tamanho da acomodação alugada. Considerou-se utilizar análise de componente principal para agrupar essas covariáveis, porém para ainda ter uma boa explicação da variabilidade, o método iria reduzir apenas uma dimensão e ainda dificultaria a interpretação.

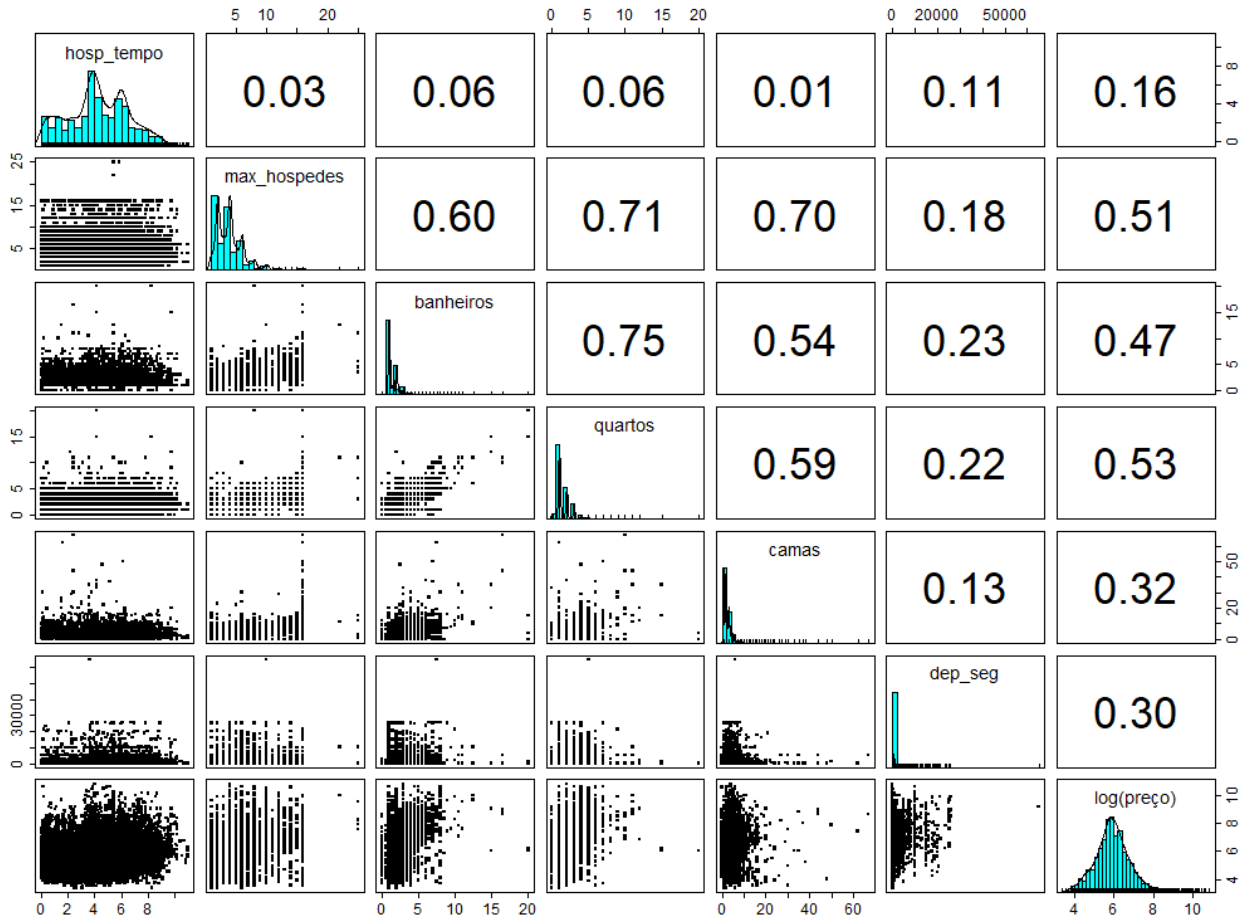


Figura 7 – Matriz de dispersão para os dados do *Airbnb* do Rio de Janeiro com a variável transformada logaritmo do preço do aluguel.

A maior parte das acomodações (67%) não exigia depósito de segurança dos hóspedes, o que justifica os histogramas correspondentes apresentados nas Figuras 6 e 7 serem concentrados em zero. Agora, das acomodações que exigiam o depósito de segurança, 82% pediam que o seu valor fosse maior do que a diária.

Nota-se pelas Figuras 8 e 9 que as comodidades (comod_TRI) da acomodação tiveram uma correlação positiva moderada com o número de avaliações no último ano, o que indica que um maior número de benfeitorias para melhorar a estadia está associado com anúncios com mais avaliações. Além disso, a variável de comodidades apresentou uma correlação negativa com a variável número de meses sem atualizar (meses_sem_att), o que indica que pessoas que passam um maior tempo sem atualizar a página do anúncio tendem a disponibilizar menos itens em comodidades na acomodação, ou os estão omitindo no anúncio.

Pelas Figuras 8 e 9 vemos que a transformação logarítmica para o preço foi capaz de aumentar a correlação para a distância para a costa, enquanto que para as demais covariáveis não houve uma grande diferença no valor da correlação.

Observando as figuras que utilizam o logaritmo do preço (Figuras 7 e 9) não é possível ver uma relação funcional clara para ser utilizada entre o logaritmo do preço e as demais covariáveis

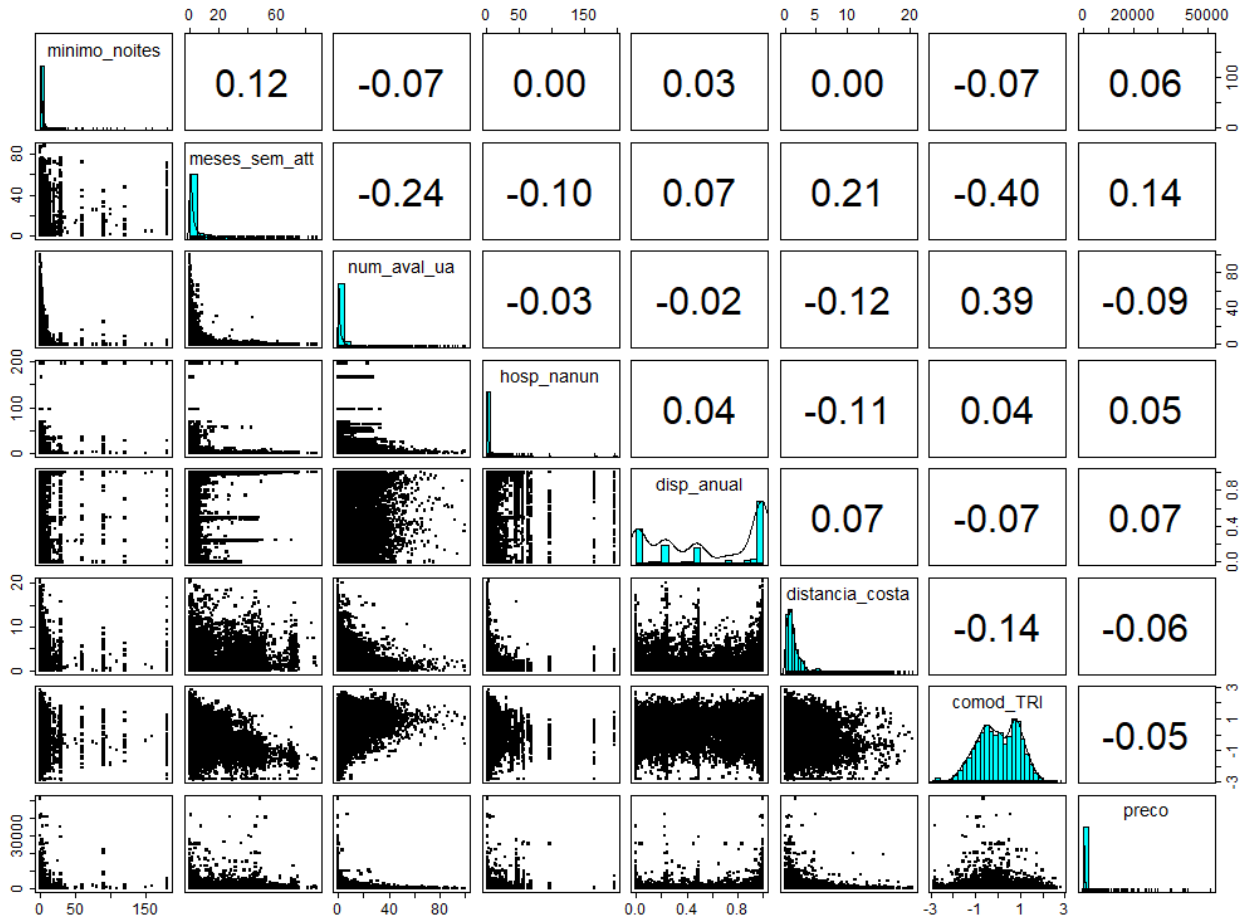


Figura 8 – Matriz de dispersão para os dados do *Airbnb* do Rio de Janeiro com a variável preço do aluguel.

contínuas no estudo.

Em geral, nota-se que as principais variáveis contínuas que influenciam o preço de aluguel da acomodação são as que se relacionam com o seu tamanho.

A Tabela 2 apresenta uma breve descrição da relação entre o preço do aluguel e as variáveis categóricas, incluindo o máximo, a média e os quantis 10%, 25%, 50%, 75%, 90% do preço do imóvel para cada categoria. Optou-se por não incluir essas covariáveis no modelo de TRI a fim de poder avaliar o impacto direto desses itens no preço do aluguel.

Na Tabela 2 podemos observar que a diferença entre as categorias é pequena para os menores preços, e essas diferenças aumentam à medida que o nível do quantil cresce. As propriedades do tipo Apartamento tiveram o valor mais alto para o seu primeiro e segundo quartil em relação as outras categorias, mas essa relação se inverte para os demais quantis. Percebe-se também uma grande diferença entre a média e a mediana do preço para as categorias, sendo que os valores médios estão, no geral, mais próximos do terceiro quartil, o que destaca como esses valores são assimétricos e os preços elevados influenciam a média.

Para as comodidades que ficaram explícitas (que não foram inclusas no modelo de TRI em *comod_TRI*), temos que acomodações com piscina tem o valor de aluguel mais elevado para

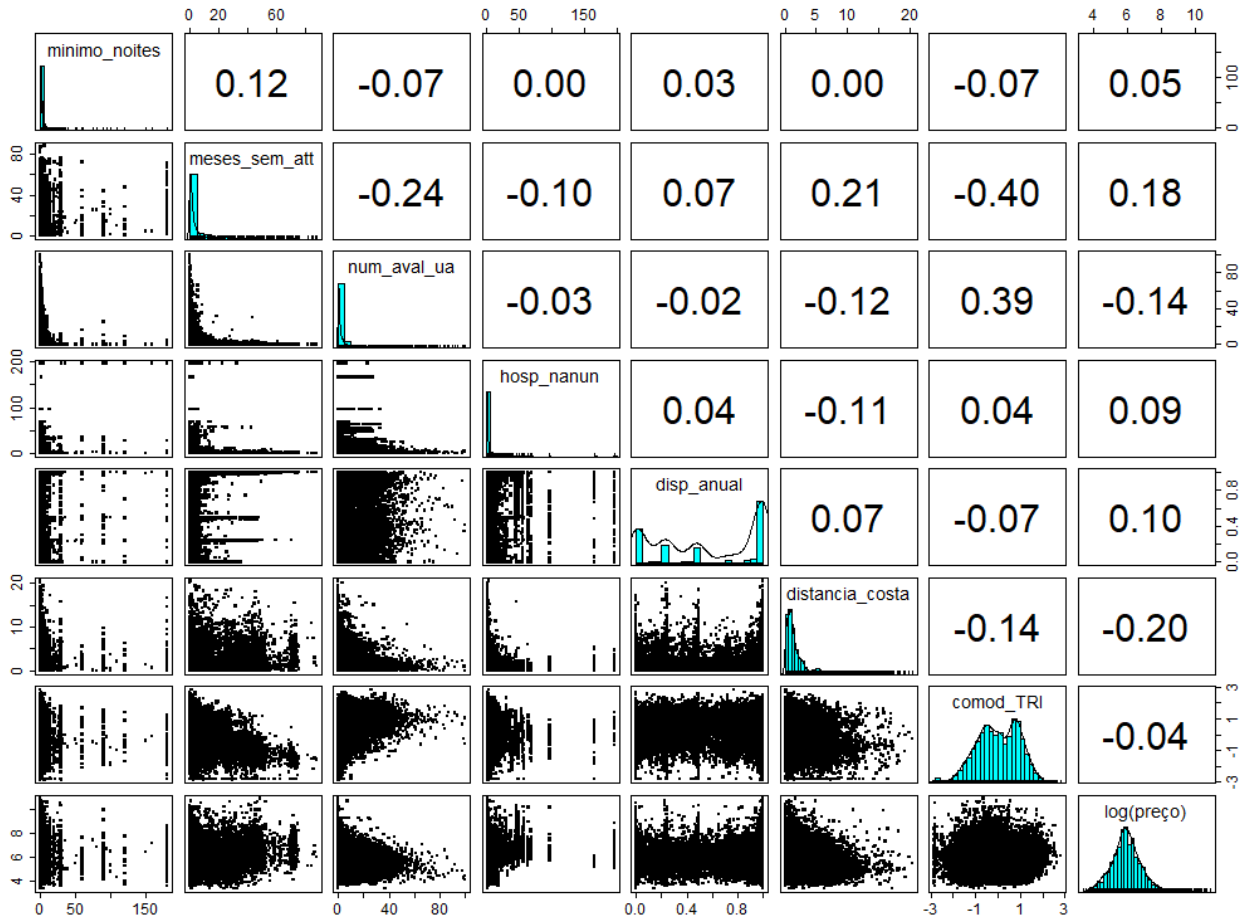


Figura 9 – Matriz de dispersão para os dados do *Airbnb* do Rio de Janeiro com a variável transformada logaritmo do preço do aluguel.

todos os quartis e para a média do preço. O mesmo acontece para o ar condicionado. Portanto, espera-se que as variáveis apresentadas na Tabela 2 ajudem a explicar o preço do imóvel no Rio de Janeiro.

A Figura 10 apresenta um mapa com o preço do aluguel dos imóveis do *Airbnb* no Rio de Janeiro e sua localização geográfica. E na Tabela 3 é apresentado esses os preços de aluguel resumidos em média e quantis por bairro.

Pode-se ver pela Figura 10 que a maior parte das acomodações se encontram próximas da costa do Rio de Janeiro, principalmente na região sul. Em relação aos preços das acomodações por grupos apresentados na Tabela 3, as acomodações que pertencem aos grupos 5 e 15 tiveram o preço bem mais baixo do que o preço dos demais grupos, isso devido ao fato desses grupos serem constituídos principalmente por favelas, como o Complexo do Alemão e Rocinha. Os grupos que se encontram próximos das principais praias, como Ipanema, Barra da Tijuca e Copacabana, apresentam valores de aluguel mais altos.

Verifica-se na Tabela 3 que a maior parte dos bairros possuem o preço da diária máximo ultrapassando o valor de R$ 10.000,00, mas pelo método de regressão quantílica utilizado essas acomodações terão um peso ínfimo na construção dos modelos. Além disso, o maior quantil

Tabela 2 – Preço das variáveis categóricas.

		Preço (R$)						
		Quantil						
Variável	Categoria	10%	25%	50%	75%	90%	máximo	média
hosp_rio	Não	132	242	399	698	1.252	50.734	673
	Sim	122	218	355	606	1.101	41.000	631
hosp_superhost	Não	126	222	378	651	1.152	50.734	637
	Sim	120	218	338	550	1.015	27.244	672
tipo_propriedade	Apartamento	145	246	383	639	1.104	50.734	618
	Casa	81	132	269	700	1.994	35.513	870
	Outros	85	159	279	479	874	41.000	569
piscina	Não	115	202	348	575	999	50.734	538
	Sim	172	291	478	874	1.808	42.000	977
ar_condicionado	Não	76	122	222	401	751	35.513	411
	Sim	161	261	401	692	1.243	50.734	702
res_inst	Não	141	245	399	693	1.280	50.734	717
	Sim	111	200	340	560	999	42.000	543
canc_restrito	Não	103	195	334	586	1.052	50.734	582
	Sim	162	272	408	699	1.248	40.587	717

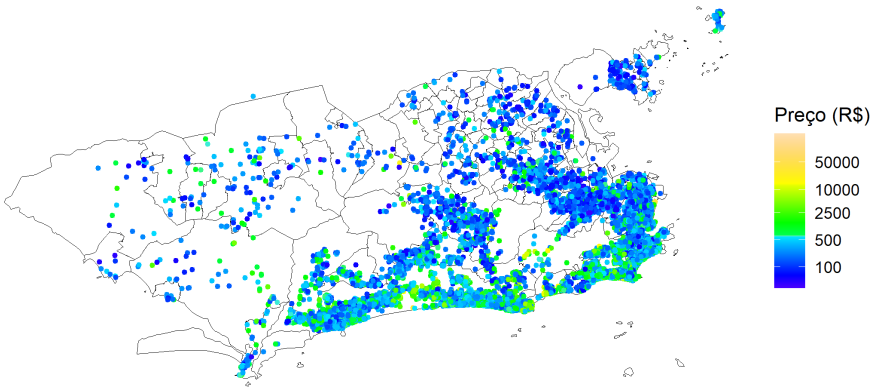


Figura 10 – Preços do aluguel do *Airbnb* no Rio de Janeiro.

modelado foi o de 90%, que tem o maior valor (R$ 5028,00) pertencendo ao grupo 13 (bairros Itanhangá e Joá), que possui apenas 216 anúncios. No geral, os menores valores de aluguel por

Tabela 3 – Preço pelos grupos de bairros.

	Preço (R$)						
	Quantil						
bairro	10%	25%	50%	75%	90%	máximo	média
Grupo 1	132	242	399	652	1.252	15.000	625
Grupo 2	86	139	248	399	739	20.000	521
Grupo 3	91	151	232	358	599	10.147	334
Grupo 4	101	170	309	502	989	42.000	517
Grupo 5	64	91	176	348	600	8.102	308
Grupo 6	71	101	198	381	841	4.221	362
Grupo 7	61	100	172	399	763	11.200	401
Grupo 8	96	152	274	501	1.360	20.000	585
Grupo 9	71	101	192	323	609	20.000	510
Grupo 10	71	108	208	349	652	5.580	331
Grupo 11	142	258	502	1.415	4.344	40.587	2.088
Grupo 12	81	154	278	490	999	3.551	431
Grupo 13	98	194	497	2.298	5.028	27.244	1.908
Grupo 14	73	160	258	599	1.015	5.073	534
Grupo 15	56	98	198	418	761	10.046	419
Grupo 16	164	252	368	600	1.030	50.734	605
Grupo 17	191	299	470	812	1.522	40.000	845
Grupo 18	241	352	536	880	1.502	25.300	838
Grupo 19	101	195	310	500	803	4.444	422
Grupo 20	271	398	591	898	1.699	40.587	903
Grupo 21	117	182	297	482	803	7.103	446
Grupo 22	101	162	299	520	838	6.666	439
Grupo 23	151	272	419	651	999	35.039	696
Grupo 24	76	122	201	401	862	4.833	382
Grupo 25	180	340	546	1.099	1.798	6.849	889
Grupo 26	95	137	263	399	700	41.000	546
Grupo 27	101	170	254	408	608	2.030	338
Grupo 28	183	252	432	763	1.601	10.147	802
Grupo 29	150	227	408	774	1.284	15.144	844
Grupo 30	132	249	368	581	1.012	9.340	540
Grupo 31	183	244	348	457	663	7.610	485
Grupo 32	124	211	409	566	1.417	5.124	608
Grupo 33	122	181	304	502	767	20.120	605

bairro vão de R$ 30,00 a R$ 50,00 a diária.

Os resultados da Tabela 3 sugerem que há correlação espacial na variável preço. Para complementar a análise foi calculado o índice de Moran I.

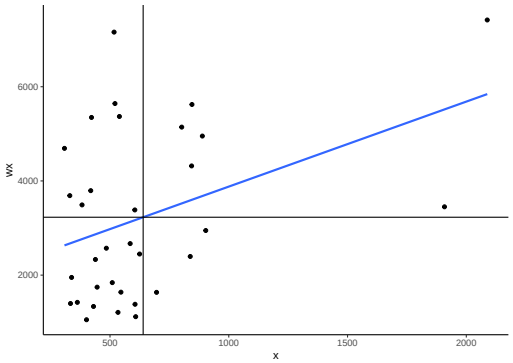
Para o cálculo do índice I de Moran foram usados os agrupamentos de bairros descritos na Tabela 9 (ver Apêndice), sendo que ele foi calculado utilizando as medidas resumos média e quartis. A matriz de proximidade espacial utilizada (vide seção 2.1) foi binária e os resultados são apresentados na Tabela 4. Todas as medidas tiveram o índice positivo e significativo ao

nível de significância de 5%, sendo que para a média e o terceiro quartil a autocorrelação foi fraca e para as demais medidas foi moderada. Ou seja, há evidências de uma estrutura de autocorrelação espacial entre os grupos apresentados para alguns quantis.

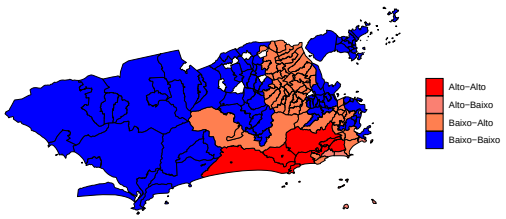
Tabela 4 – Índice I de Moran global para o preço do aluguel.

Medida	Estatística Moran I	Esperança	Variância	p-valor
Média	0,19	-0,031	0,0079	0.006
Quantil 25%	0,48	-0,031	0,0102	<0.001
Quantil 50%	0,46	-0,031	0,0104	<0.001
Quantil 75%	0,22	-0,031	0,0070	0.001

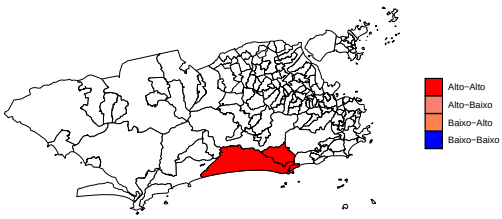
As Figuras 11 a 14 apresentam o diagrama de espalhamento e o mapa de Moran para as medidas resumo média, quantil 25%, 50% e 75%, respectivamente.



(a) Diagrama de espalhamento de Moran



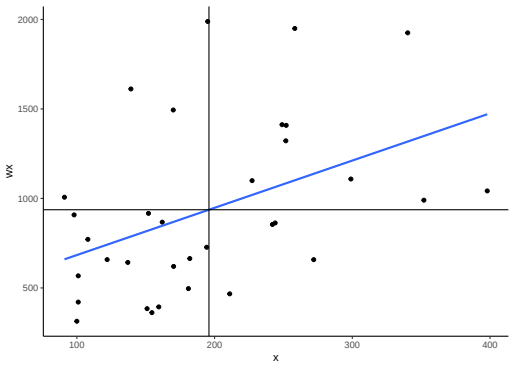
(b) Mapa de espalhamento de Moran



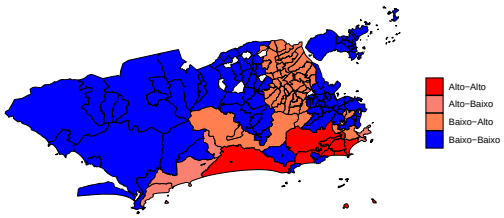
(c) Mapa de Moran 95%

Figura 11 – Estrutura de correlação espacial para a média do preço de aluguel.

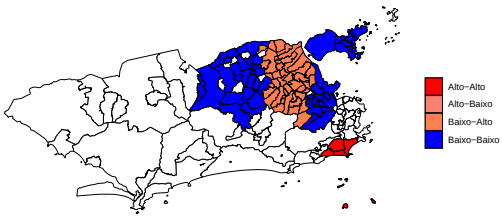
Os diagramas de espalhamento de Moran utilizando a média e o terceiro quartil (ilustrados nas Subfiguras 11a e 14a, respectivamente) ficaram muito parecidos, isso ocorre pois essas medidas foram muito próximas como é indicado na Tabela 4. Nota-se que dois grupos tiveram o preço alto, sendo eles 13 e 11, como pode ser visto na Tabela 3 e estão interferindo no valor do índice. Enquanto que para a mediana e primeiro quartil (ilustrados nas Subfiguras 13a e 12a, respectivamente), esses grupos, mesmo ainda pertencendo ao quadrante alto-alto, não se qualificam como um *outlier* e influentes.



(a) Diagrama de espalhamento de Moran

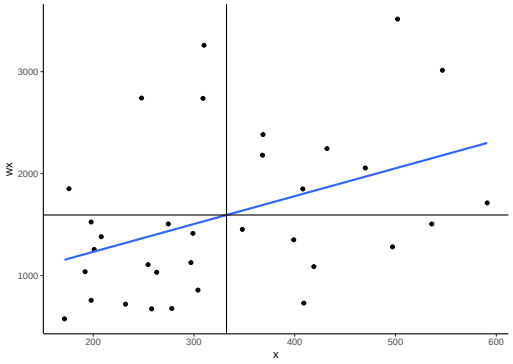


(b) Mapa de espalhamento de Moran

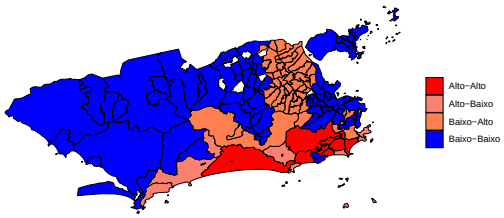


(c) Mapa de Moran 95%

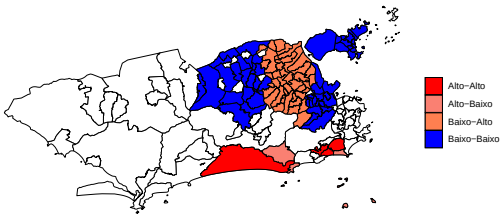
Figura 12 – Estrutura de correlação espacial para o quantil 25% do preço de aluguel.



(a) Diagrama de espalhamento de Moran

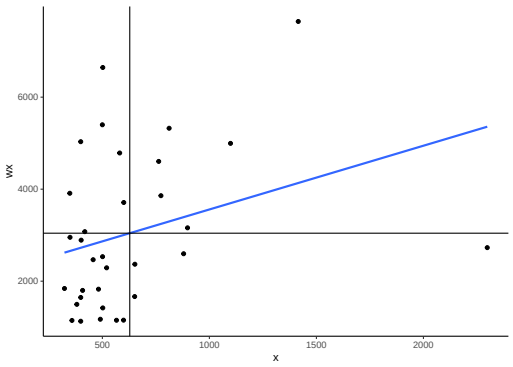


(b) Mapa de espalhamento de Moran

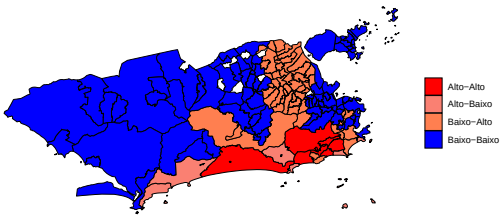


(c) Mapa de Moran 95%

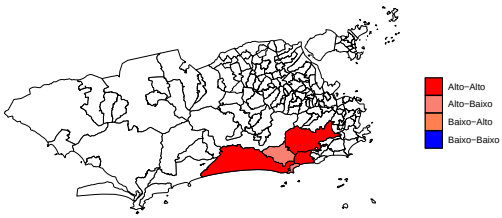
Figura 13 – Estrutura de correlação espacial para quantil de 50% do preço de aluguel.



(a) Diagrama de espalhamento de Moran



(b) Mapa de espalhamento de Moran



(c) Mapa de Moran 95%

Figura 14 – Estrutura de correlação espacial para o quantil de 75% do preço de aluguel.

O mapa para o preço das Figuras 11 a 14 tem uma predominância de valores baixo-baixo na região oeste e nordeste e alto-alto se encontram na região sudeste. Nota-se que, para a média e o terceiro quartil, poucos bairros foram significativos. Para a média (Figura 11c), apenas a Barra da Tijuca, Joá e Itanhangá foram significativos, e para o terceiro quartil, além dos bairros que foram significativos para a média, tiveram o Alto da Boa Vista e São Conrado. A presença de correlações significativas no mapa indica evidências de correlação espacial nos dados, sugerindo que um modelo espacial proporcionará melhor ajuste.

6.2 Modelos

6.2.1 Regressão quantílica não espacial

Apesar da análise descritiva indicar presença de dependência espacial, um modelo não espacial será primeiramente ajustado.

Seja y o preço do imóvel e $\boldsymbol{x}$ o vetor de dimensão $p = 22$ das covariáveis apresentadas na Tabela 1 mais o intercepto. Foram ajustados os modelos a seguir para $\tau = \{0,05; 0,10; \ldots; 0,95\}$:

1. $y_i = \boldsymbol{x_i}\boldsymbol{\beta} + \varepsilon_i, \quad P(Y \leq \boldsymbol{x}\boldsymbol{\beta}) = \tau;$

2. $\log y_i = \boldsymbol{x_i}\boldsymbol{\beta} + \varepsilon_i, \quad P(\log Y \leq \boldsymbol{x}\boldsymbol{\beta}) = \tau;$

Na Tabela 5 é apresentado o coeficiente de correlação linear R para os dois modelos descritos acima, com a variável dependente preço e log do preço para os quantis $\tau = \{0{,}05; 0{,}10; 0{,}25; 0{,}50; 0{,}75; 0{,}90; 0{,}95\}$.

Tabela 5 – Estatística R da regressão com as variáveis dependentes preço e log preço.

Quantil	Preço	Log Preço
5%	0,85	0,86
10%	0,72	0,72
25%	0,38	0,39
50%	0,21	0,30
75%	0,27	0,39
90%	0,42	0,69
95%	0,64	0,86

Os resultados apresentados na Tabela 5 mostram que a transformação logarítmica fornece um melhor ajuste no geral, principalmente para os quantis mais altos. Por isso, a variável resposta utilizada na construção dos modelos será o logaritmo do preço, e os valores estimados serão exponenciados para que o preço seja apresentado na sua escala original.

Na Tabela 6 são apresentados os coeficientes da regressão quantílica para os quantis $\tau = \{0{,}05; 0{,}25; 0{,}50; 0{,}75; 0{,}95\}$ e da regressão de mínimos quadrados ordinários (MQO), sendo que a variável resposta utilizada é o logaritmo do preço. Na Figura 15 esses coeficientes são ilustrados em subgráficos para melhor visualização.

Nota-se pela Figura 15b que para a variável comod_TRI, quanto maior o número de comodidades que a acomodação possui maior é o seu valor. O seu comportamento muda ao atingir o quantil de 30%, sendo que antes desse quantil ela assume valores positivos e após ela fica negativa. Logo, até o quantil de 30% um maior número de comodidades contribuem para um aumento do preço da acomodação, e após esse quantil um maior número de comodidades reduzem o valor da acomodação.

Verifica-se, na Figura 15, a existência de coeficientes que possuem pouca variação à medida que muda o quantil ajustado e que acabam se comportando como uma reta, como o número de anúncios do hospedeiro e depósito de segurança. Isso indica que esses coeficientes possuem o mesmo impacto para acomodações de valor baixo a alto.

O número máximo de hóspedes aceito pelo anúncio e o seu número de quartos também não aparentam sofrer uma mudança grande no seu valor para os diferentes quantis de preço, porém, o fato do anúncio aceitar um hóspede a mais consegue aumentar o seu valor em 12%, e o aumento de um quarto em 21% de acordo com o modelo estimado. O número de banheiros sofre grandes mudanças dependendo do quantil do preço estudado. Para as acomodações de preços mais baixos provoca uma redução média de 2%, para as de valor médio um aumento médio de 11% e de 23% para as de valores elevados. Essas mudanças nos valores dos coeficientes com base no quantil de interesse reforçam a indicação de um modelo de regressão quantílica ao invés da regressão de mínimos quadrados ordinários.

Tabela 6 – Estimativas do modelo linear para log do preço.

Variável	5%	25%	50%	75%	95%	MQO
Intercepto	3,2079*	4,1413*	4,6228*	5,0772*	5,7689*	4,5763*
hosp_tempo	0,0379*	0,0360*	0,0286*	0,0267*	0,0119*	0,0319*
hosp_rio	-0,0721*	-0,0769*	-0,0646*	-0,0635*	-0,0837*	-0,0795*
hosp_superhost	-0,0268	0,0322†	0,0341*	0,0577*	0,0786*	0,0470*
tipo_propriedadeApartamento	0,6190*	0,2306*	0,1220*	0,0450‡	-0,0399	0,1844*
tipo_propriedadeCasa	0,2091†	-0,1215*	-0,1374*	-0,1245*	-0,1882*	-0,0826*
max_hospedes	0,1005*	0,1071*	0,1056*	0,1017*	0,0963*	0,1056*
banheiros	-0,0192	0,0377*	0,1031*	0,1408*	0,2030*	0,0777*
quartos	0,2829*	0,2281*	0,1896*	0,1938*	0,2069*	0,2139*
camas	-0,0668*	-0,0460*	-0,0371*	-0,0364*	-0,0329*	-0,0458*
piscina	0,2201*	0,1823*	0,1763*	0,1839*	0,1782*	0,1998*
ar_condicionado	0,4731*	0,4085*	0,3273*	0,2364*	0,1797*	0,3356*
dep_seg	0,0001*	0,0001*	0,0001*	0,0001*	0,0001*	0,0001*
minimo_noites	-0,0038*	-0,0030*	-0,0017†	0,0000	0,0091*	-0,0015*
meses_sem_att	0,0056*	0,0090*	0,0117*	0,0136*	0,0160*	0,0113*
num_aval_ua	-0,0057*	-0,0096*	-0,0108*	-0,0118*	-0,0135*	-0,0113*
res_inst	0,0020	0,0086	0,0182†	0,0333*	0,0469*	0,0109
canc_restrito	0,1517*	0,1247*	0,0890*	0,0297*	-0,0238	0,0784*
hosp_nanun	0,0006*	-0,0000	-0,0005*	-0,0006*	-0,0017*	-0,0001
disp_anual	0,1200*	0,1198*	0,1009*	0,0843*	0,1001*	0,1257*
distancia_costa	-0,0868*	-0,0808*	-0,0708*	-0,0676*	-0,0648*	-0,0697*
comod_TRI	0,0205‡	0,0020	-0,0202*	-0,0554*	-0,1349*	-0,0411*

* p-valor<0,01
† p-valor<0,05
‡ p-valor<0,1

As covariáveis que mais afetam o preço da acomodação são as que informam diretamente o tamanho dela, sendo o máximo de hóspedes, número de banheiros e de quartos. Agora, a quantidade de camas acabou reduzindo o valor do preço. Isso se deve ao fato de que as acomodações que possuem mais camas do que quartos tendem a ser mais baratas.

Para todos os níveis dos quantis estudados, o impacto no preço de ter ar-condicionado na acomodação é maior do que o de uma piscina, principalmente para os quantis mais baixos. Por exemplo, no preço mediano, o fato de ter ar condicionado é capaz de aumentar o valor da diária em média 39%, enquanto uma piscina aumenta em 19% em média.

Propriedades em que o hospedeiro mora no Rio de Janeiro tendem a possuir preços menores para todos os quantis de preços estudados, como pode-se ver pelo valor negativo do coeficiente. Para a mediana, por exemplo, acomodações em que o hospedeiro mora no Rio de Janeiro tendem a ser em média 6% mais baratas em relação às acomodações dos que moram fora.

A predominância de coeficientes negativos para a variável tipo_propriedadeCasa mostra que os aluguéis de casas acaba sendo uma alternativa bem mais econômica em relação aos demais tipos de propriedade, e os apartamentos são os mais caros, exceto para os quantis superiores

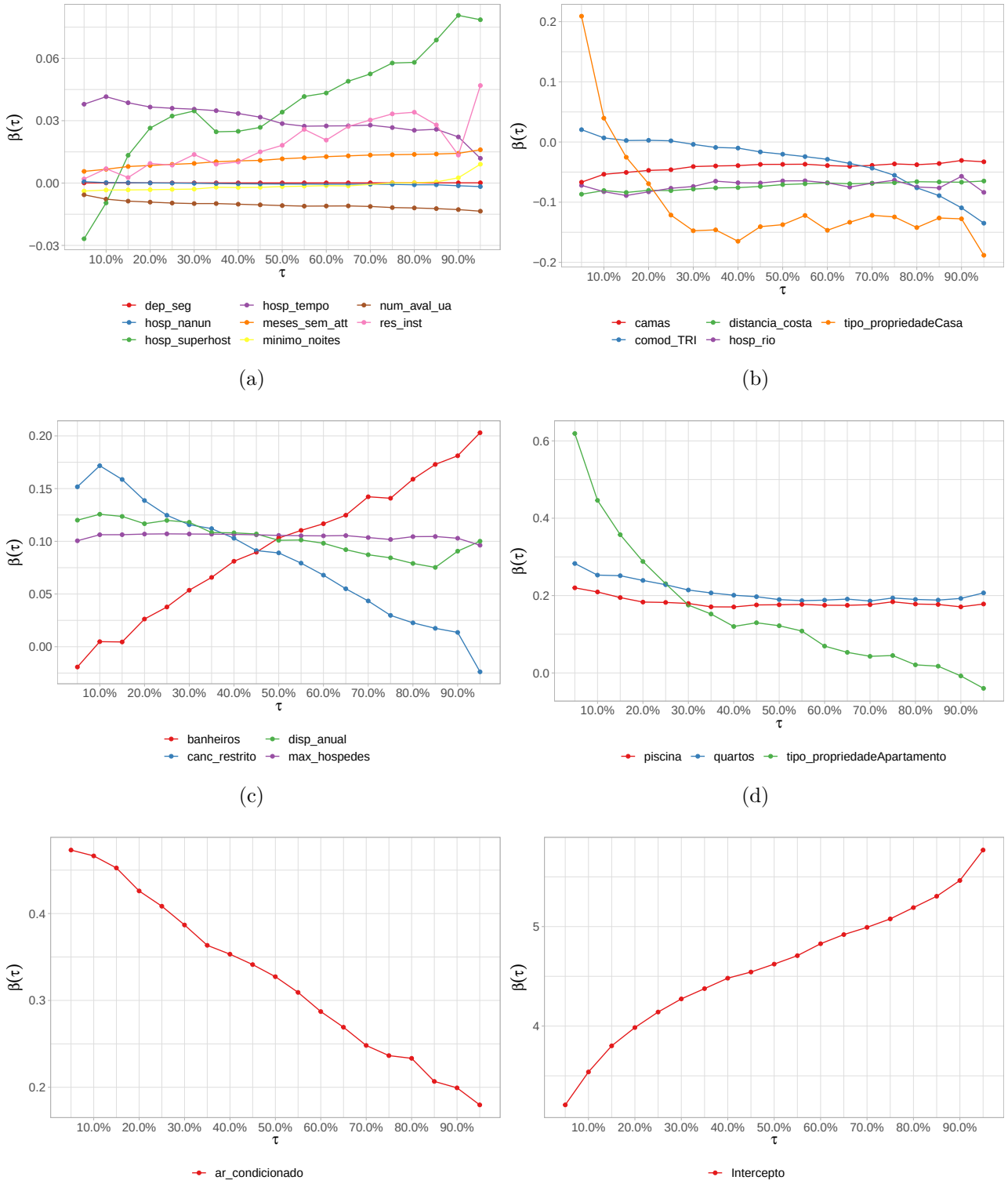


Figura 15 – Coeficiente por quantil.

extremos.

Para a mediana do preço, se uma acomodação se encontra um quilômetro mais próxima da costa há um aumento médio de 7% em seu valor de aluguel.

6.2.2 Regressão quantílica espacial

Tendo em vista a presença de dependência espacial, o modelo de regressão quantílica espacial proposto por Reich et al. (2011) será ajustado aos dados.

Seja y o preço do imóvel, $\boldsymbol{x}$ o vetor de dimensão $p = 22$ das covariáveis apresentadas na Tabela 1 mais o intercepto, e $\boldsymbol{s} = \{s_1, s_2, \ldots, s_{33}\}$ o vetor indicando o sítio do imóvel conforme a Tabela 9. Foi ajustado o modelo para os quantis $\tau = \{0{,}10; 0{,}20; \ldots; 0{,}90\}$, totalizando 6.534 $(\dim(\boldsymbol{s}) \times \dim(\tau) \times p)$ coeficientes estimados no primeiro estágio, no qual foi aplicada a regressão quantílica tradicional para cada grupo separadamente:

$$
\begin{aligned}
(\hat{\beta}_1(\tau_k, \boldsymbol{s}), \ldots, \hat{\beta}_p(\tau_k, \boldsymbol{s}))' & \\
&= \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \sum_{s_i = s, \log y_i > \boldsymbol{X_i^{*\prime}}\beta} \tau_k |\log y_i - \boldsymbol{X_i^{*\prime}}\boldsymbol{\beta}| \\
&+ \sum_{s_i = s, \log y_i < \boldsymbol{X_i^{*\prime}}\beta} (1 - \tau_k) |\log y_i - \boldsymbol{X_i^{*\prime}}\boldsymbol{\beta}|,
\end{aligned}
$$

em que $\boldsymbol{X_i^{*\prime}}$ é a covariável transformada para que pertença ao intervalo unitário, conforme sugerido em Reich et al. (2011):

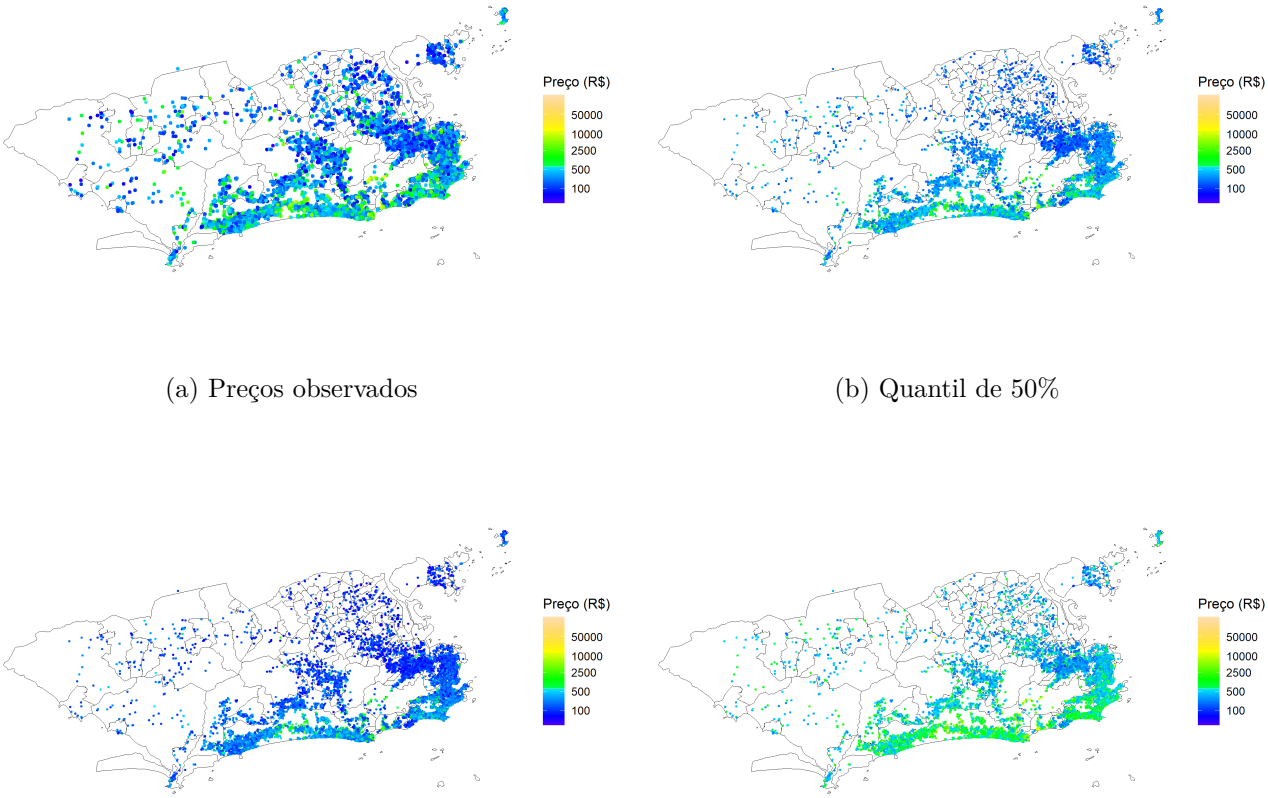
- Variável x discreta: subtraiu-se o mínimo e dividiu-se pela amplitude, ou seja, $x_i^* = \dfrac{x_i - min(x)}{max(x) - min(x)}$.
- Variável x contínua: Padronizou-se x e utilizou-se o valor da função de distribuição acumulada (FDA) da Normal , ou seja, $x_i^* = \Phi\left(\dfrac{x_i - \bar{x}}{s(x)}\right)$, em que $s(x)$ é o desvio padrão de x e $\Phi(\cdot)$ é a FDA da Normal padrão.

Em seguida, as estimativas do segundo estágio foram obtidas utilizando o modelo definido na Equação (4.16).

Na Figura 16 são apresentados o preço observado e o ajustado para 3 quantis ($\tau = \{0{,}1; 0{,}5; 0{,}9\}$) com o modelo espacial. Consegue-se ver que as regiões com as acomodações mais caras se encontram na região sudeste, que é um dos lugares mais populares do Rio de Janeiro por causa das praias. Além disso, para um apartamento com as seguintes características principais: um quarto, um banheiro, com ar-condicionado, no bairro de Copacabana, a chance de alugar com preço inferior a R$ 447,97 é de 90% e com o preço inferior a R$ 181,61 é de 10% apenas, de acordo com o modelo estimado.

A fim de visualizar as diferenças regionais, a Figura 17 apresenta preços do apartamento descrito acima variando a sua localização no mapa. Verifica-se um comportamento bastante comum em relação aos demais mapas com preço, com a região sudeste sendo a mais cara, e as acomodações que se encontram no nordeste sendo mais baratas. Em Copacabana, por exemplo, há uma chance de 50% de alugar um apartamento para um hóspede com um quarto, uma cama, e com ar-condicionado em Copacabana por até R$ 238,13, enquanto que um apartamento assim

na Ilha do Governador (pertence ao grupo 6, vide Tabela 9) custa até R$ 148,81, com base no modelo estimado.



(a) Preços observados

(b) Quantil de 50%

(c) Quantil de 10%

(d) Quantil de 90%

Figura 16 – Valores observados e quantis ajustados para o preço de aluguel.

Tem-se que a Ilha do Governador acaba sendo uma opção interessante e econômica para se viajar com o preço mediano de R$ 149 para um apartamento de solteiro, além de ter várias atrações como por exemplo a praia da Bica.

Na Figura 18 são apresentados os mapas dos valores dos coeficientes ajustados para a mediana por grupo geográfico.

Nos mapas apresentados, verifica-se que existem diversos coeficientes que variam muito pouco entre os bairros, como número de camas, quartos e número de avaliações dos usuários. Com isso, para esses coeficientes, seu impacto é muito parecido entre os bairros.

Apesar do número de quartos variar pouco no espaço geográfico, ele ainda possui um alto impacto no preço. No geral, o aumento de um quarto é capaz de aumentar o valor do aluguel mediano em 18%.

Lugares em que o preço mediano do aluguel de casa é mais barato em relação a outros tipos de acomodação são os bairros de Leme, Lagoa, Glória e Catumbi, que tendem a ser 10%

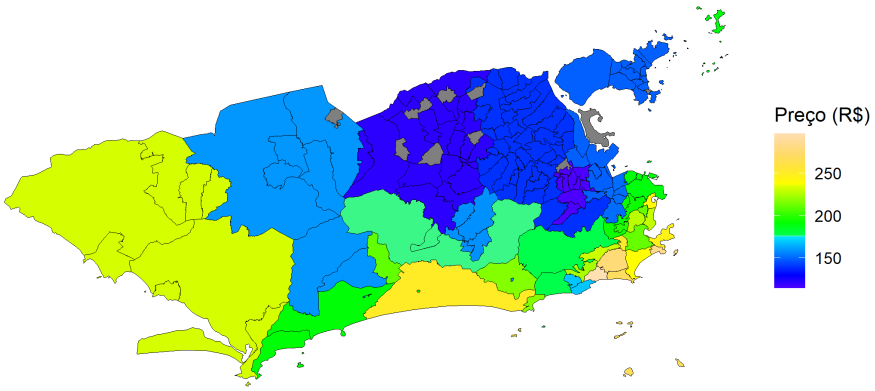


Figura 17 – Preço mediano estimado de um apartamento para um hóspede com um quarto, uma cama, e com ar-condicionado.

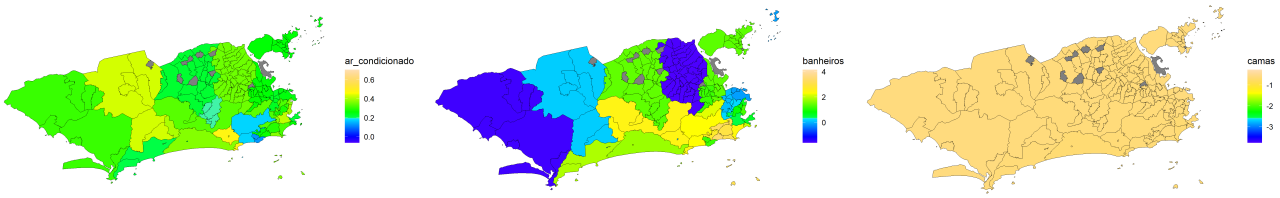
mais baratas. Em contrapartida, o preço mediano do aluguel de casa é de 15% a 24% mais caro para Leblon, Humaitá, Estácio e entorno.

O preço mediano do aluguel de apartamentos no geral é mais caro, podendo ser até 52% mais caro do que os outros tipos de propriedade. Vidigal e Rocinha foram os bairros com os apartamentos mais caros comparativamente com os demais tipos de acomodações. Enquanto que o Camorim não tem uma diferença significativa do preço do apartamento em relação aos outros tipos de propriedade, e a Tijuca possui um leve encarecimento de 5% em seus apartamentos em relação aos outros tipos de acomodações de acordo com o modelo estimado.

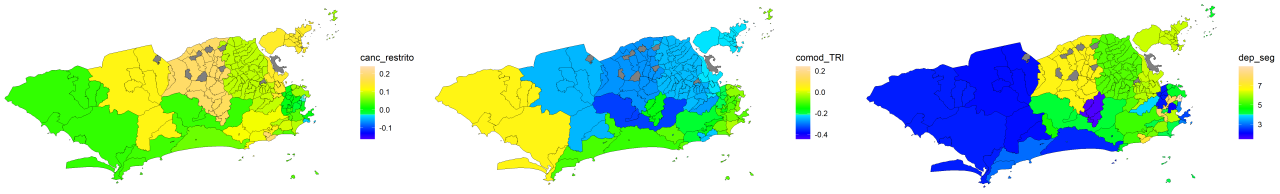
Para o intercepto (ilustrado na Figura 18m), que estabelece o preço mediano base da acomodação para dada região, verifica-se que a região nordeste apresenta os menores valores no geral. Enquanto que as regiões que apresentaram os maiores valores são Leblon, Barra da Tijuca e Copacabana (grupos 20, 17 e 16, respectivamente).

O aumento do número de banheiros teve um efeito punitivo para os grupo 5 e 15 (veja a Tabela 9 para ver os bairros que os compõem), visto que há uma desvalorização de 8%, em média, com o aumento de um banheiro (com tudo mais constante). Entretanto, no geral, ele consegue gerar uma boa valorização (em torno de 7%) no preço do aluguel, principalmente para Ipanema, Jardim Botânico e Leblon. Porém, isso não significa que simplesmente ter um banheiro a mais é o motivo do preço ser menor ou maior, ou seja, não é uma relação de causa e efeito necessariamente.

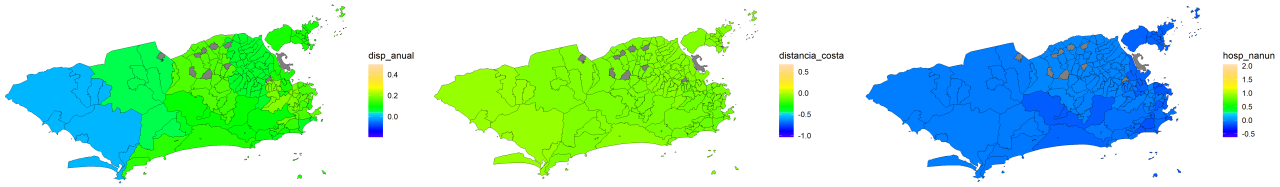
Além disso, os proprietários que tem uma política de cancelamento da hospedagem mais restrita tendem a cobrar mais caro o valor do aluguel. Os grupos que tem o maior aumento são os números 12, 6 e 15, que chegam a cobrar 20% a mais no valor da hospedagem.



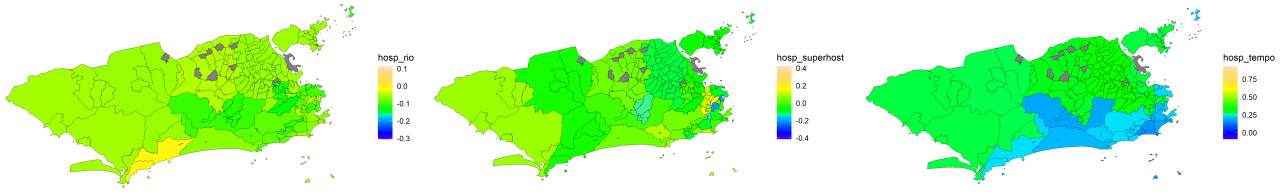
(a) ar_condicionado (b) banheiros (c) camas



(d) canc_restrito (e) comod_TRI (f) dep_seg

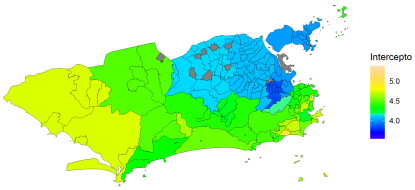


(g) disp_anual (h) distancia_costa (i) hosp_nanun

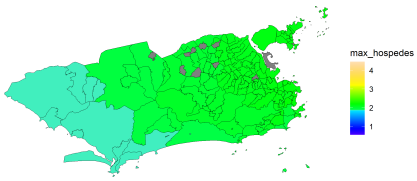


(j) hosp_rio (k) hosp_superhost (l) hosp_tempo

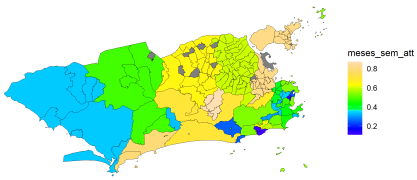
Figura 18 – Mapas de coeficientes β da regressão para a mediana.



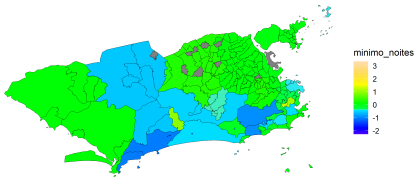
(m) Intercepto



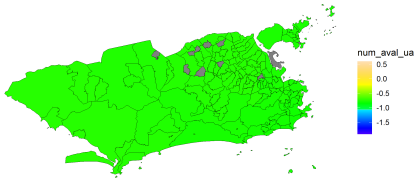
(n) max_hospedes



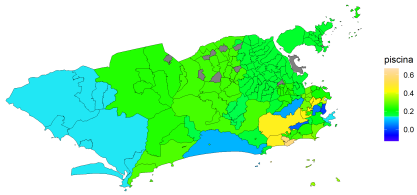
(o) meses_sem_att



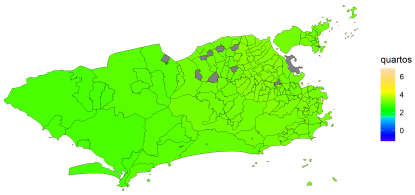
(p) minimo_noites



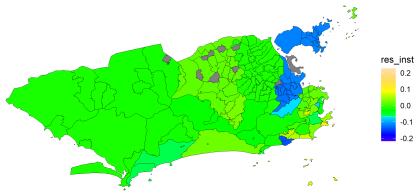
(q) num_aval_ua



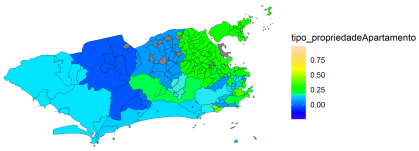
(r) piscina



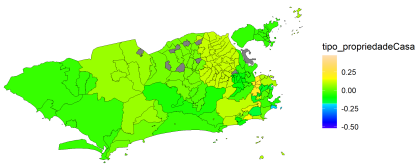
(s) quartos



(t) res_inst



(u) tipo_propriedadeApartamento



(v) tipo_propriedadeCasa

Figura 18 – (continuação) Mapas de coeficientes β da regressão para a mediana.

Na maior parte dos bairros, os hospedeiros que tem o certificado de *superhost* do *Airbnb* tendem a fornecer o valor da hospedagem mais barato, principalmente os que se localizam nos bairros Glória e Laranjeiras, que tendem a ser 19,5% mais baratos em relação aos preços medianos dos hospedeiros que não são *superhosts*.

Piscina mostrou-se ser um diferencial que tende a aumentar o valor da acomodação e impacta de forma muito diferente cada região. Por exemplo, uma piscina em Copacabana consegue aumentar o valor do aluguel em 35,7%, enquanto que no Vidigal e Rocinha aumenta em 81,6%, em média.

Na Figura 19, no Apêndice C, são apresentados mapas de coeficientes para a mediana do primeiro estágio, ou seja, antes de ser feita a suavização dos coeficientes apresentados na Figura 18. A escala de cores das duas figuras estão fixas entre os coeficientes para facilitar a comparação. No geral, ao comparar os diversos mapas de coeficientes, verifica-se que os valores das estimativas se tornam mais homogêneos entre os sítios estudados.

No Apêndice D, são apresentados os mapas de coeficientes para os quantis 10% e 90% do logaritmo do preço. Comparando os dois conjuntos de figuras observa-se diversos padrões em comum com os coeficientes da mediana, por exemplo, o valor do intercepto continua menor para a região nordeste, especialmente para a Ilha do Governador.

6.2.3 Comparação entre os modelos

Serão comparados os modelos de mínimos quadrados ordinários e os de regressão quantílica ajustados para o logaritmo do preço do aluguel. Será utilizado o logaritmo pois mesmo o modelo de regressão quantílica tendo propriedade de invariância a transformações monótonas, o modelo de mínimos quadrados ordinários não possui essa propriedade.

Na Tabela 7 é apresentado o erro quadrático médio (EQM) e o erro médio absoluto (EMA) dos modelos ajustados, em que o EQM é calculado pela média dos resíduos ao quadrado e o EMA é calculado pela média dos resíduos absolutos. Os modelos apresentados na tabela são: regressão quantílica (RQ) apresentado na subseção 6.2.1, regressão quantílica espacial Bayesiana (RQE) apresentado na subseção 6.2.2 e de mínimos quadrados ordinários (MQO) apresentado na mesma subseção do modelo de regressão quantílica.

Devido à presença de correlação espacial e assimetria nos dados de preço de aluguel, o modelo de regressão quantílica espacial melhor descreve os dados analisados, como é indicado na Tabela 7.

Tabela 7 – Comparação dos modelos ajustados.

	EQM			EMA		
Quantil	RQ	RQE	MQO	RQ	RQE	MQO
0,1	1,014	0,674	-	0,817	0,630	-
0,2	0,661	0,571	-	0,624	0,572	-
0,3	0,513	0,497	-	0,534	0,526	-
0,4	0,440	0,431	-	0,487	0,484	-
mediana/média	0,417	0,389	0,415	0,473	0,457	0,474
0,6	0,437	0,390	-	0,487	0,461	-
0,7	0,504	0,432	-	0,532	0,492	-
0,8	0,654	0,519	-	0,626	0,553	-
0,9	1,009	0,743	-	0,823	0,694	-

7 Conclusão

O presente trabalho teve como objetivo analisar os preços dos aluguéis do site *Airbnb* para a cidade do Rio de Janeiro utilizando regressão quantílica. Observaram-se diversos problemas nos anúncios do site, em que os hospedeiros mantinham o anúncio no site, mesmo não podendo alugar para nenhuma data, preços elevados de diária, que tornavam inviável o aluguel, entre outros.

Após um longo tratamento no conjunto de dados, foi possível construir uma base de dados confiável para aplicar técnicas estatísticas para analisar os anúncios. Com isso, foi possível obter algumas conclusões importantes, como, por exemplo, os principais aspectos que determinam o preço da acomodação são o número de quartos, banheiros e o máximo de hóspedes.

Foi observado na seção 6.1 que o valor do anúncio está associado também ao bairro em que a acomodação se encontra, o que justificou o uso de modelos espaciais para melhor análise dos dados.

Com o modelo espacial foi possível observar que existem aspectos da acomodação que impactam de forma semelhante entre os bairros, como é o caso do número de quartos, que aumenta o valor do aluguel a uma taxa semelhante entre os bairros considerando as demais covariáveis constantes no modelo. Já o aumento no preço devido ao acréscimo no número de banheiros mostrou-se depender da localização geográfica. E, que no geral a região sudeste do Rio de Janeiro, além de ser a que possui o maior número anúncios, é a que apresenta os aluguéis mais caros, em contrapartida, a região nordeste foi a que se apresentou como a alternativa mas econômica.

Como o conjunto de dados ainda apresenta preços de aluguel em um valor elevado, seria um tema para trabalhos futuros utilizar apenas acomodações alugadas recentemente, o que melhoraria a qualidade dos dados para apenas acomodações realmente utilizadas, poupando o serviço de olhar o tempo em que o hospedeiro não atualiza e reduzindo a discrepância nos valores de aluguel. Seria também interessante ajustar o modelo com outras combinações de covariáveis para avaliar a mudança na estimação dos coeficientes da regressão. Por fim, variar os valores definidos para as prioris utilizadas no modelo espacial Bayesiano também seria importante para uma análise de sensibilidade.

Referências

ANDRADE, D. F. de; TAVARES, H. R.; VALLE, R. da C. *Teoria da Resposta ao Item: conceitos e aplicações*. [S.l.]: Associação Brasileira de Estatística, 2000.

CLIFF, A. D.; ORD, J. K. *Spatial processes: models & applications*. [S.l.]: Taylor & Francis, 1981.

DRUCK, S.; CARVALHO, M. S.; CÂMARA, G.; MONTEIRO, A. M. V. *Análise espacial de dados geográficos*. Planaltina: Empraba Cerrados, 2004. ISBN 8573832606. Disponível em: <http://www.dpi.inpe.br/gilberto/livro/analise/>.

DUNSON, D. B.; PARK, J.-H. Kernel stick-breaking processes. *Biometrika*, Oxford University Press, v. 95, n. 2, p. 307–323, 2008.

GELFAND, A. E.; KOTTAS, A.; MACEACHERN, S. N. Bayesian nonparametric spatial modeling with dirichlet process mixing. *Journal of the American Statistical Association*, Taylor & Francis, v. 100, n. 471, p. 1021–1035, 2005.

GRIFFIN, J. E.; STEEL, M. J. Order-based dependent dirichlet processes. *Journal of the American statistical Association*, Taylor & Francis, v. 101, n. 473, p. 179–194, 2006.

HAO, D. Q. N. L. *Quantile Regression*. [S.l.]: SAGE PUBN, 2007. ISBN 1412926289.

HOPE, A. C. A. A simplified monte carlo significance test procedure. *Journal of the Royal Statistical Society: Series B (Methodological)*, Wiley, v. 30, n. 3, p. 582–598, sep 1968.

KOENKER, R. *Quantile Regression*. [S.l.]: Cambridge University Press, 2005. (Econometric Society Monographs).

KOENKER, R. *quantreg: Quantile Regression*. [S.l.], 2020. R package version 5.75. Disponível em: <https://CRAN.R-project.org/package=quantreg>.

KOENKER, R.; BASSETT, G. Regression quantiles. *Econometrica: journal of the Econometric Society*, JSTOR, p. 33–50, 1978.

KUTNER, M. H.; NACHTSHEIM, C. J.; NETER, J.; LI, W. et al. Applied linear statistical models. McGraw-Hill Irwin New York, v. 5, 2005.

MORAN, P. A. P. Notes on continuous stochastic phenomena. *Biometrika*, JSTOR, v. 37, n. 1/2, p. 17, jun 1950.

ODOM, M. D.; SHARDA, R. A neural network model for bankruptcy prediction. In: IEEE. *1990 IJCNN International Joint Conference on neural networks*. [S.l.], 1990. p. 163–168.

ORDIERES, J.; VERGARA, E.; CAPUZ, R.; SALAZAR, R. Neural network prediction model for fine particulate matter (pm2. 5) on the us–mexico border in el paso (texas) and ciudad juárez (chihuahua). *Environmental Modelling & Software*, Elsevier, v. 20, n. 5, p. 547–559, 2005.

PEREIRA, R. H. M.; GONCALVES, C. N. *geobr: Loads Shapefiles of Official Spatial Data Sets of Brazil*. [S.l.], 2021. R package version 1.5-1. Disponível em: <https://CRAN.R-project.org/package=geobr>.

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2020. Disponível em: <https://www.R-project.org/>.

RATHER, A. M.; AGARWAL, A.; SASTRY, V. Recurrent neural network and a hybrid model for prediction of stock returns. *Expert Systems with Applications*, Elsevier, v. 42, n. 6, p. 3234–3241, 2015.

REICH, B. J.; FUENTES, M.; DUNSON, D. B. Bayesian spatial quantile regression. *Journal of the American Statistical Association*, Taylor & Francis, v. 106, n. 493, p. 6–20, 2011.

REICH, B. J.; FUENTES, M. et al. A multivariate semiparametric bayesian spatial modeling framework for hurricane surface wind fields. *The Annals of Applied Statistics*, Institute of Mathematical Statistics, v. 1, n. 1, p. 249–264, 2007.

RODRIGUES, T. C. V. Regressão binomial negativa geograficamente ponderada: modelando superdispersão espacial. 2012.

APÊNDICE A – Comodidades da variável comod_TRI

Tabela 8 – Comodidades de comod_TRI.

Comodidade	Quantitativo	Percentual	a	b
Cozinha	26560	91,1	0,7299	2,54
Internet	26494	90,8	0,7826	2,54
Essenciais	25077	86,0	0,9984	2,14
TV	24976	85,6	0,5425	1,89
Cabide	20304	69,6	1,2117	1,04
Elevador	19782	67,8	0,3364	0,76
Ferro	19064	65,4	0,9759	0,74
Lava-Roupa	18671	64,0	0,2967	0,59
Espaço para laptop	16878	57,9	0,7965	0,35
Água quente	16216	55,6	1,9741	0,27
Prato e talher	12560	43,1	5,6030	-1,50
Geladeira	12074	41,4	7,3426	-2,30
Secador de cabelo	11804	40,5	0,8496	-0,46
TV a cabo	11217	38,5	0,7007	-0,53
Espaço familiar	11045	37,9	0,1976	-0,50
Microondas	10789	37,0	4,5634	-1,98
Fogão	10487	36,0	6,6305	-3,06
Permitido Fumar	10388	35,6	-0,0758	-0,59
Estacionamento Grátis	10052	34,5	0,1168	-0,64
Equipamento básico de cozinha	10039	34,4	5,5247	-2,78
Roupa de cama	9910	34,0	2,5102	-1,39
Cafeteira	9299	31,9	3,8391	-2,24
Quarto com tranca	9166	31,4	0,4278	-0,81
Forno	8969	30,7	4,4926	-2,73
Shampoo	7891	27,1	0,6046	-1,07
Extintor de incêndio	7648	26,2	0,4351	-1,08
Porteiro	7637	26,2	0,0955	-1,04
Interfone	7461	25,6	0,0853	-1,07
Estadia longa permitida	7199	24,7	2,2134	-1,98
Hospedeiro te cumprimenta	6329	21,7	0,8910	-1,49
Travesseiro e lenço extra	5956	20,4	2,4012	-2,49

Tabela 8 – Comodidades de comod_TRI (Continuação).

Comodidade	Quantitativo	Percentual	a	b
Entrega de bagagem permitida	5772	19,8	1,9325	-2,22
Animais permitidos	5489	18,8	0,0389	-1,46
Entrada privada	5467	18,7	0,4750	-1,53
Secadora	4670	16,0	0,4059	-1,71
Estacionamento na rua gratuito	4651	15,9	1,4195	-2,21
Academia	4522	15,5	0,2892	-1,72
Kit de primeiros socorros	4370	15,0	0,5463	-1,83
Estacionamento pago	4253	14,6	1,4388	-2,35
Sacada	4021	13,8	2,1779	-3,00
Detector de fumaça	3712	12,7	0,8425	-2,16
Check-in próprio	3353	11,5	1,1739	-2,48
De frente para praia	2834	9,7	1,0044	-2,58
tom de escurecimento da sala	2825	9,7	2,1644	-3,52
Check-in 24 horas	2786	9,6	0,3667	-2,30
Faz tudo	2591	8,9	1,0546	-2,72
Sala de estar privada	2367	8,1	0,6688	-2,60
Acessível a cadeirante	2304	7,9	0,1722	-2,47
Detector de monóxido de carbono	2223	7,6	1,2102	-3,01
Banheira	2140	7,3	0,5043	-2,64
Apropriado para evento	2124	7,3	-0,0836	-2,55
Lixeira	2027	6,9	1,7669	-3,59
Jardim	1872	6,4	1,9211	-3,84
Banheira aquecida	1851	6,3	0,3679	-2,75
Aquecedor	1786	6,1	0,1391	-2,74
Café da manhã	1713	5,9	-0,2495	-2,80
Limpar antes de sair	1401	4,8	1,6840	-3,96
Animais na propriedade	1299	4,5	-0,0028	-3,07
Churrasqueira	1201	4,1	1,9472	-4,41
cartão de segurança	1155	4,0	0,5099	-3,30
Grade na janela	1136	3,9	1,9179	-4,45
Lava-louça	909	3,1	1,9036	-4,69
Estacionamento pago próximo	893	3,1	1,2776	-4,08
Berço	830	2,8	1,2305	-4,12
Equipamento de praia	774	2,7	2,6075	-5,73
Brinquedo e livro para criança	759	2,6	1,7729	-4,75
Vista para o mar	661	2,3	1,8293	-4,96
Casa de um andar	632	2,2	2,0917	-5,32

APÊNDICE B – Grupos de bairros

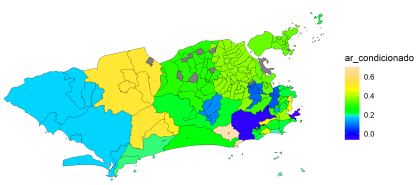
Tabela 9 – Grupos de bairros.

Grupo	Bairros	Número de acomodações
Grupo 1	Recreio dos Bandeirantes, Grumari	1156
Grupo 2	Santa Teresa, Catumbi	1064
Grupo 3	Centro, Paquetá, Gamboa, Saúde	967
Grupo 4	Jacarepaguá, Curicica	873
Grupo 5	Grajaú, Engenho Novo, Engenho de Dentro, Méier, Irajá, Todos os Santos, Brás de Pina, Cachambi, Del Castilho, Penha, Bonsucesso, Encantado, Lins de Vasconcelos, Ramos, Piedade, Maria da Graça, Quintino Bocaiúva, Vila da Penha, Cascadura, Olaria, Penha Circular, Pilares, Vicente de Carvalho, Tomás Coelho, Cordovil, Engenho da Rainha, Inhaúma, Abolição, Higienópolis, Parada de Lucas, Água Santa, Complexo do Alemão, Colégio, Engenheiro Leal, Vaz Lobo, Vigário Geral, Vila Kosmos, Cavalcanti, Jardim América, Parque Columbia, Vista Alegre	497
Grupo 6	Maracanã, São Cristóvão, Jardim Guanabara, Portuguesa, Tauá, Vasco da Gama, Cacuia, Jardim Carioca, Cocotá, Galeão, Moneró, Freguesia (Ilha do Governador), Pitangueiras, Bancários, Manguinhos, Ribeira, Caju, Maré, Praia da Bandeira	390
Grupo 7	Vila Isabel, Andaraí, São Francisco Xavier, Rocha, Mangueira, Riachuelo, Sampaio, Benfica, Jacaré	314
Grupo 8	Vargem Pequena, Vargem Grande, Campo Grande, Bangu, Senador Vasconcelos, Santíssimo, Senador Camará	296

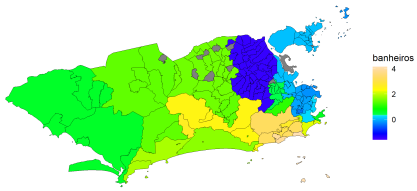
Tabela 9 – Grupos de bairros (Continuação).

Grupo	Bairros	Número de acomodações
Grupo 9	Taquara, Praça Seca, Tanque, Realengo, Vila Valqueire, Bento Ribeiro, Campinho, Marechal Hermes, Parque Anchieta, Madureira, Padre Miguel, Guadalupe, Barros Filho, Jardim Sulacap, Pavuna, Coelho Neto, Anchieta, Oswaldo Cruz, Rocha Miranda, Vila Militar, Deodoro, Honório Gurgel	259
Grupo 10	Rio Comprido, Estácio, Praça da Bandeira, Cidade Nova, Santo Cristo	249
Grupo 11	São Conrado, Alto da Boa Vista	246
Grupo 12	Freguesia (Jacarepaguá), Pechincha, Anil, Gardênia Azul, Cidade de Deus	235
Grupo 13	Itanhangá, Joá	216
Grupo 14	Barra de Guaratiba, Guaratiba, Santa Cruz, Cosmos, Pedra de Guaratiba, Sepetiba, Paciência, Inhoaíba	175
Grupo 15	Vidigal, Rocinha	171
Grupo 16	Copacabana	8806
Grupo 17	Barra da Tijuca	2954
Grupo 18	Ipanema	2878
Grupo 19	Botafogo	1477
Grupo 20	Leblon	1449
Grupo 21	Flamengo	789
Grupo 22	Laranjeiras	585
Grupo 23	Leme	561
Grupo 24	Tijuca	491
Grupo 25	Lagoa	330
Grupo 26	Glória	305
Grupo 27	Catete	282
Grupo 28	Gávea	252
Grupo 29	Jardim Botânico	250
Grupo 30	Humaitá	242
Grupo 31	Camorim	177
Grupo 32	Urca	123
Grupo 33	Cosme Velho	109

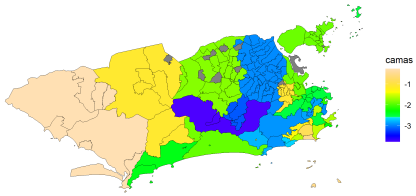
APÊNDICE C – Mapa de coeficientes para a primeira etapa



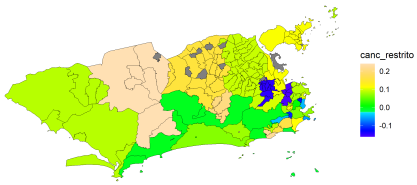
(a) ar_condicionado



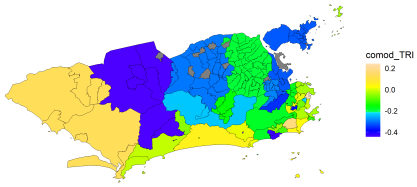
(b) banheiros



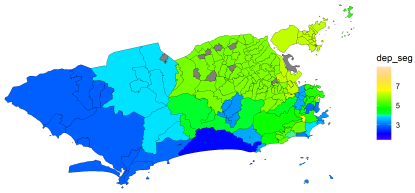
(c) camas



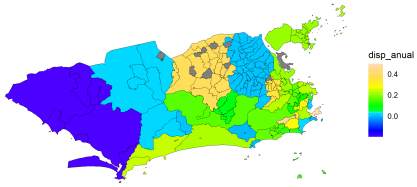
(d) canc_restrito



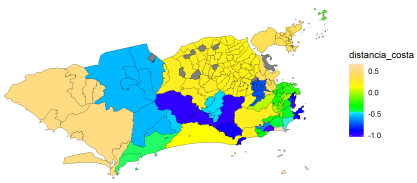
(e) comod_TRI



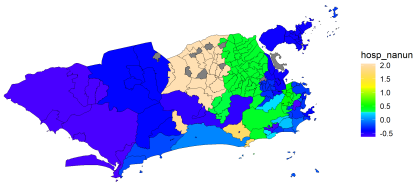
(f) dep_seg



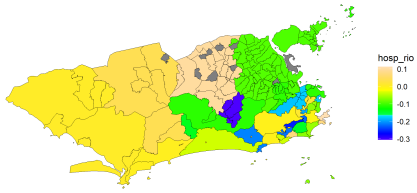
(g) disp_anual



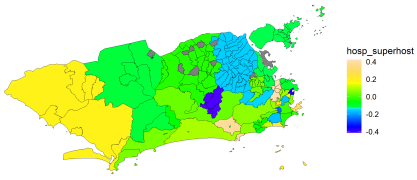
(h) distancia_costa



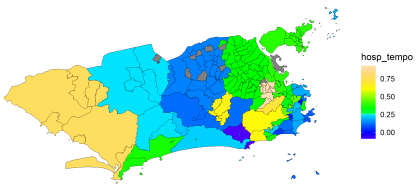
(i) hosp_nanun



(j) hosp_rio

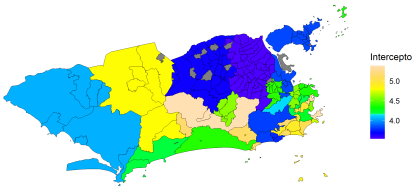


(k) hosp_superhost

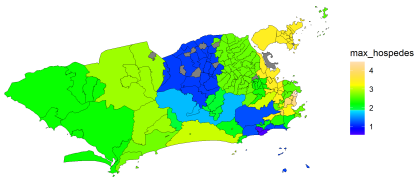


(l) hosp_tempo

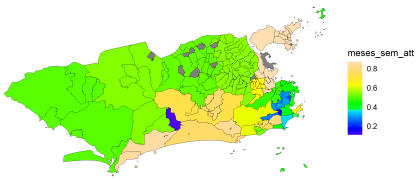
Figura 19 – Mapa de coeficientes β da regressão para a mediana.



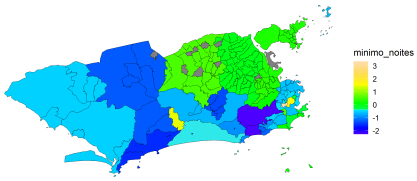
(m) Intercepto



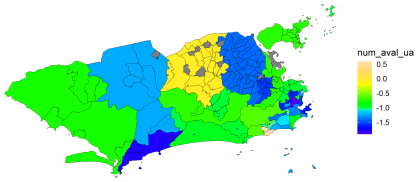
(n) max_hospedes



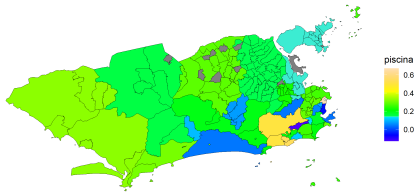
(o) meses_sem_att



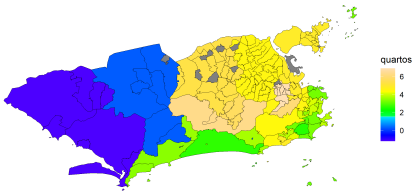
(p) minimo_noites



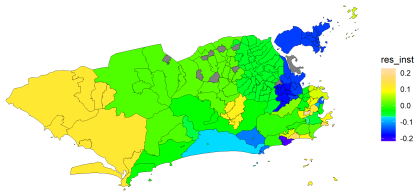
(q) num_aval_ua



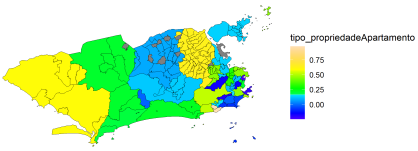
(r) piscina



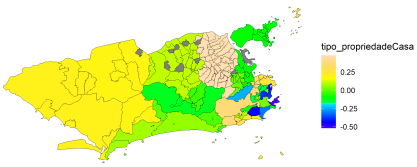
(s) quartos



(t) res_inst



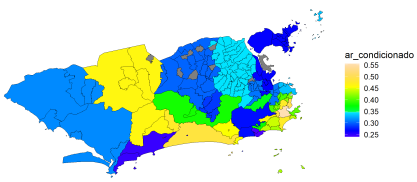
(u) tipo_propriedadeApartamento



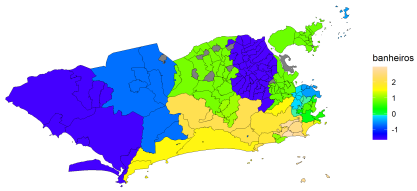
(v) tipo_propriedadeCasa

Figura 19 – (continuação) Mapa de coeficientes β da regressão para a mediana.

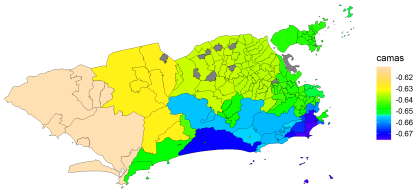
APÊNDICE D – Mapa de coeficientes para a segunda etapa



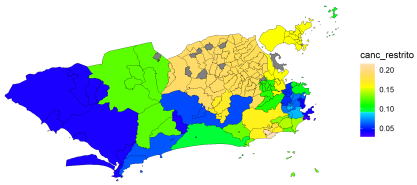
(a) ar_condicionado



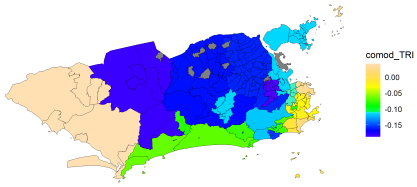
(b) banheiros



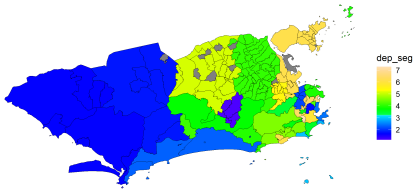
(c) camas



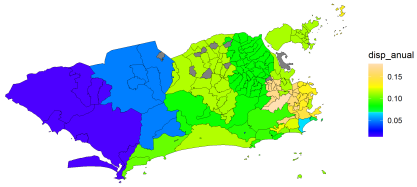
(d) canc_restrito



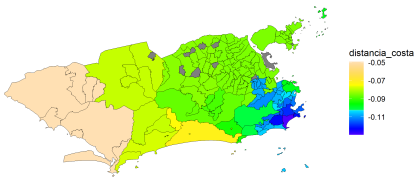
(e) comod_TRI



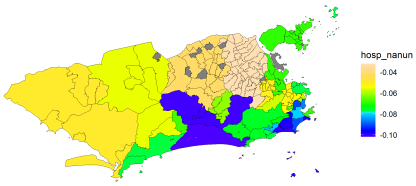
(f) dep_seg



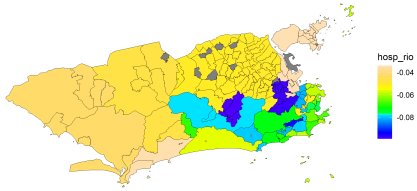
(g) disp_anual



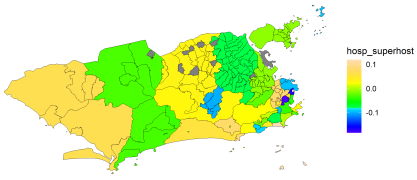
(h) distancia_costa



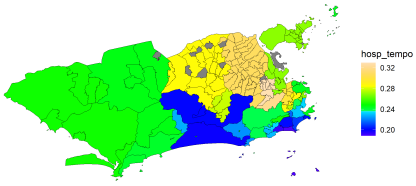
(i) hosp_nanun



(j) hosp_rio

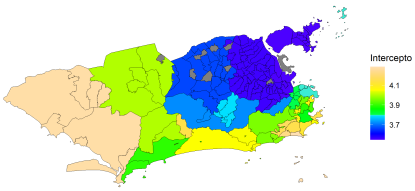


(k) hosp_superhost

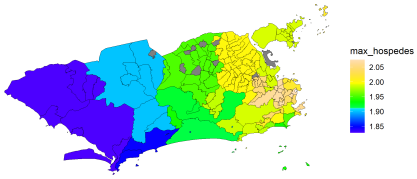


(l) hosp_tempo

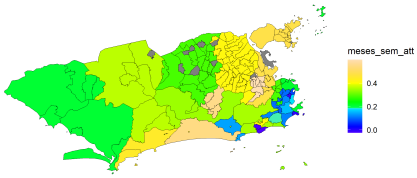
Figura 20 – Mapa de coeficientes β da regressão para o quantil 10% do preço de aluguel.



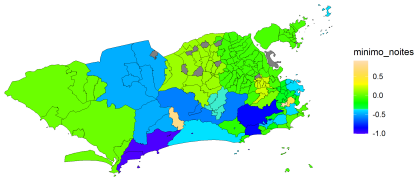
(m) Intercepto



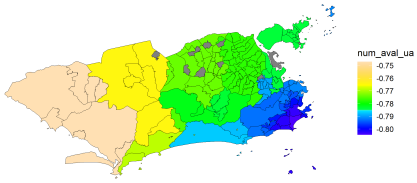
(n) max_hospedes



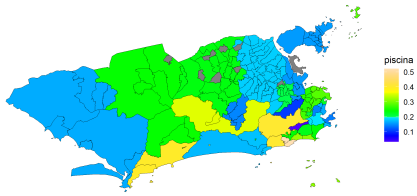
(o) meses_sem_att



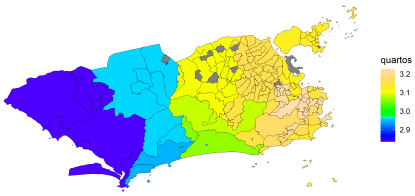
(p) minimo_noites



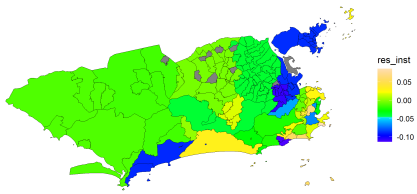
(q) num_aval_ua



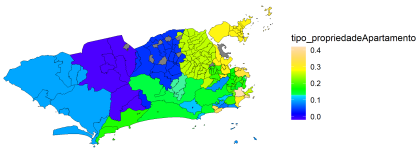
(r) piscina



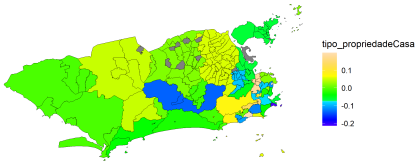
(s) quartos



(t) res_inst

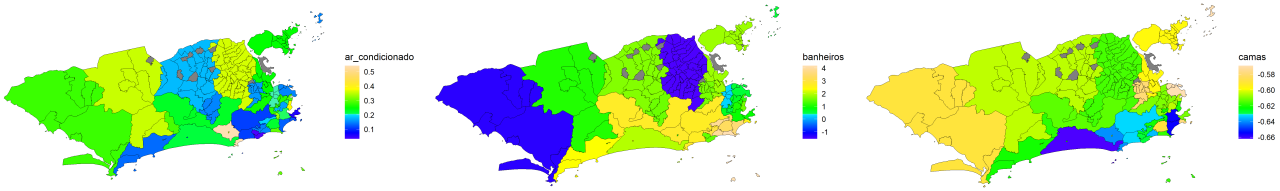


(u) tipo_propriedadeApartamento

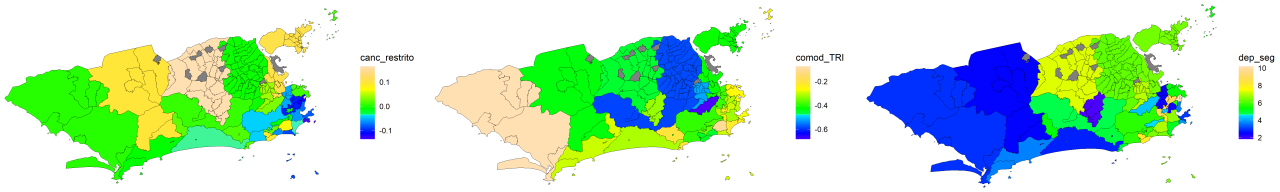


(v) tipo_propriedadeCasa

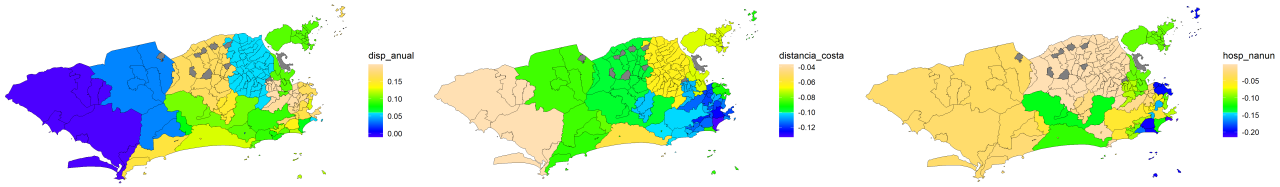
Figura 20 – (continuação) Mapa de coeficientes β da regressão para o quantil 10% do preço de aluguel.



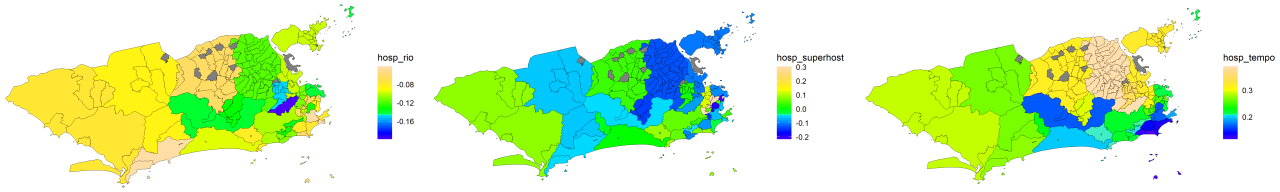
(a) ar_condicionado (b) banheiros (c) camas



(d) canc_restrito (e) comod_TRI (f) dep_seg

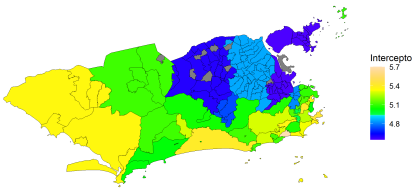


(g) disp_anual (h) distancia_costa (i) hosp_nanun

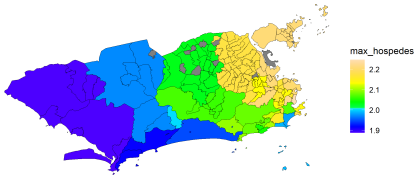


(j) hosp_rio (k) hosp_superhost (l) hosp_tempo

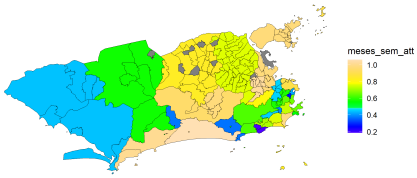
Figura 21 – Mapa de coeficientes β da regressão para o quantil 90% do preço de aluguel.



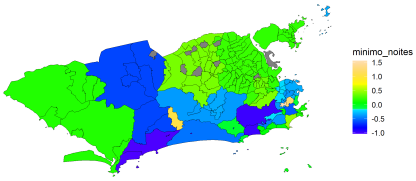
(m) Intercepto



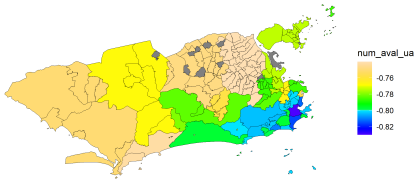
(n) max_hospedes



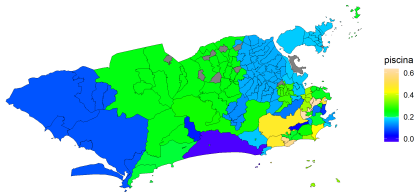
(o) meses_sem_att



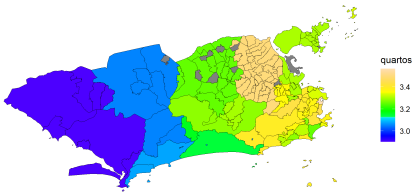
(p) minimo_noites



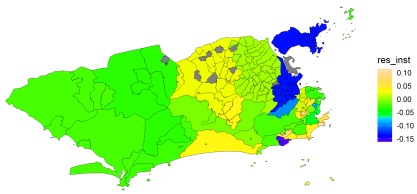
(q) num_aval_ua



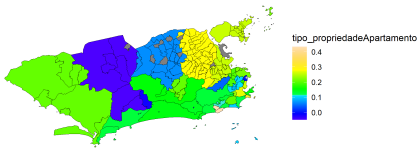
(r) piscina



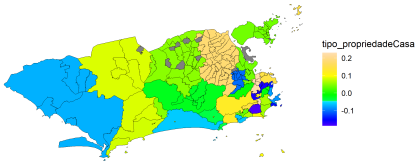
(s) quartos



(t) res_inst



(u) tipo_propriedadeApartamento



(v) tipo_propriedadeCasa

Figura 21 – (continuação) Mapa de coeficientes β da regressão para o quantil 90% do preço de aluguel.