DETERMINANTES DO DESLOCAMENTO PENDULAR NA REGIÃO METROPOLITANA DE LONDRINA – PR: UMA ANÁLISE UTILIZANDO MODELO GRAVITACIONAL E *XGBOOST*

DETERMINANTS OF COMMUTING PATTERNS IN THE METROPOLITAN REGION OF LONDRINA – PR: AN ANALYSIS USING THE GRAVITY MODEL AND XGBOOST

Saulo Murilo de Sousa Nepomuceno [1]
Augusta Pelinski Raiher[2]

RESUMO: Esta pesquisa tem como objetivo analisar os determinantes do deslocamento pendular entre os pares de municípios da Região Metropolitana de Londrina – Pr (RML), em 2010. Para isso, uma amostra de 600 deslocamentos entre os municípios da RML foi utilizada, visando modelar uma equação gravitacional com estimação Poisson Pseudo-Máxima Verossimilhança e um modelo *XGBoost*. Além disso, foram inseridos efeitos espaciais em ambos os modelos e realizada comparação quanto à capacidade preditiva. Os resultados apontam que a distância é um fator de repulsão de trabalhadores pendulares; a variável que representa a dinâmica do mercado de trabalho do município de origem e de destino são fatores de repulsão e de atração, respectivamente. Os efeitos espaciais se apresentaram relevantes, apesar de não elevarem a capacidade preditiva dos modelos. Por fim, o modelo gravitacional, após o refinamento proposto, demonstrou conseguiu alcançar o desempenho comparável ao do *XGBoost*, conforme dadas as métricas de comparação e as características da amostra estudada.

Palavras-chave: Pendularidade; Modelo gravitacional; XGBoost.

ABSTRACT: This research aims to analyze the determinants of commuting patterns between pairs of municipalities in the Metropolitan Region of Londrina – PR (RML) in 2010. A sample of 600 commutes between RML municipalities was used to model a gravity equation with Poisson Pseudo-Maximum Likelihood estimation and an XGBoost model. Spatial effects were incorporated into both models, and their predictive capabilities were compared. The results indicate that distance is a repulsive factor for commuting workers; the variable representing the labor market dynamics of the origin and destination municipalities are repulsive and attractive factors, respectively. Spatial effects were found to be relevant, although they did not enhance the predictive capability of the models. Finally, the gravity model, after the proposed refinement, demonstrated performance comparable to that of XGBoost, according to the comparison metrics and the characteristics of the sample studied.

Keywords: Commuting Patterns; Gravity Model; XGBoost.

Área temática: 14 - População, migração e desenvolvimento

[1] Doutorando pela Universidade Estadual de Maringá. E-mail: saulomurilosn@gmail.com
[2] Professora no Programa de Pós-Graduação em Ciências Sociais, no Programa de Pós-Graduação em economia e no curso de economia da UEPG. Bolsista Produtividade CNPQ. Email: apelinski@gmail.com

1. INTRODUÇÃO

Ainda que o trabalho remoto tenha ganhado mais adeptos durante os últimos anos, o mercado de trabalho ainda é composto, majoritariamente, por pessoas que saem de suas residências para realizar atividades e depois retornam com alguma frequência. Esses deslocamentos são chamados de movimentos pendulares e envolvem aspectos que transcendem a decisão individual sobre onde trabalhar. Durante um tempo considerável a abordagem predominante sobre a mobilidade dos trabalhadores era a análise da migração, porém, o estudo da pendularidade ganhou relevância após o aumento da mobilidade e do alcance das áreas urbanizadas (Tavares; Monteiro, 2019).

Dentro de uma região metropolitana (RM), dinâmicas socioeconômicas vão se formando e se modificando entre as cidades ao longo do tempo. Quando um indivíduo decide que não deve migrar e apenas se deslocar para trabalhar em outro município, isto se constitui uma pendularidade intermunicipal. Pode-se dizer que o conjunto destes indivíduos transitando nessa região representa uma daquelas dinâmicas e, no caso da Região Metropolitana de Londrina (RML) com seus 25 municípios, o objeto de estudo nesta pesquisa.

Em uma determinada RM, entre um dado par de cidades onde há algum movimento de pendulares, uma delas pode receber mais trabalhadores do que envia. Este diferencial na força de trabalho disponível pode afetar os mercados de trabalho de ambos os municípios de maneira diferente bem como o aumento do deslocamento com essa finalidade pode refletir um estrangulamento da infraestrutura de transporte (Moura, 2010). Essas são algumas das implicações da pendularidade, que pode ser benéfica, ao reduzir o diferencial de salários e elevar toda a renda da região (Hazans, 2003), ou desafiadora, ao concentrar atividades econômicas e criar cidades-dormitório com baixo nível de desenvolvimento (Moura, 2010).

Dessa forma, os efeitos da pendularidade devem ser considerados durante a produção de políticas de planejamento urbano e de desenvolvimento regional e, devido à influência da RML que apresenta a maior dinâmica de desenvolvimento econômico e a maior população do interior paranaense, a análise dos deslocamentos pendulares se torna uma tarefa ainda mais importante.

Considerando que um certo número de indivíduos saia de sua cidade para trabalhar em outra e depois retorne, buscou-se analisar a pendularidade entre as cidades na RML a partir da interação de aspectos espaciais e econômicos. A primeira variável representativa do aspecto espacial é a distância entre cada par de municípios, que serve como uma *proxy* para os custos de deslocamento do trabalhador. A segunda diz respeito à contiguidade dos territórios e a terceira relaciona esta contiguidade ao fato de que o município de destino do trabalhador tenha um dado patamar do indicador econômico analisado.

Há várias formas de analisar a pendularidade com modelos estatísticos variados, porém, Stefanouli *et al*. (2017) apontam que o modelo gravitacional tem sido o mais popular para essa finalidade. Além disso, observou-se um crescimento recente do uso de técnicas de *machine learning* (ML) como o *XGBoost* sob o argumento de que possuem maior capacidade preditiva frente ao gravitacional (Robinson; Dilkina, 2018). Neste contexto, esta pesquisa testa ambos os modelos em uma tentativa de, não somente analisar os condicionantes da pendularidade na RML, mas, investigar se o modelo gravitacional, após mudanças em sua especificação, consegue alcançar o desempenho do *XGBoost* e permanecer relevante na análise do fenômeno.

Isto posto, este trabalho analisou os determinantes do deslocamento pendular entre os pares de municípios da Região Metropolitana de Londrina – Pr (RML), em 2010. De forma mais específica, buscou: i) identificar o padrão da pendularidade entre os municípios da RML; ii) analisar como o dinamismo de mercado de trabalho, a distância geográfica e os aspectos espaciais influenciam a decisão de deslocamento dos *commuters*[5]; iii) comparar o desempenho das estimativas com e sem controle de efeitos espaciais dos modelos gravitacional e *XGBoost*.

Cabe ressaltar que uma das principais contribuições deste artigo refere-se à inclusão da variável de interação entre os aspectos econômicos e espacial, a qual se apresenta como inédita na

[5] Indivíduo que se desloca entre sua casa e seu trabalho regularmente. Alguém que realiza a pendularidade.

literatura na análise da pendularidade. O diferencial se encontra na estimação de modelo gravitacional que incorpora a dimensão espacial, além de *XGBoost* que também inclui tal dimensão. Busca-se não somente comparar o desempenho de ambos os modelos, mas, compreender como a inclusão de aspectos espaciais pode aprimorá-los e criar uma interpretação diferente sobre os padrões de deslocamentos pendulares na RML.

O artigo está estruturado em cinco seções, incluindo esta. Na segunda, tem-se o referencial teórico sobre determinantes da pendularidade. Na sequencia é apresentada a metodologia. A quarta seção apresenta os resultados; por fim, tem-se as considerações finais.

2. REFERENCIAL TEÓRICO

A pendularidade tem despertado interesse de autores que analisam diferentes países, utilizando várias metodologias. É interessante notar como muitos estudos sobre pendularidade acabam chegando a resultados parecidos e como como existe similaridade no uso de algumas variáveis, como: distância entre os municípios, custos de transporte e habitação, indicadores de qualidade de vida, disponibilidade de infraestrutura e níveis salariais.

Brito e Ramalho (2019) investigam a pendularidade na RM de Recife em 2010 e como ela é influenciada pelos IDH's[6] municipais e pela distância entre os pares de municípios através de modelagem gravitacional. Seus resultados apontam que a distância é uma força repulsiva de pendulares, o IDH do município de destino, quando alto, é uma força atrativa.

Spadon *et al* (2019) utilizam variáveis como distância, PIB, área, renda, acidentes de trânsito e mais outras características relacionadas à urbanização, para prever a pendularidade dos municípios brasileiros em 2010, utilizando o modelo *XGBoost*. Seus resultados apontam para alta capacidade preditiva, ainda melhor do que o modelo gravitacional, inferindo que distância e PIB são, respectivamente, as características mais importantes na tarefa de previsão dos fluxos pendulares.

Ahrens e Lyons (2020) investigam se o diferencial de preços dos aluguéis entre os centros de emprego e as áreas residenciais são capazes de prever mudanças no tempo médio de deslocamentos pendulares. Para essa tarefa, os autores empregam um modelo gravitacional para os dados da RM de Dublin e a um nível nacional na Irlanda, entre 2011 e 2016. Um resultado de destaque é que um aumento nos aluguéis dentro das áreas de emprego estão associados a uma elevação no tempo de deslocamento pendular.

Chen, Voigt e Fu (2021) usam técnicas de ML, como *Random Forests*, para prever a quantidade de pendulares nos municípios alemães entre 1994 e 2018. Os autores concluíram que o PIB, a distância da área metropolitana, a renda média e a distância trabalho-domicílio são variáveis que influenciam a decisão de realizar o deslocamento pendular. Além disso, o modelo apresentou boa capacidade preditiva e os autores ainda sugeriram sua aplicação para outras regiões.

Liang *et al* (2021) utilizam modelo *XGBoost* para realizar previsão do destino de pendulares usuários do transporte público de Pequim. Segundo os autores, a distância e o tempo de viagem, as localizações da residência e do trabalho e a renda são fatores mais importantes na previsão dos destinos.

De acordo com Zhao *et al* (2023), o modelo gravitacional consegue explicar a pendularidade entre cidades mais próximas, porém, é problemático ao explicar dinâmicas em espaços maiores com muitas grandes cidades interagindo. O modelo, que foi aplicado para a Região Jing-jin-Ji, que inclui a capital chinesa, apontou o PIB municipal como uma força de atração de trabalhadores e a população como uma força de atração de indivíduos que se deslocam por qualquer outro motivo que não seja o trabalho.

Alguns autores, como Morton, Piburn e Nagle (2018), Robinson e Dilkina (2018), Spadon *et al* (2019) e Liang *et al* (2021) comparam modelos gravitacional e *XGBoost* no contexto da pendularidade. Isto é um indicativo de que a comparação entre eles é um tema pertinente. Além disso,

[6] Índice de Desenvolvimento Humano

é possível notar que algumas dessas comparações são realizadas utilizando uma versão tradicional do modelo gravitacional, estimado por mínimos quadrados ordinários (MQO).

Em resumo, pode-se perceber a relevância do estudo da pendularidade, principalmente o destaque do modelo gravitacional e a ascensão de abordagens baseadas em ML. Percebe-se ainda a recorrência do uso de variáveis como distância entre os municípios, nível de renda e características de mercado de trabalho. A análise da pendularidade ainda é predominantemente realizada em um nível intermunicipal, relevando dimensões mais locais, como bairros ou distritos e até maiores, como países ou estados. Todos esses fatores dão subsídios para a escolha atual das variáveis, dos municípios inseridos em uma RM e do uso das ferramentas estatísticas bem como a opção pelo confronto entre técnicas, os quais serão detalhados na sequencia.

3. METODOLOGIA

Esta pesquisa analisou os municípios da Região Metropolitana de Londrina (RML), utilizando os dados do Censo de 2010. Optou-se por esse ano devido à indisponibilidade dos microdados do Censo 2022.

Os microdados do Censo 2010 para os 25 municípios da RML foram filtrados para obter a quantidade de pessoas ocupadas na semana de referência que se deslocaram do município *i* para o município *j* com o objetivo de trabalhar (IBGE, 2010). Entre esses 25 municípios, foram observados 600 fluxos diferentes de trabalhadores entre pares de cidades. Esses dados foram organizados e cada observação foi dividida pela população ocupada do município de origem do movimento pendular. A razão entre o número de trabalhadores que se deslocaram de *i* para *j*, ponderada pela taxa de ocupação de *i*, foi denominada "taxa de deslocamento pendular" do par *i*,*j*.

Dado que o objetivo deste artigo era de analisar a influência do mercado de trabalho e da distância espacial nos movimentos pendulares intermunicipais da RML, as estimativas econométricas tiveram como variável dependente a "taxa de deslocamento pendular". Como variáveis explicativas, teve-se quatro, sendo:

- ✓ IFDM-ER (Índice FIRJAN de Desenvolvimento Municipal-Emprego e Renda), utilizando-o como uma *proxy* para o dinâmica de mercado de trabalho de cada município. A ideia do seu uso é identificar se um município com maiores oportunidade no mercado de trabalho é capaz de atrair trabalhadores de outros lugares, supondo que os trabalhadores dão mais importância ao potencial ganho de renda do que à fatores socioculturais. Além disso, defasou o índice em um ano (2009), visando mitigar possíveis problemas de endogeneidade.
- ✓ Distância geográfica (em quilômetros) entre os pares de municípios, cuja fonte correspondeu ao IBGE[7]. Dentre as pesquisas sobre a pendularidade, a forte influência negativa da distância parece ser um consenso entre os pesquisadores (Abouelhamd, 2021).
- ✓ Variável binária, referindo-se à matriz de contiguidade (do tipo rainha), atribuindo valor igual a 1 para o caso de um par de municípios serem contíguos e valor 0 caso não sejam. O objetivo é medir o efeito de ser vizinha na determinação da pendularidade.
- ✓ Variável de interação espacial entre "ser vizinha" e o IFDM-ER, mensurando o peso de ser vizinha e de ter melhores (ou piores) condições de desenvolvimento do mercado de trabalho neste processo. Um argumento favorável à introdução de efeitos espaciais na análise da pendularidade é de que não basta ser um município próximo, mas ser vizinho ou suficientemente próximo de um município com maior dinâmica de mercado de trabalho (Moura, 2010).

Na sequencia, são apresentados os dois modelos estimados na determinação da pendularidade: modelo Gravitacional e modelo *XGBoost*.

3.1 Modelo gravitacional

[7] As coordenadas geográficas corresponderam às áreas mais urbanizadas de cada município.

A equação gravitacional (1) é definida como uma força de atração entre duas regiões onde a atividade humana é produzida pela interação das suas massas populacionais enquanto uma força repulsiva é gerada pela magnitude do espaço entre elas (Carrothers, 1956; Simini *et al*, 2012):

$$T_{i,j} = \frac{m_i^{\alpha} n_j^{\beta}}{f(r_{ij})} \qquad (1)$$

Em que: $T_{i,j}$ representa o número de indivíduos que se deslocam entre os municípios *i* e *j*, m_i é a população ocupada no município de origem *i*, n_j é a população ocupada do município de destino *j* e r_{ij} é a distância entre os municípios *i* e *j*. α, β são parâmetros ajustáveis e $f(r_{ij})$ é a função de dissuasão que ajusta os dados.

Head e Mayer (2014) alertam para problemas nos modelos gravitacionais: o primeiro ponto a se considerar é o termo do erro que deve receber atenção especial para evitar heterocedasticidade e o segundo ponto é a presença de observações com valor nulo na variável dependente.

Sobre o termo do erro, na maioria das pesquisas se estima a equação gravitacional por MQO (Mínimo Quadrado Ordinário) e, de acordo com Santos Silva e Tenrenyo (2006), os erros provenientes da modelagem, mesmo que controlados por efeitos fixos, serão heterocedásticos, gerando forte viés.

Wölver, Burgard e Breßlein (2018) apontam para a vantagem de modelos gravitacionais na forma multiplicativa com modelos lineares generalizados, como PPML e GPML[8] como o único grupo capaz de lidar com valores zero na variável dependente. Nesta pesquisa, o modelo gravitacional é estimado por PPML, seguindo Santos Silva e Tenreyro (2006), além da utilização de matriz de covariância robusta.

Devido à forma não linear da equação (1), muitos autores a linearizam através da aplicação de logaritmo e a estimam por MQO. Um grave problema resultante desse procedimento é que, como há muitos fluxos "zero" na base de dados desta pesquisa, o logaritmo de zero não é definido matematicamente. Dessa forma, o modelo PPML é capaz de lidar com essa característica da base de dados sem precisar recorrer à log-linearização. Um outro problema evitado é que, com dados de contagem, a equidispersão se torna uma preocupação. Como a variável dependente deixou de ser um dado de contagem após se tornar uma taxa de pendularidade, a equidispersão deixou de ser um problema, porém, mesmo assim, as estimativas de um PPML se tornam consistentes independentemente da distribuição dos dados, se são ou não Poisson, ou até mesmo se são ou não dados de contagem (Arvis, Shepherd, 2013). Então, a taxa de deslocamento pendular estimada por PPML apresenta a seguinte forma:

$$Tx.Pendulares_{ij} = \exp\left(\beta_0 + \beta_1 IFDM_i + \beta_2 IFDM_j + \beta_3 Distância_{ij} + \gamma Z_{ij} + \varepsilon_{ij}\right) \qquad (2)$$

Onde: $Tx.Pendulares_{ij}$ é a taxa de deslocamento pendular entre o par de municípios i e j; $IFDM_i$ e $IFDM_j$ são os índices IFDM-ER; $Distância_{ij}$ é a distância entre os pares de municípios, Z_{ij} é o termo que engloba as variáveis espaciais; ε_{ij} é o termo do erro e $\beta_0,\ \beta_1, \beta_2,\ \beta_3, \gamma$ são os parâmetros estimados.

3.2 Modelo *XGBoost*

XGBoost consiste em uma biblioteca com um conjunto de técnicas de modelagem baseados na pesquisa de Friedman (2001) e apresentado na forma atual por Cheng e Guestrin (2016). Trata-se de um conjunto de árvores de regressão e classificação ou CART[9] (Chen; Guestrin, 2016). Define-se

[8] Gamma Pseudo Máxima Verossimilhança
[9] Classification and regression trees

uma função objetivo, que pode assumir qualquer forma, representando esse conjunto de K árvores, dada uma amostra com n exemplos e m características para prever um resultado $\hat{Y}_i$.

$$\hat{Y}_i = \phi(x_i) = \sum_{k=1}^{K} f_k(x_i)\,,\ f_k \in \mathcal{F}, \tag{3}$$

$$\text{onde } \mathcal{F} = \left\{f(x) = w_{q(x)}\right\}\,,\ (q\colon \mathbb{R}^m \to \mathrm{T}, w \in \mathbb{R}^{\mathrm{T}}) \tag{3.1}$$

Onde: T é o número de folhas na árvore, q(x) é uma função que representa a correspondência entre cada observação e o valor da sua folha, w é o vetor dos valores das folhas, f_k é a k-ésima árvore com estrutura q e valor das folhas w e $\mathcal{F}$ é o conjunto de todos os CART's.

Segundo Chen e Guestrin (2016), minimiza-se a função objetivo regularizada $\mathcal{L}(\phi)$ para o aprendizado do modelo de conjunto de árvores:

$$\mathcal{L}(\phi) = \sum_i l(y_i, \hat{y}_i) \sum_k \Omega(f_k)\,, \tag{4}$$

$$\Omega(f_k) = \gamma \mathrm{T} + \frac{1}{2}\lambda ||w||^2 \tag{5}$$

Em que: $l()$ é uma função de perda que mede a diferença entre a previsão $\hat{y}_i$ e o alvo y_i, Ω é uma função de regularização que penaliza a complexidade do modelo e ajuda a evitar sobreajuste (Chen; Guestrin, 2016), γ representa a complexidade de cada folha multiplicado pelo total de folhas T, λ é um parâmetro que mede a penalização sobre o valor das folhas e w representa o vetor com o peso das folhas (Ma *et al*, 2021).

O modelo de *tree boosting* em (6) tem parâmetros nas árvores a serem encontrados. Adiciona-se uma nova árvore f_t na t-ésima iteração enquanto o modelo treina, em outras palavras, o treinamento ocorre de maneira aditiva (Osman *et al*, 2021).

$$\mathcal{L}^{(t)} = \sum_n l(y_i, \hat{y}_i^{(t)}) \sum_k \Omega(f_k) \tag{6}$$

Portanto, o modelo adiciona uma nova árvore a cada novo estágio para melhorar a previsão final. A previsão realizada, de maneira aditiva, da iteração t em $\hat{y}_i^{(t)}$ será:

$$\begin{aligned}
&\hat{y}_i^{(0)} = 0\\
&\hat{y}_i^{(1)} = f_1(x_i) = \hat{y}_i^{(0)} + f_1(x_i)\\
&\hat{y}_i^{(2)} = f_1(x_i) + f_2(x_i) = \hat{y}_i^{(1)} + f_2(x_i)\\
&\hat{y}_i^{(3)} = f_1(x_i) + f_2(x_i) + f_3(x_i) = \hat{y}_i^{(2)} + f_3(x_i)\\
&\hat{y}_i^{(t)} = \sum_{k=1}^{t} f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i)
\end{aligned} \tag{7}$$

Então, seja $\hat{y}_i^{(t)}$ a previsão do i-ésimo instante na t-ésima iteração, soma-se f_t para minimizar a função objetivo:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_k) \tag{8}$$

Isso implica que o algoritmo adiciona a melhor árvore f_t possível que, de forma míope[10], consiga melhorar o modelo de acordo com a equação de *tree boosting* (6). Então, a otimização ocorre por expansão de Taylor até a segunda ordem. Na t-ésima iteração, a função objetivo será:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l[g_i f_t(x_i) + \frac{1}{2} h_i \partial_t^2(x_i)] + \Omega(f_k) \tag{9}$$

[10] Uma escolha ótima local em cada estágio, ou seja, sem considerar os estágios futuros.

com $g_i = \partial_{\hat{y}_i^{(t-1)}} l\left(y_i, \hat{y}_i^{(t-1)}\right)$ e $h_i = \partial^2_{\hat{y}_i^{(t-1)}} l\left(y_i, \hat{y}_i^{(t-1)}\right)$ gradientes de 1º e 2º ordem da função de perda.

Substituindo (5) em (9) e, sendo $I_j = \{i|q(x_i) = j\}$ o conjunto dos dados observados relacionados à j-ésima folha, a nova função objetivo do *XGBoost* é:

$$\hat{\mathcal{L}}^{(t)} = \sum_{j=1}^{T}[(\sum_{i\in I_j} g_i)\, w_j + \frac{1}{2}(\sum_{i\in I_j} h_i + \lambda)\, w_j^2] + \gamma T \qquad (10)$$

Da equação (10), para uma determinada estrutura $q(x)$, calcula-se o valor ótimo w_j^* da folha j:

$$w_j^* = -\frac{\sum_{i\in I_j} g_i}{\sum_{i\in I_j} h_i + \lambda} \qquad (11)$$

e calcula-se o valor ótimo da função que define a qualidade da estrutura da árvore q:

$$\hat{\mathcal{L}}^{(t)}(q) = -\frac{1}{2}\sum_{j=1}^{T}\frac{(\sum_{i\in I_j} g_i)^2}{\sum_{i\in I_j} h_i + \lambda} + \gamma T) \qquad (12)$$

Segundo Chen e Guestrin (2016), para encontrar $\hat{\mathcal{L}}^{(t)}(q)$ é necessário apenas as estatísticas dos gradientes de primeira e de segunda ordem dados por cada folha e então aplicar a fórmula da equação (12) para obtê-lo e, quão menor for o valor, melhor é a estrutura da árvore para a modelagem.

Sabendo que a modelagem consegue medir a qualidade das árvores, teoricamente seria possível elencar, dentre todas as árvores possíveis, as que trariam mais benefícios em termos de previsão e escolhê-las para integrar o modelo, porém, normalmente isto é impossível. Então, no *XGBoost*, um algoritmo míope começa a otimização a partir de uma folha e vai adicionando ramificações à árvore de maneira iterativa (Ma *et al*, 2021).

Dividindo uma folha qualquer em duas e supondo I_L e I_R os conjuntos das instâncias da esquerda e da direita, respectivamente, ou seja, das amostras de dados que foram divididas, após uma divisão dentro de uma árvore e, sabendo que $I = I_L \cup I_R$, a função de perda utilizada para avaliar o ganho com essa divisão é:

$$\mathcal{L}_{split} = \frac{1}{2}\left[\frac{(\sum_{i\in I_L} g_i)^2}{\sum_{i\in I_L} h_i + \lambda} + \frac{(\sum_{i\in I_R} g_i)^2}{\sum_{i\in I_R} h_i + \lambda} - \frac{(\sum_{i\in I} g_i)^2}{\sum_{i\in I} h_i + \lambda}\right] - \gamma \qquad (13)$$

Onde: a primeira razão dentro dos colchetes representa o valor da nova folha da esquerda, a segunda razão representa o valor da nova folha da direita, a última representa o valor da folha original antes da divisão e o parâmetro γ representa a regularização sobre a folha adicional. Assim, se o ganho com a divisão for menor do que γ, o modelo não vai adicionar tal divisão na árvore e este procedimento é chamado de "poda" nos modelos de árvores em ML (XGBoost, 2023), além disso, Chen e Guestrin (2016) afirmam que um dos maiores problemas enfrentados pelos usuários de aprendizado de árvores é encontrar a melhor divisão possível.

Conforme descrito acima, *Gradient boosting* é uma técnica poderosa de ML para problemas de regressão e classificação que produz um modelo preditivo na forma de um conjunto de modelos de previsão mais fracos, tipicamente árvores de decisão, e funciona construindo o modelo de forma sequencial: cada novo modelo tenta corrigir os erros do modelo anterior. No coração do algoritmo está a ideia de otimizar uma função de perda diferençável (Ramaneswaran *et al*, 2021). Em cada estágio, o algoritmo se ajusta ao gradiente negativo da função de perda associada ao modelo de aprendizado. Isso é feito minimizando a soma dos erros quadráticos dos pseudo-resíduos (os gradientes negativos), movendo o modelo na direção de maior melhoria. Este processo é repetido até

que uma quantidade predeterminada de iterações seja alcançada ou até que uma melhoria adicional seja mínima, resultando em um modelo final que é uma combinação ponderada dos modelos mais fracos (Friedman, 2001).

Neste estudo, adota-se a modelagem *XGBoost*, com treinamento do modelo para previsão dos fluxos de pendulares entre os pares de municípios. Vale destacar que a distinção da modelagem aplicada nesta pesquisa em relação às anteriores reside na incorporação da *dummy* de interação, considerando a vizinhança e o desempenho do mercado de trabalho da vizinhança como variáveis-chave da determinação da pendularidade.

É essencial, após a definição do modelo e a sua aplicação, conseguir extrair alguma informação sobre as variáveis estudadas, interpretando seus resultados. Para isso, faz-se necessário o uso de ferramental específico para que se conclua algo sobre os efeitos das variáveis explicativas na variável de interesse, quaisquer que sejam aqueles. Os *SHAP values* cumprem este papel nesta pesquisa para o *XGBoost*.

De acordo com Li (2022), valores *Shapley* são interessantes para analisar a relação entre as variáveis e para auxiliar na previsão, quando se trata de modelos de aprendizado de máquina, sobretudo com modelos *XGBoost*.

Lundberg e Lee (2017) criaram a estrutura SHAP a partir da teoria dos valores *Shapley* para amenizar um problema dos modelos de ML, que apesar de serem acurados, são difíceis de serem interpretados, pois, no caso do *XGBoost*, as estimativas são previsões elencadas por ordem de importância sobre a variável dependente. Diferentemente da interpretação usual das características de um modelo de ML, a ferramenta permite identificar (globalmente) se cada variável possui uma contribuição positiva ou negativa na variável de interesse e identificar (localmente) tal contribuição para cada observação de cada característica (Mangalathu *et al*, 2020). Dessa forma é mais fácil entender o motivo de uma determinada previsão do modelo e identificar quais as características mais importantes na tomada de decisão do modelo.

Wang *et al.* (2021) acrescentam que se o valor SHAP é positivo, então a característica em questão deve ser entendida como uma força que impulsiona positivamente o valor da variável alvo enquanto um valor negativo gera um efeito inverso nesta, além de que, ao observar-se o valor em módulo do valor SHAP, tem-se uma noção da intensidade do impacto da característica sobre a saída do modelo, em outras palavras, quanto maior o valor em módulo, mais intensa é essa força.

3.3 Comparação entre os Modelos Gravitacional e *XGBoost* e Estratégia Empírica

Robinson e Dilkina (2018), assim como Morton, Piburn e Nagle (2018), apresentam algumas métricas comumente utilizadas para a avaliação comparativa de modelos de aprendizado de máquina, como o *XGBoost*, em comparação com modelos mais tradicionais, como o modelo gravitacional. Dessa maneira, a performance de ambos os modelos será testada utilizando as seguintes métricas propostas nas duas pesquisas mencionadas:

- ✓ *Common Part of Commuters* (CPC): Compara a matriz estimada do número de pendulares T com a matriz da quantidade observada $\tilde{T}$ (Robinson; Dilkina, 2018). A métrica apresenta valores entre 0 e 1, sendo desejáveis valores o mais próximo de 1 possível, com este valor indicando que as matrizes estimada e observada são idênticas. Ela mede a proporção do total de pendulares que são corretamente previstos pelo modelo. O CPC é, portanto, uma métrica de volume de fluxo, mostrando o quão precisamente o modelo captura a magnitude dos fluxos de pendulares. A equação é descrita como:

$$CPC\left(T,\tilde{T}\right) = \frac{2\sum_{i,j=1}^{n}\min\left(T_{ij},\tilde{T}_{ij}\right)}{\sum_{i,j=1}^{n}T_{ij}+\sum_{i,j=1}^{n}\tilde{T}_{ij}} \qquad (14)$$

- ✓ *Common Part of Links* (CPL): Definição idêntica à da CPC, mede a similaridade dos fluxos observados T com os fluxos estimados $\tilde{T}$ (Morton; Piburn; Nagle, 2018). Varia entre 0 e 1 e valores mais próximos de 1 são desejáveis. A CPL se concentra na estrutura da rede em si,

avaliando a proporção de conexões entre os pares que são corretamente identificados pelo modelo, independentemente do número de pendulares que passam por esses *links*. Assim, CPL indica quão bem o modelo captura a estrutura da rede de fluxos em termos de conexões existentes. A equação é descrita como:

$$CPL\left(T,\tilde{T}\right)=\frac{2\sum_{i,j=1}^{n}\left(\mathbb{I}_{T_{ij}>0}\cdot\mathbb{I}_{\tilde{T}_{ij}>0}\right)}{\sum_{i,j=1}^{n}\mathbb{I}_{T_{ij}>0}+\sum_{i,j=1}^{n}\mathbb{I}_{\tilde{T}_{ij}>0}} \qquad (3.4.2)$$

- ✓ *Root Mean Squared Error* (RMSE): Mede a acurácia da previsão, valores próximos de zero são desejáveis, indicando que estão mais próximos dos valores reais (Morton; Piburn; Nagle, 2018). RMSE é uma medida robusta pois dá mais peso a erros maiores por estar elevada ao quadrado, tornando-a sensível a outliers.

$$RMSE\left(T,\tilde{T}\right)=\sqrt{\frac{1}{n}\sum_{i,j=1}^{n}\left(T_{ij}-\tilde{T}_{ij}\right)^{2}} \qquad (3.4.3)$$

Onde T_{ij} é o fluxo real de pendulares entre os municípios i e j, $\tilde{T}_{ij}$ é o fluxo estimado e n é o número total de pares de municípios.

Morton, Piburn e Nagle (2018) destacam que o modelo *XGBoost* apresenta um desempenho significativamente superior, conforme evidenciado por algumas das métricas mencionadas anteriormente, em comparação com os modelos de radiação e gravitacional. No entanto, em relação ao RMSE, observaram que o *XGBoost* pode gerar estimativas discrepantes em relação aos valores observados. Diante dessa constatação, recomendam a comparação desta ferramenta com modelos de pendularidade mais complexos. Este é exatamente o enfoque adotado nesta pesquisa ao incorporar efeitos espaciais no modelo gravitacional.

No Quadro 1 tem-se um resumo das variáveis explicativas selecionadas e suas nomenclaturas (em parênteses). Ressalta-se que todas as estimativas foram efetuadas no software R.

Quadro 1: Descrição das variáveis utilizadas nas estimações

Variável dependente	Variáveis explicativas				
Taxa de deslocamento pendular entre os municípios i e j (T_{ij})	Índice Firjan – Emprego e Renda do município de origem do fluxo ($if0\_renda$)	Índice Firjan – Emprego e Renda do município de destino do fluxo ($if1\_renda$)	Distância entre o município i e o município j (D_{ij})	*Dummy* de contiguidade ($contig$)	Variável de interação espacial ($varinter$)

Fonte: Elaborado pela pesquisa.

Produziu-se um total de seis modelos, a saber: três estimativas com modelo gravitacional e três com modelo *XGBoost*, alterando apenas a dimensão do efeito espacial entre eles. O Quadro 2 resume as modelagens.

Foram realizadas estimações MQO, GPML e PPML na equação gravitacional para se certificar de que os modelos PPML são bem especificados, os quais validaram tal hipótese. Os modelos gravitacionais também foram estimados com matriz de covariância robusta para evitar problemas de heterocedasticidade.

Quanto ao modelo XGBoost, inicialmente ocorre a seleção das variáveis independentes (Quadro 2) consideradas preditoras para o modelo. Após essa etapa, os dados selecionados são transformados em uma matriz, num formato adequado para a utilização no modelo de aprendizado de máquina. Em seguida, a variável dependente, que o modelo visa prever, é definida a partir do conjunto de dados. O próximo passo envolve a configuração dos parâmetros do XGBoost, os quais incluem configurações que influenciam o tipo de modelo, a taxa de aprendizagem, os critérios de

minimização de perda, a profundidade das árvores de decisão, entre outros aspectos relevantes para o treinamento do modelo.

O modelo é então treinado utilizando a matriz de preditores e a variável alvo. Durante o treinamento, são ajustados aspectos como o número de iterações, a frequência de atualizações durante o treinamento e critérios para a parada antecipada, visando otimizar o desempenho do modelo. Por fim, é realizada uma análise dos valores SHAP, que são utilizados para entender a importância e a contribuição de cada variável explicativa nas previsões do modelo.

Quadro 2: Descrição das estimativas econométricas

Modelos	Variável dependente	Variáveis explicativas
	Modelo gravitacional	
G1	T_{ij}	$if0\_renda + if1\_renda + D_{ij}$
G2	T_{ij}	$if0\_renda + if1\_renda + D_{ij} + contig$
G3	T_{ij}	$if0\_renda + if1\_renda + D_{ij} + varinter$
	Modelo XGBoost	
X1	T_{ij}	$if0\_renda + if1\_renda + D_{ij}$
X2	T_{ij}	$if0\_renda + if1\_renda + D_{ij} + contig$
X3	T_{ij}	$if0\_renda + if1\_renda + D_{ij} + varinter$

Fonte: Elaborado pela pesquisa.

4. RESULTADOS E DISCUSSÃO

A RML foi criada em 1998 por meio da Lei Estadual Complementar nº. 81, instituída pelo Governo do Estado do Paraná e, originalmente, compunha apenas seis municípios: Londrina, Jataizinho, Cambé, Rolândia e Tamarana (SEDU, 2022). Após duas décadas de desenvolvimento regional, a RML passou a ter 25 municípios incluindo os supracitados, com população combinada de 1.000.062 habitantes e, destes, 378.216 pessoas declararam que trabalhavam fora dos seus domicílios e retornavam diariamente (IBGE, 2010).

Os municípios da RML, em sua maioria, eram de pequeno porte, apresentando baixa quantidade de pessoas ocupadas (Figura 1a). Eram 20 municípios, dos 25, com menos de 20 mil habitantes. Londrina se apresenta como um *outlier* nas observações (município-polo), ressaltando que os municípios que possuíam mais de 20 mil habitantes (quais sejam: Arapongas, Cambé, Ibiporã e Rolândia) faziam necessariamente fronteira com o município-polo.

Grande parte da população ocupada da RML se concentrava ao redor do município-polo, sobretudo, na região de conurbação[12] entre Londrina e as cidades de Cambé e Ibiporã (Figura 1a e b). Porecatu apresentava uma quantidade maior de população ocupada em relação aos outros municípios do noroeste da região. Além disso, outros dois municípios com maior quantidade de população ocupada da região, Rolândia e Arapongas, se encontravam em processo de conurbação, além de Jataizinho (SEDU,2022). Este último processo pode ser observado tanto na figura 1a quanto na figura 1b e evidencia o importante papel da Rodovia Federal BR-369 como um eixo de ocupação que perpassa esta região.

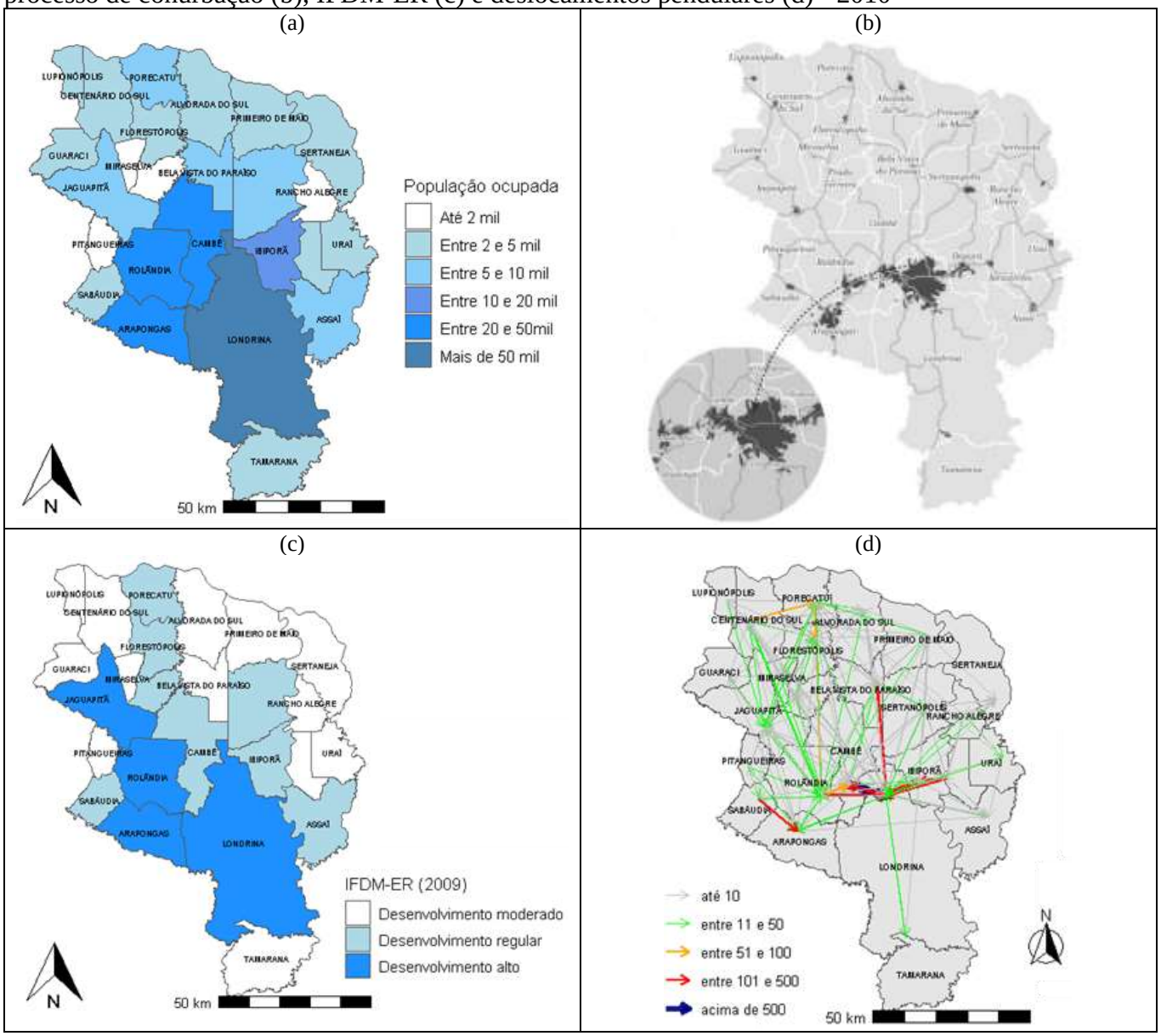
Sobre o índice Firjan – Emprego e Renda, não havia municípios com baixo nível de desenvolvimento em termos de dinamismo econômico (com enfoque no mercado de trabalho), ou seja, todos apresentavam um nível, pelo menos, moderado, como se pode observar na Figura 1c.

[12] Quando cidades limítrofes crescem até o ponto em que suas regiões urbanas se encontram e se tornam homogêneas. A existência de uma conurbação não é condição necessária para a formação de uma RM.

Agora é possível perceber que algumas das cidades com menos de 20 mil habitantes e população ocupada menor que 5 mil trabalhadores, principalmente em direção ao noroeste da região, tinham IFDM-ER melhores em relação aos seus pares de mesma característica populacional.

Ao comparar a entrada e a saída de trabalhadores pendulares, apenas cinco municípios eram receptores líquidos de trabalhadores (entrada maior que saída): Arapongas, Jaguapitã, Londrina, Porecatu e Rolândia. Isso indica a força de atração exercida pelo município-polo e pelos municípios que apresentavam dinâmica própria dentro da RML.

Figura 1: Disposição espacial dos municípios da RML por quantidade de trabalhadores ocupados (a), processo de conurbação (b), IFDM-ER (c) e deslocamentos pendulares (d) - 2010



Fonte: (a) e (d) IBGE (2010), (b) SEDU(2022), (c) Firjan, com dados organizados pela pesquisa

A magnitude dos deslocamentos dos trabalhadores (Figura 1d) tende a convergir com o tamanho do mercado de trabalho dos municípios receptores (Figura 1a). É possível observar a grande concentração de deslocamentos que ocorrem na região de conurbação e no município-polo, recebendo trabalhadores pendulares de praticamente todos os demais municípios da RML, deixando mais explicito a força de atração de *commuters* nesta região. Os municípios de Rolândia, Bela Vista do Paraíso, Jataizinho, Ibiporã e Cambé eram os maiores responsáveis pelos 2656 trabalhadores que se deslocavam até Londrina, com destaque para Cambé, o maior doador de *commuters* da amostra, cedendo 1497 de seus trabalhadores somente para o município-polo. Já Londrina, era a maior

receptora de *commuters*, e era a quarta doadora, enviando quase 70% de seus trabalhadores pendulares às duas cidades conurbadas, Cambé e Ibiporã.

É interessante observar a existência de outras dinâmicas de pendularidade mais afastadas e que não envolviam o município-polo. Exemplo disso é a região noroeste da RML, a qual apresentava dinâmica própria. Outra dinâmica que merece destaque refere-se a de Rolândia, a qual era a segunda maior receptora de pendulares da RML, com fluxos intensos vindo de municípios como Cambé, Florestópolis e Centenário do Sul. Esses movimentos pendulares que não envolviam Londrina possivelmente decorriam de algum tipo de demanda por mão de obra específica, por exemplo, em decorrência da presença de algum tipo de indústria ou de negócio.

Dessa forma, percebe-se que municípios mais afastados do município-polo e com características mais modestas, em termos de população ocupada e de mercado de trabalho, apresentavam poucos fluxos, especialmente de entrada de trabalhadores, enquanto municípios próximos com maior mercado de trabalho e população ocupada, tenderam a gerar maiores deslocamentos. Isso sugere que a menor distância dos municípios com maiores dinâmicas do mercado de trabalho desempenhava um papel de atração para o deslocamento dos trabalhadores (e vice-versa).

Ademais, é possível observar que municípios não somente próximos, mas, fronteiriços, tenderam a apresentar os maiores movimentos de trabalhadores entre si. Então, além da distância geográfica, o efeito espacial pode decorrer da dinâmica do mercado de trabalho do envoltório.

A partir destas observações iniciais, testou econometricamente a importância destas variáveis no processo pendular dos municípios da RML. Na Tabela 1, tem-se os resultados dos modelos gravitacionais.

Tabela 1: Coeficientes do modelo gravitacional - PPML – modelos (G1), (G2) e (G3) do Quadro 2

Variáveis	Modelo (G1) Coeficientes	Modelo (G2) Coeficientes	Modelo (G3) Coeficientes
Intercepto	1,53*	-0,77	-0,28
Distância entre pares	-2,03*	-1,46*	-1.61*
IFDM-ER cidade de origem	-2,27*	-2,15*	-2,22*
IFDM-ER cidade de destino	6,65*	6,59*	6,23*
Variável de interação	-	-	0,84*
Dummy de contiguidade	-	0,91*	-

Fonte: Elaborado pela pesquisa.
(*) estatisticamente significativo a 1%

Os três modelos gravitacionais apresentam estimativas significativas e convergentes em suas conclusões. Todos indicam que a distância é um fator de repulsão para os fluxos pendulares, ou seja, municípios mais próximos uns dos outros tendem a atrair trabalhadores vizinhos e municípios distantes provocam efeito inverso em seus trabalhadores. Além disso, as condições de mercado de trabalho (IFDM-ER) atuam como força repulsiva para trabalhadores nos municípios de origem (sinal negativo), enquanto, para o município de destino, exercem uma força de atração (sinal positivo), sugerindo que municípios com baixa dinâmica de mercado de trabalho tendem a fornecer trabalhadores para aqueles com alta dinâmica. Em outras palavras, uma cidade com baixo IFDM-ER tende a ser doadora de *commuters*, enquanto outra que apresente o indicador mais elevado tende a ser receptora. Essa análise vale para os três modelos gravitacionais.

Já no modelo (G2) há uma interpretação adicional, uma vez que existe uma nova variável. O coeficiente da *dummy* de contiguidade é positivo e significativo, indicando que, na média, existe um efeito espacial de vizinhança. Ou seja, ser um município fronteiriço a outro eleva a força de atração de trabalhadores pendulares entre esses municípios. Esse resultado condiz não somente com os fluxos observados entre os municípios conurbados, mas com os municípios vizinhos no noroeste e no oeste da RML, como Porecatu-Florestópolis e Sabáudia-Arapongas.

Sobre o último modelo gravitacional, a variável de interação espacial entre “apresentar contiguidade e mercado de trabalho” também foi positiva e significativa. Nesse contexto, o efeito espacial que influencia a atração entre dois municípios não se limita apenas à proximidade

geográfica, mas também está associado, na média, à qualidade do mercado de trabalho do envoltório. De fato, ser um município vizinho com características laborais mais favoráveis aumenta, na média, a atração de trabalhadores pendulares para essa localidade. Consequentemente, essa constatação reflete uma conclusão semelhante à variável IFDM-ER do município de destino, com a interpretação adicional de que, na média, a proximidade geográfica e a dinâmica econômica combinadas exercem uma força atrativa sobre os trabalhadores pendulares quando comparados aos demais municípios não vizinhos.

De maneira geral, os resultados das três estimativas gravitacionais (Tabela 1) estão em consonância com os achados de outros autores que utilizaram equação gravitacional na análise dos movimentos pendulares, como Renkow e Hoover (2000), Brito e Ramalho (2019) e Zhao et al. (2023). Além disso, esses resultados também convergem com as conclusões de pesquisas que utilizaram outros métodos, como Polyzos, Tsiotas e Minetos (2013) e Brito e Silva (2020).

Na sequência, estimou-se três modelos *XGBoost*: o primeiro sem efeitos espaciais (X1); o segundo introduzindo a *dummy* de contiguidade (X2) e; o último introduzindo a variável do tipo SLX (X3). Os três modelos foram aplicados através de dois métodos diferentes de escolha de parâmetros. Na primeira aplicação, foram ajustados inicialmente com os parâmetros no padrão do *software*, seguido de um segundo ajuste de acordo com o nível de RMSE das iterações. Este segundo ajuste ocorreu da seguinte forma: a cada nova iteração o modelo consegue reduzir mais um pouco o erro da iteração anterior, porém, há um limite para isso e, em um determinado ponto, o modelo para de reduzir o erro e pode até mesmo voltar a elevá-lo. Neste ponto o modelo passa a sofrer de sobreajuste e a partir daí as previsões serão prejudicadas.

Já o segundo método utiliza um protocolo de validação cruzada (Spadon *et al*, 2019), que consiste em dividir o conjunto de dados em várias partes, treinando o modelo em algumas dessas partes e validando-o nas restantes. Esse processo é repetido várias vezes, com diferentes partes usadas para validação a cada iteração e então os parâmetros são determinados de maneira automática, também utilizando o menor RMSE possível como critério de escolha dos melhores parâmetros.

Dessa forma, nos três modelos, o método padrão do software com escolha manual de iterações foi superior: no caso do modelo sem efeitos espaciais (X1) foram escolhidas 22 iterações; no modelo com *dummy* de contiguidade (X2) foram escolhidas 23 iterações e; no modelo com variável de interação (X3), 27 iterações.

Utilizou-se a ferramenta SHAP, que estimou as contribuições das características à previsão da variável alvo, resultando na figura 2 para o modelo sem efeitos espaciais (X1).

Cada ponto escuro é uma observação da característica (Dij, if0_renda, if1_renda), a qual, no eixo horizontal, é atribuída ao seu valor SHAP correspondente, representando assim uma explicação local do modelo. A linha azul em cada gráfico é a forma suavizada que se ajusta aos pontos individuais e ajuda a visualizar a tendência geral de cada característica em relação à sua contribuição para as previsões. A curva representa uma explicação global, onde se pode verificar a relação entre a característica e o impacto na previsão e evidenciar tendências ou padrões.

Inicialmente, ao se observar as linhas percebe-se que se trata de relações não lineares. Na figura 2(a) - referente à distância (Dij) - mostra que até aproximadamente 35km, quando a linha atravessa o eixo vertical no valor zero, a relação deixa de ser positiva entre pendularidade e distância. Ou seja, os trabalhadores pendulares preferem se deslocar para trabalhar em outro município até essa distância. A partir disso o espaço percorrido se torna um fator de redução da pendularidade.

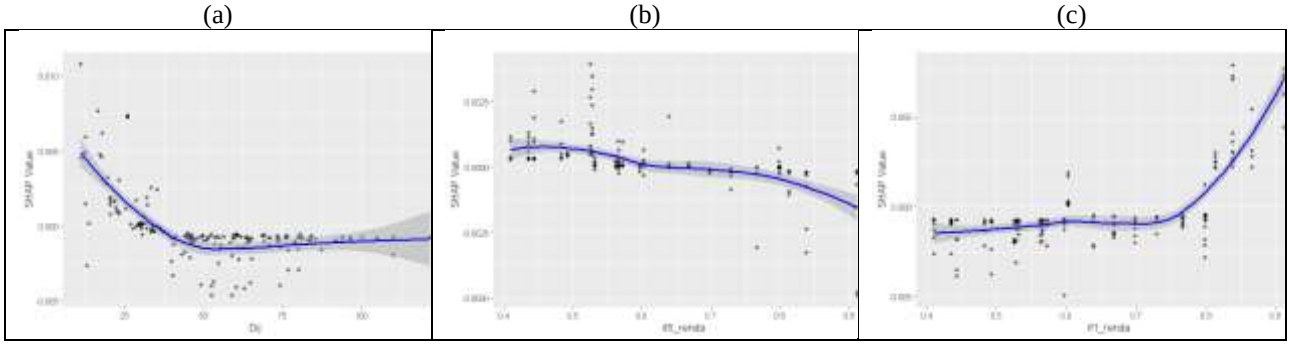
Na figura 2(b) é possível perceber que o município de origem doa trabalhadores até o ponto em que seu IFDM-ER atinge o valor aproximado de 0,65. A partir disso, dado que o município de origem passa a ter uma melhor dinâmica de mercado de trabalho, supõe-se que o *commuter* prefira permanecer trabalhando no município de residência. Ou seja, ele só se desloca para trabalhar em outro município quando o município em que vive apresenta baixa dinâmica de mercado de trabalho em relação aos outros.

Se considerar que a RML tinha quinze municípios com um IFDM-ER inferior a 0,65, é possível inferir que a maioria desses municípios, correspondendo a 60% do total, demonstrava indícios de um fenômeno de força centrífuga, no qual trabalhadores são repelidos para outras

localidades. Essa tendência sugere que as condições desfavoráveis do mercado de trabalho local na maioria dos municípios da RML desempenham um papel significativo na determinação dos padrões de deslocamento pendular dos trabalhadores

Figura 2: Valores SHAP - modelo (X1)

(a) (b) (c)



Fonte: Elaborado pela pesquisa.

Sobre a Figura 2(c), o efeito se mostrou inverso, de modo que, enquanto o município de destino mantiver uma dinâmica de mercado de trabalho desfavorável, o trabalhador pendular tende a evitar o deslocamento. No entanto, a partir de um determinado patamar do IFDM-ER, aproximadamente 0,77[15], ele opta por realizar o movimento. Em outras palavras, o município de destino passa a atrair trabalhadores pendulares quando apresenta uma dinâmica de mercado de trabalho favorável, enquanto deixa de exercer tal atração quando suas condições são desfavoráveis.

Na figura 3a obtêm-se explicações tanto locais quanto globais dos valores SHAP, em que, é possível realizar uma análise da importância relativa de cada variável sobre as previsões. Isso é viabilizado pela escala apresentada logo à direita das características no eixo vertical, onde os valores representam as médias dos valores SHAP absolutos, organizados por ordem decrescente de impacto na previsão da variável alvo.

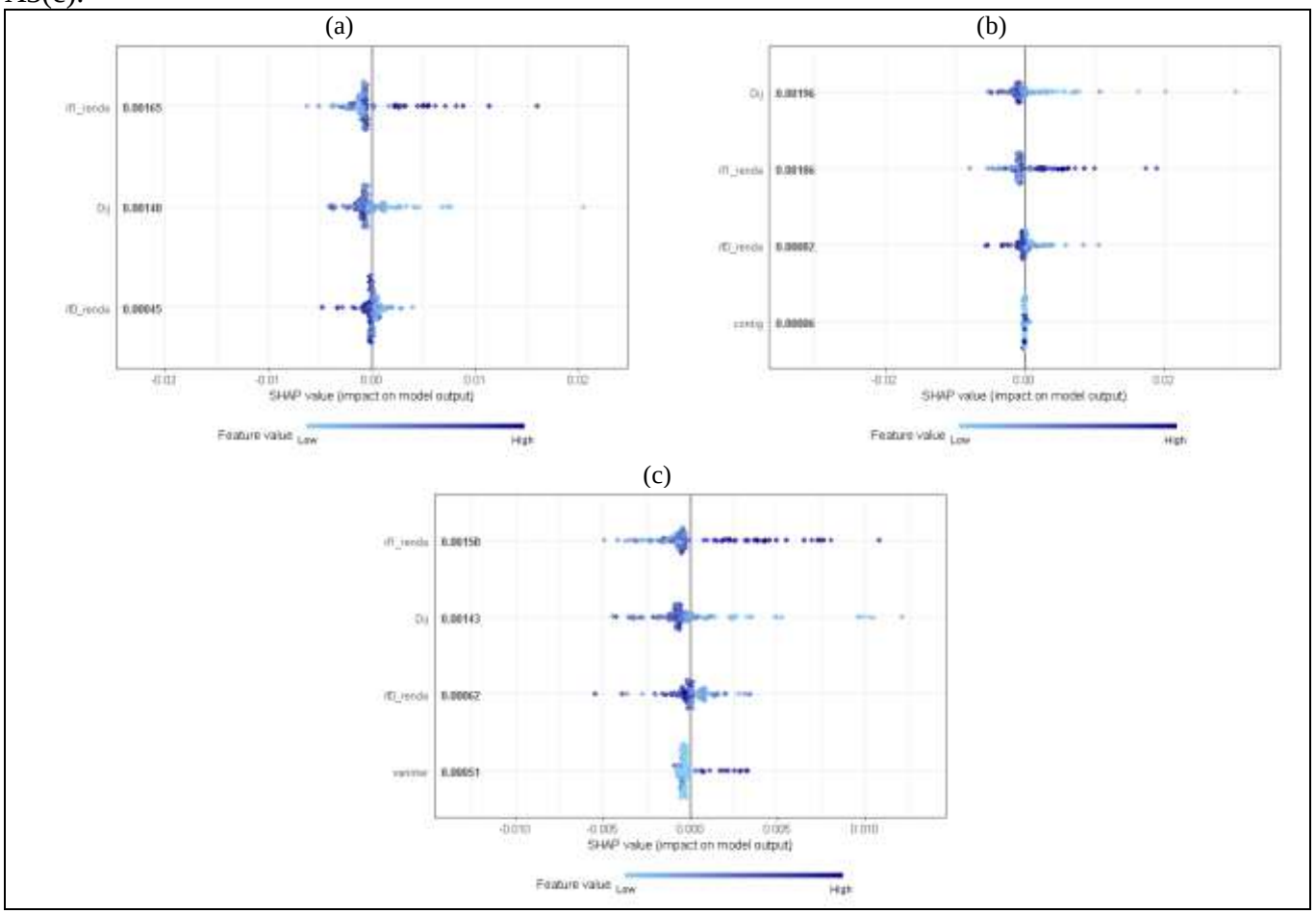
Os pontos mais escuros representam valores mais altos das características e, à medida em que vão ficando mais claros, as observações correspondentes vão reduzindo de tal maneira que o tom mais claro de azul representa o valor da menor observação. Já no eixo horizontal, encontram-se os valores SHAP correspondentes de cada variável. Os resultados do modelo (X1) apontam o seguinte:

- ✓ If1_renda: Interpretando o gráfico de maneira global, esta é a variável de maior importância na previsão de deslocamento pendulares, sendo responsável, em média, por aproximadamente 0,00165 na previsão absoluta média do modelo. Agora localmente, valores altos do IFDM-ER do município de destino (pontos escuros) estão associados, em sua maioria, a valores SHAP positivos, o que faz com que sejam responsáveis por aumentos nos valores das previsões individuais. Valores baixos (pontos mais claros) estão associados à valores SHAP negativos, o que faz com que sejam responsáveis por reduções nos valores das previsões individuais. Em termos práticos, isto significa que municípios de destino com alto índice Firjan recebem mais trabalhadores pendulares e municípios com baixo índice recebem menos.
- ✓ Dij: Esta é a segunda variável mais relevante na previsão, com valor absoluto médio de 0,0014. Ao contrário da primeira característica, pode-se perceber que a maioria dos pontos mais claros está associada a valores SHAP positivos, enquanto os pontos mais escuros estão associados a valores SHAP negativos. Isto significa que distâncias mais curtas entre os municípios aumentam os deslocamentos pendulares entre eles e distâncias mais longas reduzem-nos.
- ✓ If0_renda: A maioria das observações está próxima de zero no eixo horizontal, isso faz com que o IFDM da cidade de origem do deslocamento seja a característica com menor influência sobre a previsão, mas não é nula e possui relevância. É possível perceber

[15] Esse valor do IFDM-ER seria o ponto em que a linha atravessa o eixo horizontal de valor SHAP zero da Figura 6(c).

efeitos antagônicos de acordo com a observação da característica, pois os pontos mais claros estão à direita de zero no eixo horizontal e os pontos mais escuros estão à esquerda, ou seja, municípios de origem dos deslocamentos com baixo índice Firjan doam mais trabalhadores pendulares e, quando possuem alto índice, tendem a manter seus trabalhadores no território.

Figura 3: Médias dos valores SHAP absolutos e impacto na variável explicada-modelo X1(a), X2(b), X3(c).



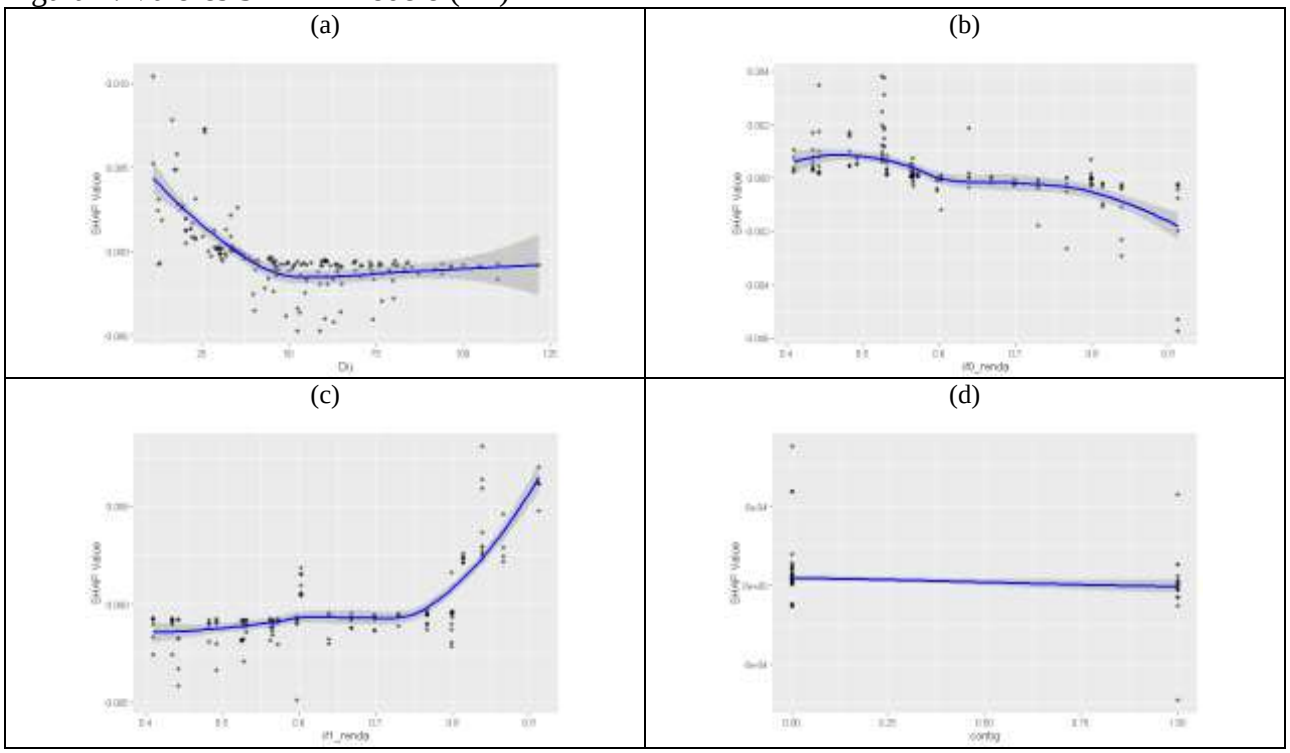
Fonte: Elaborado pela pesquisa.

No modelo (X2) a variável de contiguidade foi incorporada e a Figura 4 mostra os valores SHAP correspondentes a cada variável explicativa, as quais apresentaram similaridade com o observado no modelo (X1) – Figura 2. No caso da variável que mensura a contiguidade, como se trata de uma *dummy*, as observações assumiram valores 0 ou 1. Percebe-se (Figura 4d) que, dentre os valores 0 das observações, a maioria possui valor SHAP positivo, isto significa que municípios que não são vizinhos fronteiriços mesmo assim tendem a trocar pendulares. Esta conclusão faz sentido uma vez que, mesmo não sendo vizinhos, todos os municípios da amostra estão suficientemente próximos e inseridos na mesma RM.

Em contrapartida, do outro lado da Figura 4d, apesar da curva dos valores SHAP se encontrar um pouco abaixo de zero no eixo vertical, as observações dos municípios que são vizinhos parecem estar dispostas de maneira simétrica em relação a esse eixo. Ou seja, há municípios vizinhos que possuem maior dinâmica de deslocamentos pendulares enquanto outros pares de vizinhos possuem menor dinâmica.

Isto não é necessariamente contraintuitivo pois, por exemplo, no caso do município Bela Vista do Paraíso e Sertanópolis, o efeito provocado sobre a pendularidade pelo fato de ambas as cidades serem vizinhas não se sobrepõe ao efeito que o município-polo exerce sobre elas. Dessa forma, ambas doam mais pendulares à Londrina e trocam menos pendulares entre si, fazendo com que, mesmo sendo vizinhas por contiguidade, possuam uma dinâmica menor de deslocamentos pendulares.

Figura 4: Valores SHAP - modelo (X2)



Fonte: Elaborado pela pesquisa.

Apesar disso, o efeito da variável de contiguidade não parece ser tão claro, não podendo inferir se a média do valor do seu SHAP representa um efeito majoritariamente positivo ou negativo, além do seu grau de importância para a previsão média. Portanto, a figura 3b contribuiu com essa interpretação.

No modelo (X1), a variável IFDM-ER do município de destino era a mais influente na previsão (Figura 3a). No modelo (X2), a distância entre os pares de cidades assume esse papel (Figura 3b). Observa-se que distâncias maiores reduzem a previsão média dos deslocamentos pendulares, enquanto um alto IFDM-ER do município de destino tende a aumentá-la. O IFDM-ER do município de origem tem efeito inverso.

A média do valor SHAP para a contiguidade é 0.00006, com observações próximas de zero, indicando a heterogeneidade do efeito da vizinhança na previsão. Isso sugere que ser um município vizinho não é necessariamente um fator de atração para trabalhadores pendulares, já que municípios mais distantes, com maior dinâmica de mercado de trabalho, exercem uma força de atração maior.

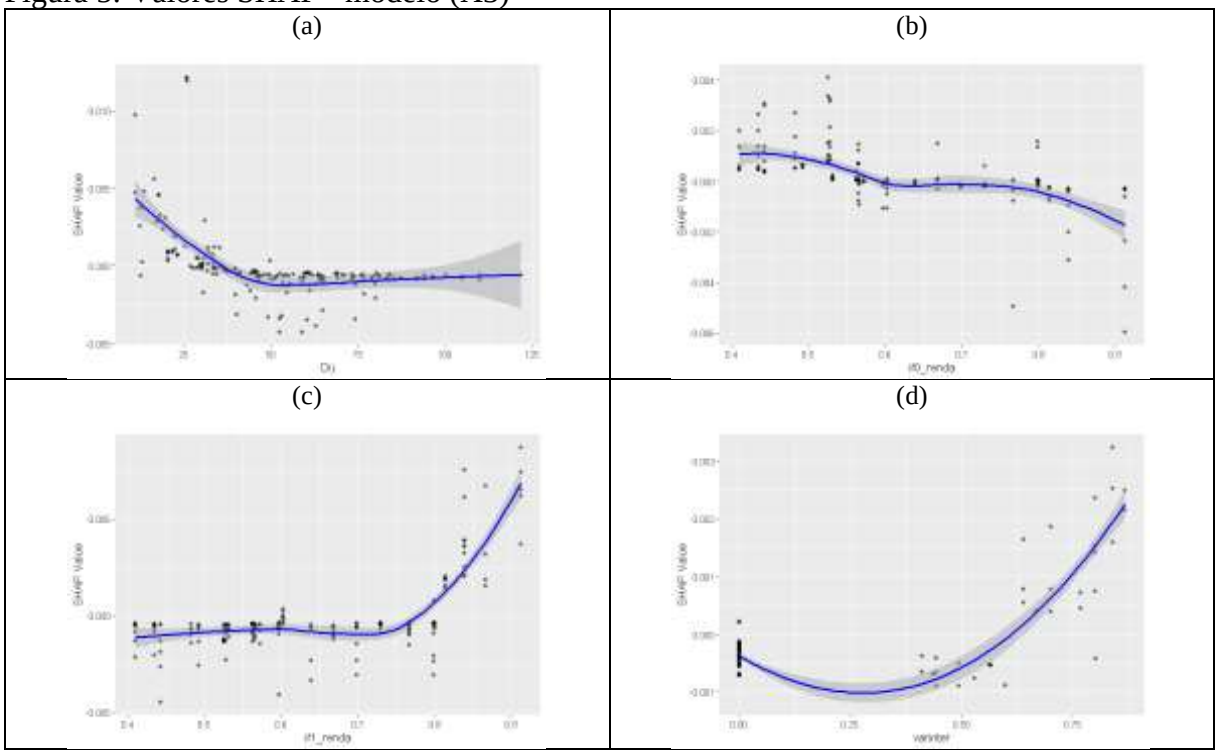
Por fim, na Figura 5 apresenta os valores SHAP do modelo (X3), incluindo a variável de interação entre “ter contiguidade” e IFDM-ER do município de destino. A distância e os índices Firjan dos municípios de origem e de destino se comportam da mesma maneira que nos outros dois modelos (X1) e (X2), sendo as interpretações idênticas às anteriores.

No caso da variável de interação, observa-se (Figura 5d) que quase todos os valores zero das suas observações estão associados à valores SHAP negativos, um indicativo de que, quando a variável de interação não está presente, a tendência é de que o nível de pendularidade seja menor do que o esperado. Por outro lado, a partir de um valor de aproximadamente 0,6 para essa característica, os valores SHAP correspondentes são positivos. Isto significa que municípios de destino que, além de serem vizinhos do município de origem do deslocamento pendular, possuem também um nível de desenvolvimento de mercado de trabalho mais elevado, possuem um diferencial para atrair pendulares em relação aos que são somente vizinhos ou somente possuem alto nível de IFDM-ER.

Para completar a análise do modelo, a Figura 3c ilustra a média do valor SHAP para cada característica de (X3). No caso da variável de interação “mercado de trabalho versus contiguidade”, sua influência na previsão da variável alvo é a menor em comparação com as demais características

analisadas. Entretanto, observa-se de forma mais proeminente o efeito positivo associado aos valores mais elevados dessa variável. Esses achados sugerem que a presença de vizinhos com um mercado de trabalho altamente desenvolvido tende a gerar um fluxo pendular mais significativo, quando comparado aos municípios não vizinhos.

Figura 5: Valores SHAP - modelo (X3)



Fonte: Elaborado pela pesquisa.

Todos os modelos gravitacionais apresentaram coeficientes estatisticamente significativos e convergentes. Os modelos *XGBoost* também apresentaram resultados que corroboram toda a análise realizada nos modelos gravitacionais. Dessa forma, como todos os experimentos lograram êxito a trazer conclusões idênticas, pode-se partir para a comparação do desempenho destes e determinar qual foi o mais adequado para prever os deslocamentos pendularidade da RML. Essa comparação se encontra na Tabela 2.

Tabela 2: Métricas de desempenho dos modelos Gravitacional e XGBoost – (G1), (G2), (G3), (X1), (X2) e (X3).

Modelos	CPC	CPL	RMSE
(G1)	0,43	0,42	0,005
(G2)	0,42	0,42	0,005
(G3)	0,42	0,42	0,005
(X1)	0,44	0,45	0,004
(X2)	0,41	0,47	0,004
(X3)	0,42	0,51	0,004

Fonte: Elaborado pela pesquisa.

Os modelos gravitacionais são equivalentes, com métricas praticamente idênticas. Dessa forma, a busca pela modelagem de efeitos espaciais através da *dummy* contiguidade e da variável de interação, apesar de significativas e relevantes, não foram suficientes para elevar a capacidade preditiva. Já dentre os modelos *XGBoost*, observou-se um pequeno ganho de desempenho em (X3) com a inserção da variável de interação, obtido pela métrica CPL.

Comparando os modelos gravitacional e *XGBoost*, pode-se perceber que possuem métricas parecidas, com destaque para o modelo (X3) que performou um pouco melhor. Dessa forma, conclui-

se que o modelo gravitacional proposto, pelo menos para a amostra desta pesquisa, foi capaz de desempenhar de maneira satisfatória frente ao modelo *XGBoost.*

Por fim, os efeitos espaciais foram significativos, o que é interessante pois traz à tona novas análises como as realizadas nesta pesquisa; contudo, apesar de melhorarem de maneira tímida a capacidade preditiva do modelo *XGBoost*, não foram suficientes para elevar o desempenho do modelo gravitacional, apenas mantendo-o no mesmo patamar em termos das métricas estabelecidas, o que não inviabiliza a sua utilização dentro da modelagem dado que foram significativos e relevantes para as interpretações.

Isto leva a conclusão de que provavelmente os efeitos espaciais não tiveram tanta influência sobre os deslocamentos quanto poderiam porque o espaço limitado da RML impede que se observe outras relações.

CONSIDERAÇÕES FINAIS

Esta pesquisa analisou o efeito da distância geográfica e da característica de mercado de trabalho entre os pares de municípios da Região Metropolitana de Londrina em 2010, bem como analisou a introdução de efeitos espaciais, lançando mão de dois modelos distintos. Além disso, realizou comparação do desempenho dos modelos e determinou a estimação mais adequada.

Os experimentos e os resultados confirmaram as evidências apresentadas na literatura sobre pendularidade, porém, comprovaram que o modelo gravitacional, após alguma sofisticação, ainda possui capacidade preditiva adequada quando comparado com um modelo de aprendizado moderno como o *XGBoost*. Além disso, o presente estudo trouxe uma nova análise do modelo gravitacional ao produzir uma variável de interação espacial em um regressor e, portanto, respondeu à questão de pesquisa levantada.

Morton, Piburn e Nagle (2018) sugeriram que modelos *XGBoost*, no contexto da pendularidade analisada em sua pesquisa, deveriam ser comparados com modelos mais complexos do que o gravitacional utilizado por eles. Esta tarefa foi realizada através da sofisticação desta modelagem. Simini *et al.* (2012) criticam a baixa capacidade preditiva do modelo gravitacional, porém, o modelo proposto na pesquisa foi capaz de alcançar métricas de desempenho pelo menos tão boas quanto as do *XGBoost.* Como as comparações entre *XGBoost* e modelo gravitacional realizadas por outros autores, como Spadon *et al* (2019) e Liang *et al* (2021) foram realizadas com este em sua forma tradicional, a forma proposta nesta análise se mostrou satisfatória frente ao desempenho do modelo de ML uma vez que as métricas de ambos apresentaram valores muito próximos.

Isto significa que o modelo gravitacional ainda possui relevância na pesquisa sobre pendularidade, principalmente considerando o fato de que o modelo *XGBoost* é mais complexo e sua aplicação possui o obstáculo, que não existe no modelo gravitacional, da necessidade da estimação correta dos parâmetros *ex ante*. Além disso, o modelo gravitacional possui uma estrutura teórica mais consolidada, baseada na econometria tradicional, ou seja, pode-se utilizar mais ferramentas de diagnóstico, o que faz com que se tenha mais confiança nos seus resultados, tornando a inferência estatística um processo mais fácil e mais claro.

Apesar de modelar adequadamente os deslocamentos pendulares, é necessário destacar algumas limitações da pesquisa. A primeira limitação diz respeito ao espaço estudado. Os efeitos espaciais se restringiram à vizinhança dentro da RML, não considerando outros municípios fora desta região. Existe fluxo de pendulares entre municípios dentro da RML e municípios situados fora da região que estão à uma distância adequada para essa pesquisa, como Arapongas e Apucarana, por exemplo. Esta interação é de fato relevante, mas, foi eliminada do estudo sob a hipótese de se analisar a RML.

Dado que uma maior dinâmica econômica do mercado de trabalho da cidade de destino e/ou uma menor dinâmica do mercado de trabalho da cidade de origem e/ou uma menor distância entre as cidades, são fatores que afetam positivamente a pendularidade, conforme observado em todos os

modelos estimados, então é fundamental que as políticas públicas se concentrem em abordar essas questões de forma integrada.

REFERÊNCIAS

ABOUELHAMD, Islam. The Relationship between Urban Spatial Structure & Commuting Patterns: Literature Review. JES. **Journal of Engineering Sciences**, v. 49, n. 5, set. 2021, p. 662–78.

AHRENS, Annette; LYONS, Seán. Do rising rents lead to longer commutes? A gravity model of commuting flows in Ireland. **Urban Studies**, v. 58, n. 2, p. 264-279, fev. 2021.

ARVIS, Jean-François; SHEPHERD, Ben. The Poisson Quasi-Maximum Likelihood Estimator: A Solution to the 'Adding up' Problem in Gravity Models. **Applied Economics Letters**, v. 20, n. 6, abr. 2013, p. 515-519.

BRITO, Danyella Juliana Martins de; BRITO RAMALHO, Hilton Martins de. "Caracterização e Determinantes dos Movimentos Pendulares na Região Metropolitana do Recife: Evidências a partir de um Modelo Gravitacional". **Revista Econômica do Nordeste**, v. 50, n. 2, ago. 2019, p. 95–113.

BRITO, Danyella Juliana; SILVA, Marcus Vinicius Amaral e. "Determinantes dos movimentos pendulares no Brasil: uma análise espacial". **Estudios económicos**, vol. 38, no 76, fevereiro de 2021, p. 95–122.

CARROTHERS, Gerald A. P. An Historical Review of the Gravity and Potential Concepts of Human Interaction. **Journal of the American Institute of Planners**, v. 22, n. 2, p. 94-102, jan. 1956.

CHEN, Hui et al. "Data-Driven Analysis on Inter-City Commuting Decisions in Germany". **Sustainability**, vol. 13, no 11, janeiro de 2021, p. 6320.

CHEN, Tianqi; GUESTRIN, Carlos. XGBoost: **A Scalable Tree Boosting System. In: Proceedings of the 22nd ACM SIGKDD** *In:* **International Conference on Knowledge Discovery and Data Mining**. Association for Computing Machinery, 2016, p. 785-794.

DELGADO, Paulo et al. "Caracterização dos movimentos pendulares nas regiões metropolitanas do Paraná". **Caderno IPARDES - Estudos e Pesquisas**, vol. 3, no 1, julho de 2013, p. 1–24.

FRIEDMAN, Jerome H. Greedy function approximation: A gradient boosting machine. **The Annals of Statistics**, v. 29, n. 5, p. 1189-1232, out. 2001.

HAZANS, Mihails. **Commuting in the Baltic States**: Patterns, determinants and gains. ZEI working paper, No. B 02-2003, 2003.

HEAD, Keith; MAYER, Thierry. Gravity Equations: Workhorse, Toolkit, and Cookbook. In: GOPINATH, Gita et al. (Org.). **Handbook of International Economics**. vol. 4. Elsevier, 2014, p. 131-195.

LI, Zhipeng; NIU, Xinyi. Exploring Spatial Nonstationarity in Determinants of Intercity Commuting Flows: A Case Study of Suzhou–Shanghai, China. **ISPRS International Journal of Geo-Information**, v. 11, n. 6, p. 335, junho de 2022.

LIANG, Quan et al. Travel Destination Prediction of Public Transport Commuters by Integrating XGBoost Algorithm and Graph Adjustment Method, v. 39, n. 4, p. 68-76, 2021.

LUNDBERG, Scott; LEE, Su-In. **A Unified Approach to Interpreting Model Predictions**. arXiv:1705.07874v2 [cs.AI], 25 de novembro de 2017.

MA, Meihong, et al. XGBoost-based method for flash flood risk assessment. **Journal of Hydrology**, vol. 598, julho de 2021, p. 126382.

MANGALATHU, Sujith, et al. “Failure mode and effects analysis of RC members based on machine-learning-based SHapley Additive exPlanations (SHAP) approach”. **Engineering Structures**, vol. 219, setembro de 2020, p. 110927.

MORTON, April et al. Need A Boost? **A Comparison of Traditional Commuting Models with the XGBoost Model for Predicting Commuting Flows**. *In:* **Oak Ridge National Laboratory (ORNL**), Oak Ridge, TN (United States), 1o de agosto de 2018.

MOURA, Rosa. Movimento pendular da população no Paraná: evidência da desconexão moradia/trabalho. **Cadernos Metrópole**, v. 12, n. 23, jan. 2010.

OSMAN, Ahmedbahaaaldin Ibrahem Ahmed, et al. Modelo Extreme Gradient Boosting (XGBoost) para Prever os Níveis de Água Subterrânea em Selangor, Malásia. **Ain Shams Engineering Journal**, vol. 12, no. 2, junho de 2021, p. 1545–56.

POLYZOS, Serafeim; TSIOTAS, Dimitrios; MINETOS, Dionissios. Determining the Driving Factors of Commuting: An Empirical Analysis from Greece**. Journal of Engineering Science and Technology Review**, v. 6, p. 46-55, 2013.

RAMANESWARAN, S., et al. Modelo Híbrido Inception v3 XGBoost para Classificação de Leucemia Linfoblástica Aguda. **Computational and Mathematical Methods in Medicine**, vol. 2021, julho de 2021, p. e2577375.

RENKOW, Mitch; HOOVER, Dale. Commuting, Migration, and Rural-Urban Population Dynamics. **Journal of Regional Science**, v. 40, n. 2, p. 261-287, maio de 2000.

ROBINSON, Caleb; DILKINA, Bistra. **A Machine Learning Approach to Modeling Human Migration**. *In:* Proceedings of the 1st ACM SIGCAS **Conference on Computing and Sustainable Societies**. Association for Computing Machinery, 2018, p. 1-8.

SANTOS SILVA, J. M. C.; TENREYRO, Silvana. The Log of Gravity. **The Review of Economics and Statistics**, v. 88, n. 4, p. 641-658, nov. 2006.

SIMINI, Filippo *et al*. A Universal Model for Mobility and Migration Patterns. **Nature**, v. 484, n. 7392, p. 96-100, abr. 2012.

SPADON, Gabriel *et al*. Reconstructing commuters network using machine learning and urban indicators. **Scientific Reports**, v. 9, n. 1, p. 11801, agosto de 2019.

STEFANOULI, Maria; POLYZOS, Serafeim. Gravity vs Radiation Model: Two Approaches on Commuting in Greece. **Transportation Research Procedia**, v. 24, p. 65-72, janeiro de 2017.

TAVARES, Érica; MONTEIRO, Jéssica. Movimentos pendulares para trabalho e estudo: estratégias metodológicas a partir dos censos demográficos de 2000 e 2010. **Geosul**, v. 34, n. 73, p. 33-58, dezembro de 2019.

WANG, Feier, et al. “Spatial Heterogeneity Modeling of Water Quality Based on Random Forest Regression and Model Interpretation”. **Environmental Research**, vol. 202, novembro de 2021, p. 111660.

WÖLWER, Anna-Lena *et al*. Gravity Models in R. **Austrian Journal of Statistics**, v. 47, n. 4, p. 16-35, jun. 2018.

ZHAO, Pengjun, et al. "Revisiting the Gravity Laws of Inter-City Mobility in Megacity Regions". **Science China Earth Sciences**, vol. 66, no. 2, fevereiro de 2023, p. 271-281.