UNIVERSIDADE ESTADUAL DE PONTA GROSSA
SETOR DE CIÊNCIAS SOCIAIS APLICADAS
DEPARTAMENTO DE ECONOMIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ECONOMIA

SAULO MURILO DE SOUSA NEPOMUCENO

# MODELO GRAVITACIONAL *VERSUS XGBOOST*: APLICAÇÕES PARA A PENDULARIDADE NA REGIÃO METROPOLITANA DE LONDRINA

PONTA GROSSA
2024

SAULO MURILO DE SOUSA NEPOMUCENO

# MODELO GRAVITACIONAL *VERSUS XGBOOST*: APLICAÇÕES PARA A PENDULARIDADE NA REGIÃO METROPOLITANA DE LONDRINA

Dissertação apresentada como requisito parcial à obtenção do título de Mestre em Economia, no Programa de Pós-Graduação em Economia, Setor de Ciências Sociais Aplicadas da Universidade Estadual de Ponta Grossa.

Orientadora: Prof.ª Dr.ª Augusta Pelinski Raiher

PONTA GROSSA

2024

Nepomuceno, Saulo Murilo de Sousa

N441 Modelo gravitacional *versus XGBoost*: aplicações para a pendularidade na Região Metropolitana de Londrina / Saulo Murilo de Sousa Nepomuceno. Ponta Grossa, 2024.

94 f.

Dissertação (Mestrado em Economia - Área de Concentração: Economia), Universidade Estadual de Ponta Grossa.

Orientadora: Profa. Dra. Augusta Pelinski Raiher.

1. Modelo gravitacional. 2. Pendularidade. 3. Xgboost. I. Raiher, Augusta Pelinski. II. Universidade Estadual de Ponta Grossa. Economia. III.T.

CDD: 330.1

Ficha catalográfica elaborada por Maria Luzia Fernandes Bertholino dos Santos- CRB9/986

UNIVERSIDADE ESTADUAL DE PONTA GROSSA
Av. General Carlos Cavalcanti, 4748 - Bairro Uvaranas - CEP 84030-900 - Ponta Grossa - PR - https://uepg.br

# TERMO

SAULO MURILO DE SOUSA NEPOMUCENO

**MODELO GRAVITACIONAL VERSUS XGBOOST: APLICAÇÕES PARA PENDULARIDADE NA REGIÃO METROPOLITANA DE LONDRINA**

**D**issertação aprovada como requisito parcial para obtenção do grau de Mestre no Programa de Pós-Graduação em Economia, da Universidade Estadual de Ponta Grossa, pela seguinte banca examinadora:

Ponta Grossa, 22 de março de 2024

Augusta Pelinski Raiher – Universidade Estadual de Ponta Grossa

Cristiano Stamm - Universidade Estadual do Oeste do Paraná

Luma de Oliveira - Universidade Estadual de Ponta Grossa

Documento assinado eletronicamente por **Augusta Pelinski Raiher**, **Professor(a)**, em 22/03/2024, às 22:15, conforme Resolução UEPG CA 114/2018 e art. 1º, III, "b", da Lei 11.419/2006.

Documento assinado eletronicamente por **Luma de Oliveira**, **Professor(a)**, em 25/03/2024, às 16:29, conforme Resolução UEPG CA 114/2018 e art. 1º, III, "b", da Lei 11.419/2006.

A autenticidade do documento pode ser conferida no site https://sei.uepg.br/autenticidade informando o código verificador **1888177** e o código CRC **F2F4DBAB**.

## AGRADECIMENTOS

Sou profundamente grato à minha mãe, Valnice, pelo suporte inabalável e à minha esposa, Nataly, por seu companheirismo, seu ombro amigo e por acreditar em mim até mesmo quando nem eu acreditei. Agradeço ao meu pai, Maurinélio, pela sua integridade e seu apoio constante que sempre me inspiraram. Ao meu irmão, Paulo Mateus, que tem sido uma fonte de amizade e de incentivo.

Agradeço aos professores do PPGECO/UEPG, em especial à minha orientadora, Profª Drª. Augusta Raiher, por suas orientações valiosas e por contribuir para meu crescimento acadêmico. Por fim, agradeço a todos os familiares e os amigos que influenciaram minha jornada até aqui.

## RESUMO

Esta pesquisa tem como objetivo principal analisar os determinantes dos deslocamentos pendulares entre os pares de municípios da Região Metropolitana de Londrina em 2010. Para cumprir o objetivo, uma amostra de 600 deslocamentos entre as cidades foi utilizada para 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 da capacidade preditiva. Os resultados apontam que a distância é um fator de repulsão de trabalhadores pendulares, as variáveis de mercado de trabalho do município de origem e de destino são fatores de repulsão e de atração, respectivamente. Os efeitos espaciais são relevantes, apesar de não elevarem a capacidade preditiva dos modelos e, por último, o modelo gravitacional, após a sofisticação proposta, conseguiu alcançar o desempenho do *XGBoost* dadas as métricas de comparação e a característica da amostra estudada.



**Palavras-chave:** Pendularidade; Modelo gravitacional; *XGBoost*.

## ABSTRACT

This research aims to analyze the determinants of commuting movements between pairs of cities in the Metropolitan Region of Londrina in 2010. To achieve this goal, a sample of 600 pairs of cities was used to model a gravitational equation with pseudo-maximum likelihood Poisson estimation and an XGBoost model. Furthermore, spatial effects were included in both models, and their predictive capacity was compared. The results indicate that distance is a repelling factor for commuting workers, while the labor market variables of the origin and destination cities are repelling and attracting factors, respectively. Spatial effects are relevant, although they do not enhance the predictive capacity of the models. Lastly, the gravitational model, after the proposed sophistication, managed to match the performance of the XGBoost given the comparison metrics and the characteristics of the studied sample.



**Keywords:** Commuting; Gravity model; XGBoost.

## LISTA DE FIGURAS

Figura 1: Mapa de rodovias e de coordenadas geográficas da RM de Londrina.....62

Figura 2: Disposição espacial dos municípios da RML por quantidade de trabalhadores ocupados.....68

Figura 3: Mapa do processo de conurbação na RML.....68

Figura 4: Mapa dos índices Firjan - Emprego e Renda (IFDM-ER) da RML.....69

Figura 5: Mapa de deslocamentos pendulares da RML .....70

Figura 6: Valores SHAP - modelo (X1).....75

Figura 7: Médias dos valores SHAP absolutos e impacto na variável explicada - modelo (X1). .....76

Figura 8: Valores SHAP - modelo (X2).....78

Figura 9: Médias dos valores SHAP absolutos e impacto na variável explicada - modelo (X2). .....79

Figura 10: Valores SHAP - modelo (X3).....80

Figura 11: Médias dos valores SHAP absolutos e impacto na variável explicada - modelo (X3). .....81

## LISTA DE TABELAS

Tabela 1: Resumo de algumas evidências empíricas acerca dos determinantes da pendularidade .....26

Tabela 2: Descrição das variáveis explicativas utilizadas nas estimações. .....64

Tabela 3: Descrição dos experimentos econométricos. .....64

Tabela 4: Características selecionadas dos municípios da RML - 2010 .....67

Tabela 5: Coeficientes do modelo gravitacional PPML – modelos (G1), (G2) e (G3).....73

Tabela 6: Métricas de desempenho dos modelos Gravitacional e XGBoost – (G1), (G2), (G3), (X1), (X2) e (X3). .....82

Tabela 7: Coeficientes do modelo gravitacional MQO (Y+1) – modelos (M1), (M2) e (M3) 94

Tabela 8: Coeficientes do modelo gravitacional GPML – modelos (GP1), (GP2) e (GP3).....94

**SUMÁRIO**


**1. INTRODUÇÃO ..... 10**

**2. A PENDULARIDADE NA LITERATURA ECONÔMICA ..... 16**

**2.1 O conceito de movimento pendular ..... 16**

**2.2 As análises empíricas sobre os movimentos pendulares ..... 20**

**2.3 Percepções empíricas sobre os modelos gravitacional e *XGBoost* ..... 29**

**3. PROCEDIMENTOS METODOLÓGICOS ..... 33**

**3.1 O modelo gravitacional ..... 34**

**3.2 Os efeitos espaciais na pendularidade. ..... 42**

**3.3 *XGBoost* ..... 45**

**3.4 As métricas de comparação ..... 53**

**3.5 O espaço analisado ..... 55**

**3.6 As variáveis Selecionadas e Fonte dos dados ..... 56**

3.6.1 Taxa de deslocamento pendular ..... 57

3.6.2 Índice Firjan – Emprego e Renda (IFDM-ER) ..... 59

3.6.3 Distância entre os pares de municípios. ..... 61

3.6.4 Resumo da Estratégia Empírica utilizada nas Estimações ..... 63

**4. PENDULARIDADE NA RML: EVIDÊNCIA EMPÍRICA DOS SEUS DETERMINANTES ..... 66**

**4.1 Os movimentos pendulares na RM de Londrina ..... 66**

**4.2 Resultados dos modelos gravitacional e *XGBoost*: Determinantes da Pendularidade da RML ..... 72**

**CONSIDERAÇÕES FINAIS ..... 84**

**REFERÊNCIAS ..... 87**

**APÊNDICE A – ESTIMAÇÕES ADICIONAIS DO MODELO GRAVITACIONAL ... 94**

# 1. INTRODUÇÃO

A distância, o tempo e o custo despendidos por uma pessoa para realizar alguma atividade fora de sua residência e retornar compreendem o conceito de movimento ou deslocamento pendular. Este fenômeno, também chamado de pendularidade, já tem sido objeto de estudo da comunidade científica há algumas décadas, entretanto, o interesse em analisá-lo cresce à medida em que se observa cada vez mais pessoas tomando a decisão de trabalhar em outro município e um processo de adensamento urbano que aproxima as cidades em vários aspectos.

O movimento pendular específico que esta pesquisa investigará é a quantidade de indivíduos que se deslocam de suas residências para trabalhar em municípios diferentes, constituindo a entrada de trabalhadores oriundos de uma determinada cidade para outra. Esta tendência tem se intensificado notavelmente no Brasil desde o início do século XXI, especialmente em cidades de porte médio no interior do país, as quais têm demonstrado uma taxa de crescimento significativamente maior (Stamm, 2005).

A pendularidade apresenta uma gama de impactos tanto positivos quanto negativos relacionados às características da estrutura urbana para as cidades receptoras e para as cidades de origem dos trabalhadores pendulares, como mercado de trabalho, educação, lazer, infraestrutura e aspectos culturais entre os espaços (Tavares; Monteiro, 2019). A disjunção entre moradia e trabalho não somente requer políticas de melhoria do deslocamento, mas também políticas de infraestrutura social com creches, postos de saúde, programas de cultura, de moradia e de lazer posto que a família do *commuter*[1] permanece na cidade de origem (Cintra; Delgado; Moura, 2012).

No que diz respeito às cidades receptoras, o movimento pendular amplia a disponibilidade de força de trabalho, permitindo o preenchimento de lacunas de habilidades e a ocupação de empregos especializados. Isso, por sua vez, fomenta a inovação e a diversificação da economia local. No entanto, essas cidades também enfrentam desafios, como o estrangulamento da infraestrutura de transporte (Moura, 2010) conforme o aumento de veículos e de pessoas naqueles espaços e a vulnerabilidade econômica resultante da dependência de uma força de trabalho pendular dado o transbordamento da demanda por capital humano.

[1] Deriva de *commuting,* é o termo em inglês utilizado para nomear o cidadão que viaja regularmente entre casa e trabalho, ou seja, aquele que realiza o movimento pendular.

Quanto às cidades de origem do deslocamento, a pendularidade pode ser benéfica, uma vez que oferece oportunidades de emprego e mobilidade econômica para seus residentes além de reduzir o diferencial de salários e de desemprego entre regiões rurais e urbanas, elevando a renda de toda a região (Hazans, 2003). Isso pode reduzir o desemprego local e melhorar a qualidade de vida dos trabalhadores que conseguem empregos em cidades-polo. No entanto, a cidade fornecedora de *commuters* não consegue absorver a mão de obra local pois eles buscam oportunidades em outras cidades de acordo com a melhor remuneração possível ao capital humano (Borjas, 2012) e o efeito observado, na maioria delas, é a concentração de atividades econômicas menos especializadas com menos serviços disponíveis, tornando-as municípios-dormitório com baixa arrecadação (Moura, 2010).

Portanto, os efeitos líquidos da pendularidade nas economias locais podem ser ambíguos. É crucial identificar os fatores determinantes desse fenômeno para orientar políticas públicas de planejamento urbano e desenvolvimento regional. Essas políticas têm o papel de mitigar os impactos negativos da pendularidade e, se necessário, reverter os fluxos intensivos de trabalhadores a fim de promover um equilíbrio econômico sustentável.

Existe um conjunto complexo de fatores que influencia a escolha das pessoas em se deslocar diariamente para suas atividades profissionais ou educacionais. Tais fatores podem ser analisados sob diferentes níveis, considerando aspectos individuais, familiares, sociais, econômicos e espaciais (Stamm, 2013). Neste cenário, sob o ponto de vista do fluxo de pendulares entre municípios se destacam na literatura questões como distância, habitação, renda, mercado de trabalho e desenvolvimento econômico (Brito; Ramalho, 2018; Brito; Silva, 2020; Chen; Voigt; Fu, 2021; Colla, 2018). Nesta pesquisa considera-se o efeito de duas dessas dimensões sobre os deslocamentos dos trabalhadores: uma espacial e uma econômica.

A consideração do espaço é essencial ao analisar questões relacionadas a deslocamento, como migração e pendularidade, devido aos custos subjacentes. Da mesma forma, ao examinar o trabalho em determinadas regiões, é crucial incorporar fatores econômicos, como o nível de renda e o desenvolvimento daquela área. Essa abordagem motiva a compreensão da interseção entre essas duas dimensões - espaço geográfico e oportunidades econômicas - na explicação de fenômenos que envolvem o deslocamento, especialmente com o objetivo laboral.

Em vez de considerar a pendularidade apenas do ponto de vista individual, analisa-se a interação das escolhas individuais em um contexto intermunicipal. Isso permite observar uma rede completa de cidades, onde a entrada ou a saída de trabalhadores ocorre em função da distância entre elas e do seu nível de desenvolvimento econômico.

Portanto, é essencial compreender se a presença de um município com um determinado nível de desenvolvimento, situado a uma distância específica de outro município de interesse, desempenha um papel relevante no deslocamento de trabalhadores pendulares entre esses dois pontos de análise. Além disso, é fundamental investigar como tal fator espacial influencia o desenvolvimento das regiões e afeta a pendularidade.

Nesse contexto, ao formular o modelo para analisar o fenômeno entre duas cidades quaisquer, a distância entre essas cidades será empregada como uma métrica representativa do custo de deslocamento intermunicipal. Além disso, a avaliação da importância do desenvolvimento econômico dos municípios vizinhos capturará a dimensão espacial do desenvolvimento regional. Vale ressaltar que a dimensão do Índice Firjan de Desenvolvimento Municipal (IFDM) que representa características municipais de mercado de trabalho será adotada como uma variável representativa do status econômico de cada localidade.

A escolha pela distância entre os pares de municípios representa a necessidade de incorporar os custos de deslocamento enfrentados pelos *commuters*. Sabendo que essa movimentação pode ser feita de várias maneiras diferentes e que os custos de cada trabalhador e entre cada município podem variar, optou-se pela utilização da distância em quilômetros entre os pares de cidades como uma *proxy* para estimar esse efeito.

A decisão de utilizar o Índice Firjan, em particular o seu componente "Emprego e Renda" (IFDM-ER), foi motivada pela capacidade desse índice capturar de forma mais precisa os resultados econômicos relacionados ao mercado de trabalho[2].

Para analisar os condicionantes desse fenômeno e como as variáveis se relacionam há uma variedade de ferramentas estatísticas como, por exemplo, modelos gravitacionais (Masucci *et al*, 2013), modelos de radiação (Varga; Tóth; Néda, 2020), modelos de *machine learning* (ML) (Spadon *et al*, 2019; Chen; Voigt; Fu, 2019), técnicas de econometria espacial (Brito; Silva, 2020, Li; Niu, 2022) e modelos *probit* (Colla, 2018). Porém, de acordo com Stefanouli *et al.* (2017), os estudos nesse tema têm utilizado em sua maioria o modelo gravitacional.

Duas modelagens serão empregadas nesta pesquisa: um modelo gravitacional e um modelo de aprendizado de máquina. A utilização destes métodos é motivada pela necessidade de estimativas precisas, bem como pela intenção de identificar e modelar possíveis efeitos espaciais nas dimensões relevantes relacionadas à pendularidade. Além disso, é fundamental avaliar qual desses modelos demonstra melhor desempenho na previsão dos deslocamentos,

[2] Essa dimensão é mensurada por meio das seguintes variáveis: emprego formal, formalidade do mercado de trabalho, geração de renda, massa salarial do mercado formal e Índice Gini de desigualdade da renda do trabalho formal.

considerando a interação das duas dimensões mencionadas. Vale destacar que a equação de gravitação é frequentemente empregada na literatura para abordar esses aspectos, enquanto ferramentas de *machine learning*, como o *XGBoost*, representam o estado da arte na solução de diversos problemas estatísticos, por exemplo, como uma ferramenta de diagnóstico precoce de *Alzheimer* (Kadem; Noseworthy; Doyle, 2023) ou de detecção de fraude financeira (Nti; Somanathan, 2022), com a vantagem de consumir menos recursos do que as alternativas disponíveis (Chen; Guestrin, 2016).

A comparação dos modelos é feita através de algumas métricas aplicadas após a especificação e a estimação, como a capacidade preditiva, a raiz do erro quadrático médio e outras métricas presentes em diversos estudos sobre pendularidade que mensuram o quão próximo o deslocamento pendular estimado se encontra do valor observado e que serão descritas com mais detalhes no capítulo dos procedimentos metodológicos.

Essa comparação só pode ser feita após a especificação correta dos modelos. A modelagem gravitacional é estimada por uma técnica de pseudo-máxima verossimilhança assumindo distribuição *Poisson*, o que a torna consistente na presença de heterocedasticidade e adequada para lidar com a existência de observações-zero na variável dependente. Também conta com a inserção de variável explicativa defasada espacialmente, tornando-se um modelo *Poisson* do tipo regressivo cruzado espacial (SLX[3]). Já o segundo é o *XGBoost*, que incorpora as mesmas variáveis espaciais e se utiliza da ferramenta SHAP[4] para interpretar o efeito das variáveis explicativas sobre a dependente. Dessa maneira, além de comparar dois tipos de métodos, avalia-se se a consideração de efeitos espaciais em ambos gera algum tipo de diferencial na pendularidade.

A escolha pela análise dos municípios da Região Metropolitana (RM) de Londrina se deu por alguns motivos específicos: i) a região possui uma característica desejável, sendo um *cluster* de desenvolvimento não tão extenso, o que torna a estimação do modelo gravitacional um pouco melhor, como sugerem Masucci et al (2013) ao afirmarem que o modelo gravitacional superestima a pendularidade quando se analisa distâncias maiores que 100 quilômetros; ii) A RM de Londrina possui o maior PIB (Produto Interno Bruto) e a maior população dentre as cidades do interior do Paraná. A região representa uma fatia considerável da atividade econômica estadual e carece de pesquisa sobre pendularidade; iii) Municípios

[3] Spatial Lag of X.
[4] *SHapley Additive exPlanations*: Ferramenta que traz interpretabilidade aos modelos de ML. Mais detalhes na seção 3.3.

paranaenses de porte médio, de acordo com Stamm e Stadutto (2008), têm apresentado tendência de crescimento de deslocamentos pendulares.

Dadas estas considerações e a dinâmica do deslocamento intermunicipal enfrentado pelos trabalhadores, como o espaço geográfico e o desenvolvimento econômico podem influenciar a intensidade dos movimentos pendulares entre diferentes municípios? Portanto, o objetivo principal consiste em analisar a importância da distância espacial e do nível de desenvolvimento econômico no movimento pendular de trabalhadores considerando pares de cidades da RM de Londrina.

Especificamente, busca-se:

i) Analisar a distribuição espacial do movimento dos *commuters* entre os municípios da RM de Londrina;

ii) Modelar os deslocamentos de trabalhadores pendulares entre os municípios selecionados;

iii) Identificar como os fatores econômicos e a distância geográfica influenciam a decisão de deslocamento dos *commuters*;

iv) Comparar o desempenho das estimativas com e sem controle de efeitos espaciais dos modelos gravitacional e *XGBoost.*

A relevância desta pesquisa se encontra na compreensão aprofundada dos movimentos pendulares, um fenômeno crescente que reflete e impacta a estrutura socioeconômica das regiões urbanas e metropolitanas. A importância do estudo da pendularidade reside na dependência da distribuição espacial das atividades econômicas e dos locais de moradia ligados através da infraestrutura de transporte (Sohn, 2005), o que afeta toda a cadeia produtiva de determinada região. O entendimento desses padrões é crucial para o planejamento urbano e regional, influenciando diretamente políticas públicas em áreas como transporte, habitação, emprego e desenvolvimento econômico, estimativas incorretas podem apontar um nível de pendularidade diferente do observado.

Além disso, a comparação entre os modelos, dada a sofisticação proposta na estimação da equação gravitacional, pode constituir uma evidência de que este ainda é útil frente às ferramentas de aprendizado de máquina para modelar a pendularidade. A comprovação de que, mesmo assim, os métodos de aprendizado de máquina seriam superiores, também pode constituir um resultado relevante.

Adicionalmente, uma evidência da importância desta pesquisa é a recente conclusão do Plano de Desenvolvimento Urbano Integrado (PDUI) da RM de Londrina (SEDU, 2022), que, dentre outras propostas, elabora um planejamento para melhorar a infraestrutura de transporte,

sugerindo estratégias integradas que otimizem os movimentos pendulares, típicos das RM's em geral (SEDU, 2022). Neste contexto, o PDUI, como uma ferramenta de fomento ao desenvolvimento instituído pelo Estatuto da Metrópole (SEDU, 2022), constitui uma prova de que a pendularidade é um tema atualmente debatido pelos gestores públicos.

Ademais, esta pesquisa introduz uma inovação ao aplicar efeitos espaciais nos regressores (modelagem do tipo SLX) de ambas as ferramentas estatísticas para explicar pendularidade.

Esta pesquisa está estruturada em cinco capítulos. Esta, a introdução, foi o primeiro deles. No próximo capítulo são apresentadas as reflexões a respeito da pendularidade e como esse fenômeno se interliga com aspectos sociais, econômicos e espaciais. Em seguida, é apresentado o diversificado conjunto de ferramentas estatísticas utilizadas na literatura, bem como os achados mais significativos. O terceiro capítulo apresenta os procedimentos metodológicos da pesquisa, findando com a análise dos resultados. Por fim, tem-se as considerações finais.

# 2. A PENDULARIDADE NA LITERATURA ECONÔMICA

Neste capítulo são apresentadas discussões sobre pendularidade e a maneira na qual o fenômeno se relaciona com a sociedade, com a economia e com o espaço. Logo após é exposto o variado ferramental estatístico empregado na pesquisa sobre pendularidade e os resultados mais relevantes.

## 2.1 O conceito de movimento pendular

Por muito tempo o estudo da migração era a abordagem mais frequente da mobilidade dos trabalhadores, porém, com o processo de transformação dos espaços através do espraiamento da urbanização e do aumento da mobilidade, os deslocamentos diários entre regiões (residência-trabalho-residência) ganharam relevância (Tavares; Monteiro, 2019).

Polyzos, Tsiotas e Minetos (2013) definem a pendularidade como o trajeto diário entre duas localidades com o objetivo de exercer atividade laboral e como um fenômeno que sofre influência de fatores econômicos, sociais, políticos e geográficos. Os movimentos pendulares constituem uma grande parte dos deslocamentos em todo o mundo e demandam muitos investimentos em transporte e infraestrutura que são iniciados com a finalidade de atender à demanda dos *commuters* (Haas; Osland, 2013).

As teorias clássicas de localização espacial, representadas principalmente por Lösch (1954), Von Thünen (1966) e Christaller (1966), desempenham um papel fundamental na análise da organização espacial das atividades econômicas e constituem uma base sólida para a compreensão dos fenômenos relacionados aos movimentos pendulares.

Lösch (1954) desenvolveu a Teoria das Áreas Centrais, que enfoca a organização das atividades econômicas em uma área geográfica considerando fatores como distância e custos de transporte. Sua abordagem matemática permitiu a identificação de áreas centrais que servem como polos de atração econômica, onde as atividades tendem a se concentrar devido à maior acessibilidade e menor custo de transporte. Por outro lado, à medida que se afasta dessas áreas centrais, as atividades econômicas se dispersam devido aos custos de transporte mais elevados.

A Teoria da Localização de Von Thünen (1966) se concentra na escolha do uso da terra em função da distância ao mercado. Ele estabeleceu o conceito de "anéis concêntricos", nos quais diferentes tipos de cultivos ou atividades econômicas são organizados de acordo com sua relação com o mercado. À medida que se afasta do mercado, os produtos ou atividades que requerem menor tempo e custo de transporte têm maior probabilidade de serem encontrados.

A Teoria dos Lugares Centrais de Christaller (1966) é outra contribuição relevante para a compreensão da organização espacial das atividades econômicas. Essa teoria explora como os lugares centrais, como cidades ou centros urbanos, desempenham papéis de centralização econômica, fornecendo bens e serviços para áreas circundantes. A organização hierárquica dos lugares centrais, com centros menores servindo áreas mais próximas e centros maiores atendendo áreas mais distantes, é fundamental para entender como as atividades econômicas são distribuídas no espaço em função da acessibilidade e da demanda.

Essas teorias clássicas, embora tenham sido desenvolvidas há décadas, continuam a ser relevantes para a análise dos movimentos pendulares, pois oferecem uma estrutura conceitual sólida para entender como as atividades econômicas se organizam no espaço em função de fatores como distância, custos de transporte, custo do solo e acessibilidade. Isso é particularmente importante para compreender como a localização das oportunidades de trabalho e a distribuição geográfica das atividades influenciam as decisões de deslocamento dos trabalhadores.

À medida em que os municípios crescem e se desenvolvem, vão se aproximando uns dos outros não somente com o aumento do espaço físico das cidades, mas com as suas economias cada vez mais simbióticas e com as suas infraestruturas de transportes cada vez mais interligadas, fazendo com que se observe mais pendulares entre estes municípios nas duas direções mesmo quando são polos centrais das suas respectivas regiões de desenvolvimento, o que Stamm e Staduto (2008) chamam de pendularidade cruzada. Em outras palavras, se as cidades apresentam um custo de deslocamento suficientemente baixo e oferecem condições de trabalho adequadas, a pendularidade ocorrerá entre elas. A formação destas redes urbanas cada vez mais conectadas e a facilidade crescente de mobilidade estão evidenciadas na pesquisa de Stamm e Staduto (2008) através da observação da ocorrência da pendularidade de maneira bilateral, ou seja, ambos os municípios estudados são tanto receptores quanto fornecedores de pendulares.

É possível abordar as decisões de mobilidade dos pendulares tanto a nível individual quanto familiar. Contudo, a agregação dessas decisões revela uma significativa relevância, uma vez que elucida a complexa interação entre municípios de diferentes níveis de desenvolvimento, como discutido por Brito e Ramalho (2019). Isso se traduz, por exemplo, na dinâmica onde municípios mais desenvolvidos frequentemente atuam como cidades receptoras, atraindo uma quantidade substancial de pendulares originários de cidades circunvizinhas, tornando-se polos econômicos. Em contrapartida, municípios menos centrais tendem a apresentar uma predominância de saída de *commuters* em relação à entrada.

O fenômeno de aglomeração urbana suscita preocupações abrangentes em diversos segmentos da sociedade. O aumento da densidade populacional nos espaços urbanos demanda um incremento substancial em infraestrutura, habitação, oportunidades de emprego e serviços, visando a manutenção da qualidade de vida dos habitantes (Brito; Ramalho, 2019).

Notadamente, tal expansão populacional costuma estar associada a um encarecimento dos custos habitacionais e ao agravamento do congestionamento viário, conforme discutido por Brito e Ramalho (2019). Este cenário se torna particularmente acentuado no contexto da pendularidade, uma vez que, como enfatizado por Chen, Voigt e Fu (2021), a movimentação intermunicipal de trabalhadores frequentemente desencadeia gargalos na infraestrutura de transporte regional. Nesse sentido, torna-se imprescindível a incorporação dessas considerações nos processos de planejamento de políticas públicas voltadas para o setor de transporte.

Tavares e Monteiro (2019) também apontam a relação da pendularidade com a estrutura urbana, enfatizando a importância em mensurá-la para a implementação de políticas públicas voltadas ao trabalho, à mobilidade e à educação:

> Os movimentos pendulares fazem parte dos processos de mobilidade espacial e o enfoque a partir desse conceito contribui para ressaltar a dimensão social e espacial do fenômeno, que está bastante ligado às potencialidades e aos constrangimentos da estrutura urbana – acesso à moradia, trabalho, educação, lazer – e aos aspectos socioculturais das pessoas e dos lugares de origem e destino (Tavares; Monteiro, 2019, p. 55).

Os pesquisadores também destacam que os movimentos pendulares tiveram um aumento substancial no Brasil, notadamente cobrindo distâncias relativamente curtas, devido aos padrões diários de deslocamento. Este incremento não se limitou às áreas periféricas das regiões metropolitanas, estendendo-se também para outras localidades no interior dos estados (Tavares; Monteiro; 2019). Essa observação sobre a curta distância de deslocamento é corroborada pela pesquisa conduzida por Delgado, Moura e Cintra (2013), que constataram que, em regiões de menor porte, como a Região Metropolitana de Londrina, as distâncias percorridas frequentemente não ultrapassam os 50 quilômetros. Estas distâncias mais curtas estão associadas ao deslocamento da demanda de mão de obra de municípios com atividades econômicas específicas, o que se diferencia do fenômeno de pendularidade nas regiões metropolitanas mais destacadas, como a de Curitiba, onde os deslocamentos são predominantemente moldados pela dinâmica metropolitana em si.

No caso da pendularidade paranaense, os deslocamentos aumentaram na primeira década do século XXI, porém, com um decréscimo relativo das três maiores cidades-polo (Curitiba, Londrina e Maringá) em relação ao restante dos municípios, o que indica que o

fenômeno cresceu de maneira mais acentuada nas cidades de menor porte (Delgado; Moura; Cintra, 2013).

A diferença óbvia entre os fenômenos da pendularidade e da migração, no âmbito intermunicipal, é a decisão do trabalhador em residir ou não onde trabalha e ele o faz ponderando desde questões de preferências individuais ou familiares, de ganho de remuneração até o aspecto habitacional, este último se tornando mais atrativo à medida em que se torna mais acessível (Haas; Osland, 2013) e, de acordo com Brito e Ramalho (2019), a decisão de moradia do trabalhador é influenciada pela localização dos empregos que remuneram mais e do custo de habitação:

> Enquanto a concentração espacial de empregos atua como uma força inibidora da demanda residencial em localidades mais distantes, as economias geradas por aglomerações urbanas favorecem o pagamento de maiores remunerações, as quais tendem a compensar os custos de residência em locais mais distantes do eixo dinâmico da economia (menores preços de habitação). Tal compensação pode ser limitada na presença de assimetria de informação e elevados custos de mobilidade (Brito; Ramalho, p. 96).

Ao investigar o fenômeno de pendularidade entre as cidades de Cascavel e Toledo, duas localidades de porte médio situadas no interior do estado do Paraná, que compartilham características semelhantes à região de interesse desta pesquisa, Stamm (2005) concluiu que fatores de natureza econômica desempenham um papel motivador no deslocamento dos trabalhadores, enquanto considerações de ordem social influenciam sua decisão de não migrar para o município onde trabalham. Em outras palavras, ambos os aspectos parecem exercer influências distintas sobre a tomada de decisão do trabalhador. Reforçando essa perspectiva, Stamm e Staduto (2008) acrescentam que a escassez de oportunidades de emprego no município de residência, aliada a uma oferta salarial superior na cidade vizinha, constitui um estímulo para que o trabalhador opte pelo movimento pendular.

Essas conclusões dos autores em relação aos fatores socioeconômicos que influenciam a escolha entre migrar ou realizar deslocamentos diários para o trabalho endossam a necessidade de desagregar o componente econômico dos demais fatores na variável explicativa. Tal abordagem permite isolar o impacto da pendularidade em oposição ao efeito da migração durante o processo de modelagem.

Conforme observado por Haas e Osland (2013), várias pesquisas têm destacado que a pendularidade resulta da crescente especialização geográfica das atividades laborais e está sujeita às influências das estruturas de transporte e infraestrutura, dinâmicas do mercado de trabalho e demanda da sociedade. Isso enfatiza a importância de uma análise abrangente que considere não apenas os mercados de habitação e emprego, mas também os custos associados

à mobilidade. Dessa maneira, a próxima seção apresenta investigações sobre como diversos aspectos, tais quais os supracitados, afetam os deslocamentos pendulares.

## 2.2 As análises empíricas sobre os movimentos pendulares

O deslocamento pendular é um fenômeno amplamente investigado em diversos países, utilizando uma variedade de abordagens metodológicas. É notável a recorrência de resultados semelhantes em múltiplos estudos que se dedicam a compreender a pendularidade. Nesta seção, documenta-se a relevância de variáveis-chave, tais como a distância entre os municípios, os custos de transporte, o preço da habitação, a qualidade de vida, a disponibilidade de amenidades e infraestrutura, os níveis salariais e outros fatores pertinentes, que têm sido objeto de estudo e análise nesse contexto.

Renkow e Hoover (2000) utilizam dados censitários do estado da Carolina do Norte nos anos de 1980 e 1990 para estimar como o número de pendulares que se deslocam de um distrito a outro é afetado pela diferença de salários entre as regiões, a diferença do custo de habitação, a distância entre os distritos e a migração para o distrito de origem no período anterior. Os autores produziram um modelo gravitacional para essa estimação e seus resultados são: i) A pendularidade é fortemente influenciada pelo custo de deslocamento e as diferenças de salários. Custos de habitação entre os distritos também são determinantes importantes; ii) A dinâmica de pendularidade entre regiões rurais-urbanas no estado ocorreram devido às mudanças nas preferências das famílias por causa dos custos cada vez menores de deslocamento, da maior mobilidade das pessoas e das firmas e das externalidades negativas crescentes resultantes da vida em áreas urbanas, como criminalidade e congestionamentos. Isto levou a uma desconcentração da população e consequente aumento da pendularidade.

No estudo realizado por Stamm e Staduto (2008), foi minuciosamente examinado o fenômeno de migração pendular entre os municípios de Cascavel e Toledo, localizados na região oeste do Paraná. A pesquisa, que envolveu a coleta de dados por meio de entrevistas com uma amostra de usuários de transporte metropolitano, revelou conclusões significativas.

Primeiramente, ficou evidente que a escassez de oportunidades de emprego em suas cidades de residência e a atrativa oferta de salários mais substanciais na cidade vizinha são os principais impulsionadores do movimento pendular. Isso aponta para a necessidade de trabalhadores deslocarem-se diariamente para atender às suas demandas profissionais.

Além disso, a pesquisa desvendou dois outros achados notáveis. Em cidades de médio porte, como as analisadas no estudo, os trabalhadores optam por esse deslocamento diário em

vez de uma mudança definitiva, influenciados por fatores pessoais, como preferências individuais e considerações familiares. Essa escolha de mobilidade temporária é uma característica marcante do fenômeno.

Outro aspecto intrigante é a existência de uma pendularidade cruzada. Isso significa que há uma circulação de trabalhadores entre ambos os municípios, com pessoas partindo de Cascavel para trabalhar em Toledo e vice-versa. Esse padrão de mobilidade é viabilizado pela formação de uma rede urbana de desenvolvimento econômico que aproxima suficientemente os dois polos regionais, tornando o custo e o tempo de transporte aceitáveis. Isso demonstra a interconexão entre as cidades e a importância da mobilidade na busca de oportunidades de emprego.

McArthur *et al.* (2011) fazem um teste de "transferibilidade" comparando os parâmetros estimados do seu modelo gravitacional de pendularidade de uma região sendo utilizados para outra região e os dados reais desta para identificar se os parâmetros daquela conseguem explicar a segunda área. Para isso, foram realizados testes de "transferibilidade" em três regiões da Noruega, seguidos de uma simulação de *Monte Carlo* para testar o desempenho desses testes e os resultados mais relevantes são: A "transferibilidade" dos parâmetros é bem-sucedida somente quando se trata de regiões com características socioeconômicas homogêneas (é o caso dos municípios noruegueses) e os parâmetros estimados das regiões são próximos. Tal procedimento é complexo e parece não funcionar em agrupamentos de cidades com características heterogêneas. Pode-se associar este último resultado à pesquisa de Brito e Ramalho (2019) quando afirmam que as dinâmicas diferenciadas das variadas regiões metropolitanas brasileiras dificultam as investigações para o país como um todo.

Delgado, Moura e Cintra (2013) identificam o perfil dos pendulares com a finalidade exclusiva de trabalho nas regiões metropolitanas de Curitiba, Londrina e Maringá através dos dados do Censo Demográfico 2010: níveis de renda, setor de atividade, tempo de deslocamento, entre outros. A escolha dos autores pelas regiões supracitadas se deu após descobrirem que os deslocamentos de pendulares paranaenses se concentram em sua maioria nessas aglomerações urbanas. Três conclusões do estudo que merecem destaque são a redução da concentração de movimentos pendulares nessa região, indicando uma intensificação deste fenômeno em cidades de menor porte no estado; os *commuters* alcançam geralmente distâncias até 50 quilômetros e; nas três regiões metropolitanas do estudo a taxa de mobilidade pendular aumenta conforme o rendimento do trabalho aumenta.

Outra abordagem quantitativa é o modelo de regressão ordinal utilizado por Polyzos, Tsiotas e Minetos (2013) para examinar como a distância percorrida pelos trabalhadores nas

regiões de Tessália e de *Sterea Hellas*, na Grécia, responde às variáveis relacionadas a aspectos populacionais e econômicos dentre as cidades de origem e de destino. Alguns dos seus resultados que são relevantes para esta pesquisa são: i) municípios de origem mais populosos cedem menos pendulares; ii) municípios de destino mais populosos atraem mais pendulares; iii) uma população mais educada no município de origem está mais disposta a trabalhar fora para ter maiores rendimentos; iv) cidades com maior nível de bem-estar (os autores o mensuram em termos de PIB per capita) atraem pendulares de fora e retém seus trabalhadores originais; v) uma cidade de origem com alto nível de desemprego causa saída de trabalhadores pendulares.

A pendularidade sob a ótica do tempo de deslocamento dos trabalhadores e as variáveis selecionadas que o afetam no contexto da Região Metropolitana de Recife é analisada por Barbosa (2015). Para fazê-la a autora utiliza dados do Censo 2010 do IBGE e os aplica, dentre outras tentativas de ferramentas da econometria espacial, na modelagem que apresenta autocorrelação espacial no termo do erro (SEM). As evidências apontam que a distância até a capital estadual e a densidade populacional do município analisado influenciam positivamente o tempo de deslocamento enquanto para a proporção de imóveis alugados por município e a renda domiciliar per capita tal relação é negativa.

O modelo gravitacional, o modelo de radiação e um modelo de radiação que incorpora o tempo de deslocamento como critério de busca dos pendulares foram apresentados por Stefanouli *et al.* (2017) para testar o quanto a população e a distância entre os municípios explicam o volume de pendulares na Grécia bem como identificar qual desses modelos é o melhor ajustado e qual possui melhor capacidade preditiva. Um resultado interessante encontrado nos modelos e que converge com a literatura é que a distância é um obstáculo para a pendularidade, que os aspectos de desenvolvimento econômico são fatores atrativos de pendulares e que há uma correlação entre a mobilidade dos trabalhadores e o setor terciário do município de destino.

Varga, Tóth e Néda (2017) conduziram uma análise da distribuição dos movimentos de pendulares entre as regiões da Hungria no ano de 2011. A pesquisa envolveu a comparação entre um modelo de radiação convencional e um modelo de radiação que incorpora a capacidade dos trabalhadores de fazer escolhas seletivas, considerando apenas uma fração dos postos de trabalho disponíveis. Os autores também introduziram um modelo de radiação que otimiza os custos de transporte, o qual demonstrou ser mais eficaz na modelagem dos dados em âmbito nacional.

O modelo proposto pelos autores parte do pressuposto de que os indivíduos optam por trabalhar em outras regiões somente se essa decisão resultar em um ganho líquido em seus

salários, após a subtração dos custos de transporte associados à distância percorrida. Assim, a variável de custos de transporte desempenha um papel de relevância ainda maior dentro desse modelo, uma vez que é determinante na tomada de decisão dos trabalhadores em relação aos seus deslocamentos pendulares.

Ciommi, Chelli e Salvati (2018) analisam os padrões de crescimento urbano e os padrões de pendularidade na região metropolitana de Atenas para identificar subcentros de emprego nos anos de 1991, 2001 e 2011. Os autores utilizam estatísticas descritivas e outras técnicas como análise multivariada exploratória, inferência não paramétrica e modelos de regressão. Os resultados mais relevantes indicam uma transição na estrutura espacial de Atenas em direção à uma descentralização através de subcentros de emprego com distritos comerciais que atraem trabalhadores de áreas vizinhas, elevando a pendularidade.

Em um estudo para analisar a mobilidade dentro de uma provável megalópole no centro do México, Escamilla *et al* (2018) utilizam um modelo gravitacional para identificar que a população da área metropolitana de Querétaro possui a menor mobilidade dentre as áreas estudadas devido ao fato de que a região tem os maiores salários e a melhor infraestrutura de transportes enquanto a população da região de Tianguistenco possui alta mobilidade e baixa força de atração de pendulares uma vez que essa área apresenta baixos salários e indicadores ruins de educação e saúde.

Brito e Ramalho (2019) investigam a influência dos IDH's municipais e da distância entre os pares de municípios da Região Metropolitana de Recife nos fluxos pendulares (trabalho e estudo) intrametropolitanos através de um modelo gravitacional com efeitos fixos, robusto à presença de questões multilaterais não observadas e associadas aos aspectos estudados. Seus resultados mais relevantes são, em primeiro lugar, que uma distância cada vez maior entre duas cidades reduz o fluxo de pendulares entre elas, ou seja, é uma força repulsiva da pendularidade e, em segundo lugar, quão maior for o nível de desenvolvimento humano do município de destino, maior é a sua força atrativa de pendulares e em relação à cidade de residência, um IDH elevado reduz os fluxos, ou seja, os cidadãos tendem a trabalhar e estudar no município em que moram.

Spadon *et al.* (2019) utilizam técnicas de aprendizado de máquina e 22 indicadores como distância, PIB, área, acidentes de trânsito, renda, IDH e vários outros aspectos socioeconômicos relacionados à urbanização para prever o fluxo de pendulares em todos os municípios brasileiros no ano de 2010. Os resultados mais importantes encontrados pelos autores são de que a modelagem possui boa capacidade preditiva e que os quatro indicadores

mais importantes para prever a pendularidade no Brasil são (em ordem de importância): Distância, PIB, tamanho da cidade e o número de acidentes de trânsito.

Brito e Silva (2020) Testam indicadores socioeconômicos de diferentes instituições associados às taxas de fluxos de pendulares em todos os municípios brasileiros no ano de 2010 sob a hipótese de que sejam autocorrelacionados espacialmente. Os autores detectaram autocorrelação espacial e prosseguiram com um modelo de erro espacial (SEM) estimado por máxima verossimilhança, identificando a importância do IDHM-renda, do índice *Theil-L* de desigualdade nos rendimentos do trabalho, da distância até a capital estadual e do índice IFDM-saúde na determinação da taxa dos deslocamentos pendulares. A única variável não significativa foi o índice IFDM-educação e isto se deve, segundo os autores, à vários outros aspectos municipais embutidos no indicador, o que dificultaria a análise.

A mobilidade dos trabalhadores indianos residentes das áreas rurais em direção aos postos de trabalho nas áreas urbanas é objeto de análise de Bhatt, Chandrasekhar e Sharma (2020), para o período de 2018 a 2019. De acordo com os autores, o modelo multinominal produzido indica que um nível de desemprego elevado na região rural atua como força de repulsão dos trabalhadores enquanto um baixo nível de desemprego na região urbana atua como força de atração dos pendulares que residem na região rural. Vale ressaltar ainda que a urbanização possui correlação positiva com a pendularidade, enquanto essa relação é negativa quando se trata de regiões menos urbanizadas.

Abouelhamd (2021) examina a relação entre a estrutura espacial urbana e os padrões de pendularidade através de uma revisão de literatura com uma abordagem de análise dedutiva. O autor conduz uma revisão minuciosa das contribuições empíricas e teóricas anteriores para identificar lacunas na pesquisa. Seus resultados destacam a relevância de variáveis como o tamanho da cidade, a centralidade, a concentração de atividades, a densidade urbana e a distribuição do uso da terra para entender os padrões da pendularidade.

Chen, Voigt e Fu (2021) utilizaram dados de movimentos pendulares intermunicipais entre 1994 e 2018 da Alemanha para analisar a influência do PIB, da habitação e do mercado de trabalho na decisão de trabalhar fora do município de residência através de uma regressão linear e concluíram que o PIB, a renda média, o custo de habitação entre moradia e trabalho e a distância da área metropolitana são fatores que influenciam a decisão de trabalhar em outro município. Dessa forma os trabalhadores planejam seu locais de moradia e trabalho de acordo com diversos fatores, como a renda, a distância casa-trabalho, a localização, o tipo de trabalho e outros fatores como idade, gênero, despesas domésticas, composição do domicílio e nível de bem-estar. Os autores também utilizam algoritmos de *machine learning* para fazer previsão do

número de pendulares alemães com certa precisão sugerindo o uso do modelo para outras regiões.

Liang *et al* (2021) utilizam dados de transporte público em Pequim coletados ao longo de um mês juntamente com uma pesquisa de preferência revelada (RP) para prever o destino de *commuters* usuários do transporte público local. O estudo identifica a cadeia de viagem de 563 usuários por meio de análise de associação de números de cartões inteligentes, dados de transações e de rede, criando dessa maneira a amostra analisada para o algoritmo *XGBoost*. Os autores desenvolvem um modelo de previsão do próximo destino dos *commuters* baseado em características que consideravam determinantes para influenciar um destino de viagem, encontrando alta precisão do modelo. Como corolário, Liang *et al* (2021) apontam que os fatores que afetam os destinos de viagem dos passageiros do transporte público incluem a distância da viagem, o tempo de viagem, as localizações da residência e do trabalho, a idade, o gênero, o status de emprego, a renda e o propósito da viagem. Além disso, o estudo também identificou que a frequência com a qual os *commuters* viajam é um fator importante que influencia os destinos de viagem dos passageiros do transporte público

Li e Niu (2022) utilizam dados de sinal de telefone-móvel na região do delta do rio Yangtzé entre as cidades de Sucheu e Xangai para analisar os fluxos de pendulares. Algumas técnicas de econometria espacial foram empregadas, como o modelo de regressão geograficamente ponderada (GWR), o modelo de interações espacialmente ponderadas (SWIM) e o modelo de regressão negativa binomial (NB) para efeito de comparação e para identificar quais os determinantes significativos. Os resultados encontrados pelos autores mostram que o tempo de deslocamento, a oferta e o preço da habitação são algumas das variáveis que explicam o fenômeno. Também é importante destacar que o modelo de regressão local foi o mais ajustado e que melhor explicou os determinantes da pendularidade.

Li *et al.* (2022) afirmam que a duração e a distância do movimento pendular são sensíveis em diferentes intensidades aos padrões espaciais urbanos e aos aspectos sociais, dessa maneira, utilizam análise de autocorrelação espacial, modelo de defasagem espacial (SLM) e modelo SEM para comparar o impacto de variáveis (algumas variáveis que vale mencionar: distância média de deslocamento e migração, densidade de empregos, distância do centro de negócios, porcentagem de trabalhadores com nível superior, porcentagem de famílias com carro particular e porcentagem de trabalhadores que moram de aluguel ) no tempo e na distância do fenômeno baseados numa amostra populacional de Xangai em 2015.

Mulalic e Rouwendal (2023) analisaram o impacto da introdução da rede de metrô na RM de Copenhague através de pesquisas de viagens - entre 2002 e 2010 - e regressões lineares

para modelar a pesquisa. Os pesquisadores observaram mudanças significativas no tempo de viagem dos *commuters*, na velocidade e na escolha do modo de deslocamento realizado (parcial ou totalmente pelo metrô). Segundo os autores, o impacto foi significativo na pendularidade, havendo redução no tempo de deslocamento pendular, aumento no uso de transporte público e aumento dos fluxos pendulares.

Wiese (2023) investiga o atraso dos trens ao se dirigirem para a *Penn Station*, em Nova Iorque durante a manhã, e como isso afeta a experiência dos *commuters* que utilizam o serviço de transporte ferroviário. O autor aplicou dois modelos de ML, *XGBoost* e *Random Forest Regression*, para prever o atraso individual dos trens ao chegarem à estação, o que, segundo ele, causou um impacto significativo na experiência dos passageiros. As variáveis que afetam os *commuters*, ou o tempo gasto por eles, incluem a localização de onde o trem parte, o dia do mês, o dia da semana, o percurso do trem e se o horário é de pico ou não. Além disso, a localização foi identificada como a variável mais relacionada ao atraso na chegada do trem.

Segundo Zhao *et al* (2023), o modelo gravitacional descreve bem a pendularidade entre cidades mais próximas, mas possui problemas para descrever padrões de viagem interurbana em espaços maiores com dezenas de megacidades interagindo, abrigando mais de 100 milhões de pessoas. Os autores analisaram mais de 2 bilhões de viagens de mais de 40 milhões de usuários de celulares na Região Jing-Jin-Ji, que inclui a capital chinesa, durante a primeira semana de junho de 2018. Os autores desenvolveram modelos gravitacionais mais precisos para a distribuição espacial de viagens interurbanas relacionadas e não relacionadas ao trabalho e, dentre seus resultados, o PIB municipal aparenta ser uma força de atração dos *commuters* e a população aparenta ser uma força de atração dos indivíduos que se deslocam por qualquer outro motivo.

Na Tabela 1 tem-se um resumo de alguns dos trabalhos que analisaram os determinantes da pendularidade.

Tabela 1: Resumo de algumas evidências empíricas acerca dos determinantes da pendularidade

| Autores | Metodologia | Região | Periodo | Efeito na pendularidade |
|---|---|---|---|---|
| Renkow e Hoover (2000) | Modelo gravitacional. | EUA (Carolina do Norte) | 1980 e 1990 | Distância é uma força repulsiva. Diferenças de salários são forças repulsivas. |
| Polyzos, Tsiotas e Minetos (2013) | Modelo de regressão ordinal. | Grécia (Tessália e Sterea Hellas) | 2000 | Municípios de origem mais populosos cedem menos pendulares e municípios de destino mais populosos atraem mais pendulares.<br>Desemprego causa saída de pendulares.<br>Cidades grandes atraem pendulares das cidades menores nos seus entornos. |

| | | | | |
|---|---|---|---|---|
| | | | | População mais educada é mais propensa a trabalhar em cidades mais distantes para atingir maior remuneração. |
| Barbosa (2015) | Modelo SEM | Brasil (RM de Recife) | 2010 | Distância até a capital estadual e a densidade populacional do município possuem efeito positivo no tempo de *commuting*<br>Renda domiciliar per capita possui efeito negativo. |
| Stefanouli *et al.* (2017) | Modelos gravitacional e de radiação. | Grécia | 2011 | Distância é uma força repulsiva.<br>Desenvolvimento econômico é uma força atrativa. |
| Varga, Tóth e Neda (2017) | Modelo de radiação | Hungria | 2011 | Custo de transporte reduz a pendularidade |
| Ciommi, Chelli e Salvati (2018) | Estatísticas descritivas, análise multivariada exploratória, inferência não paramétrica e modelos de regressão. | Grécia (Atenas e Pireu) | 1991, 2001 e 2011 | A descentralização do espaço através da formação de subcentros de emprego elevou a pendularidade. |
| Escamilla *et al.* (2018) | Modelos gravitacional e de radiação. | México | 2008-2016 | Região com baixo salário e baixos índices de educação e saúde têm alta saída de pendulares e baixa entrada.<br>Região com maiores salários e maior infraestrutura de transporte tem baixa saída de pendulares. |
| Brito e Ramalho (2019) | Modelo gravitacional | Brasil (RM de Recife) | 2000 e 2010 | Distância é força repulsiva dos fluxos.<br>IDH elevado no município de origem reduz os fluxos.<br>IDH elevado no município de destino eleva os fluxos. |
| Spadon et al. (2019) | Modelos gravitacional, *XGBoost* e de radiação. | Brasil (todos os municípios) | 2010 | Distância entre municípios tem efeito negativo.<br>Municípios com alto PIB atraem pendulares.<br>Cidades maiores atraem pendulares. |
| Bhatt, Chandrasekhar e Sharma (2020) | Modelo Multinominal | Índia | 2018 a 2019 | Alto desemprego rural é uma força de repulsão;<br>Baixo desemprego urbano é força de atração;<br>Alta urbanização é força de atração. |
| Brito e Silva (2020) | Modelos SAR e SEM | Brasil (todos os municípios) | 2010 | IDHM-renda alto eleva os fluxos de entrada.<br>Distância possui efeito negativo nos fluxos.<br>Municípios com alta densidade populacional possuem maior fluxo de entrada. |
| Abouelhamd (2021) | Revisão de literatura com abordagem de análise dedutiva. | Cidades entre 2 e 5 milhões de habitantes | 2012 a 2021 | Distância do polo econômico eleva o custo de pendularidade. |
| Chen, Voigt e Fu (2021) | Regressão linear, árvore de | Alemanha | 1994-2018 | PIB per capita municipal tem correlação positiva com entrada de pendulares e correlação negativa com a saída. |

| | | | | |
|---|---|---|---|---|
| | decisão e *Random forest* | | | Renda média da cidade de trabalho influencia fortemente a entrada de pendulares.<br>Distância tem relação negativa.<br>Preço do apartamento tem relação negativa. |
| Liang et al. (2021) | X*GBoost* | China (Pequim) | Abril de 2017 | Distância entre residência e trabalho reduz a pendularidade<br>Tempo de viagem reduz *commuting* |
| Li e Niu (2022) | GWR, SWIM e regressão binomial negativa. | China (Sucheu e Xangai) | Julho de 2021 | Tempo de deslocamento tem efeito negativo.<br>Setores de produção mais desenvolvidos na cidade de trabalho e setores menos desenvolvidos na cidade de residência têm efeito positivo. |
| Mulalic e Rouwendal (2023) | Modelos de regressão linear | RM de Copenhague (Dinamarca) | 2002 e 2010 | O Tempo e o custo de deslocamento provocados pela introdução do metrô elevaram a pendularidade na região. |
| Wiese (2023) | *Random Forest Regression* e *XGBoost* | Nova York (EUA) | 2022 | Local de partida do trem, percurso do trem e horário de pico são fatores que influenciam positivamente o tempo gasto pelos *commuters.* |
| Zhao et al. (2023) | Modelo Gravitacional | Região de Jing-Jin-Ji (China) | Junho de 2018 | PIB municipal atrai *commuters*<br>População do município atrai não-*commuters* |

Fonte: Resultado da Pesquisa.

Assim, a análise das pesquisas previamente elencadas evidencia a significância do estudo da pendularidade, que continua a ser objeto de investigação por acadêmicos em âmbito global, destacando-se a prevalência do modelo gravitacional para a representação desse fenômeno, bem como a emergência de abordagens baseadas em aprendizado de máquina com a mesma finalidade. Observa-se, ainda, a recorrência no uso de variáveis como a distância entre os municípios, características do mercado de trabalho e nível de renda local.

Notavelmente, a análise dos espaços urbanos se concentra predominantemente no nível municipal, em detrimento de subunidades como bairros, distritos ou estados, independentemente da localização geográfica ou período considerado. Estes fatores direcionam a seleção das variáveis, ferramentas estatísticas e escolha dos municípios específicos para esta pesquisa. Tendo em vista o foco atribuído ao modelo gravitacional e ao algoritmo *XGBoost*, conforme delineado na revisão de literatura anterior, a subsequente seção discorrerá sobre considerações práticas associadas a ambas as abordagens modelísticas.

## 2.3 Percepções empíricas sobre os modelos gravitacional e *XGBoost*

Da seção anterior é possível perceber a variedade de ferramentas disponíveis para modelar o fenômeno e a popularidade de modelos gravitacionais para este fim. Por outro lado, observa-se um número considerável de estudos recentes com o uso de ferramentas de aprendizado de máquina, sobretudo, modelos *XGBoost*. Ambos serão estimados e depois determinado o melhor modelo de acordo com as métricas estabelecidas no próximo capítulo. Neste contexto, os dois modelos serão submetidos a uma análise mais detalhada nesta seção, a fim de proporcionar uma maior compreensão de suas características e aplicabilidade.

O modelo gravitacional é uma ferramenta comumente utilizada na economia para analisar comércio internacional, fluxos migratórios e pendularidade. A forma tradicional da equação gravitacional, estimada por MQO (Mínimo Quadrado Ordinário) e aplicada na ciência econômica, foi introduzida por Tinbergen (1962) e possui a vantagem de analisar como variáveis específicas causam forças de atração ou de repulsão sobre uma variável dependente. Já o modelo *XGBoost*, proposto por Chen e Guestrin (2016), é uma técnica moderna de aprendizado de máquina que utiliza um processo de *boosting* de gradientes para otimizar modelos de árvores de decisão. O *XGBoost* melhora a construção do modelo e a velocidade de computação através de técnicas como o crescimento de árvore paralela, poda de árvores e regularização, que são parâmetros próprios, além da capacidade de lidar tanto com problemas de classificação quanto de regressão (Chen; Guestrin, 2016).

Morton, Piburn e Nagle (2018) comparam modelagem *XGBoost* com modelo gravitacional baseado em uma função exponencial de decaimento da distância e modelo de radiação, visando prever fluxos de pendulares na cidade de *Knoxville* (EUA) em 2010. Os autores concluem que, apesar de o seu modelo de ML conseguir recriar a rede de pendularidade de maneira muito mais satisfatória do que as outras ferramentas, uma parte dessas redes recriadas apresentou diferença discrepante entre os fluxos reais. Dessa maneira, o *XGBoost* deveria ser comparado com outros modelos mais complexos para checar se ainda assim performaria melhor.

*Simini et. al.* (2012) apontam algumas limitações do modelo gravitacional, como: a falta de um arcabouço teórico que força os pesquisadores a utilizarem funções de dissuasão para a distância entre as regiões e vários parâmetros para ajustar os dados empíricos; a sua baixa capacidade preditiva e; o fato de que, por ser determinístico, o modelo não leva em consideração as flutuações na quantidade de pendulares entre as regiões (Simini et al, 2012).

Sellner, Fischer e Koch (2010) também inferem que a sua especificação assume de maneira equivocada independência dos fluxos e, segundo LeSage e Pace (2008) os modelos gravitacionais tradicionais não consideram dependência espacial entre os fluxos de origem-destino, ou seja, na variável dependente. Morton, Piburn e Nagle (2018) criticam a simplicidade da especificação no sentido de que carrega poucas variáveis e limita a capacidade de capturar dinâmicas mais complexas que poderiam existir e serem estimadas através da inserção de outras variáveis, bem como é possível fazê-la com a modelagem de aprendizado de máquina.

Sellner, Fischer e Koch (2010) também apontam o excesso de simplicidade do modelo gravitacional, mas com o argumento de que assume independência dos fluxos de origem-destino quando na verdade pode haver uma dependência espacial. Os pesquisadores ainda afirmam que linearizar o modelo pode fazer com que acomode tal dependência na sua especificação.

O *XGBoost* bem como outras ferramentas de ML se apresentam em pesquisas recentes como soluções para alguns dos problemas supracitados. Por exemplo, na pesquisa de Wang *et al* (2021), os autores utilizaram *Random forests*, uma modelagem similar ao *XGBoost*, para explorar interações entre as características de uma bacia hidrográfica e os parâmetros de qualidade da água do lago *Taihu*, na China, durante o ano de 2017 e analisaram os valores SHAP do modelo, ou seja, o problema da quantidade reduzida de variáveis explicativas deixou de ser uma limitação para os pesquisadores.

A aplicação desta ferramenta revelou relação positiva inesperada entre a qualidade da água e a densidade populacional em áreas urbanas altamente povoadas, atribuída a diferenças no tratamento no tratamento de esgoto entre áreas urbanas e rurais Wang *et al* (2021).

Chen Voigt e Fu (2021) também fazem uso de técnicas de ML para fazer previsão do número de pendulares nos municípios alemães com árvores de decisão e *Random forests.* Os autores justificam o uso das ferramentas devido à alta precisão desses algoritmos de aprendizado e devido ao fato de que são capazes de modelar muitos tipos de bases de dados.

Osman *et al.* (2021) se propuseram a desenvolver um modelo preciso de previsão dos níveis de água subterrânea em áreas altamente povoadas em Selangor, Malásia. Os autores testaram três modelos de ML (*XGBoost*, rede neural artificial e regressão de vetores de suporte) com base em dados de chuva, temperatura e evaporação ao longo de 11 meses entre 2017 e 2018. Realizaram avaliação do desempenho dos modelos através de métricas como raiz do erro quadrático médio, erro absoluto médio e coeficiente de determinação em três cenários simulados. Os autores destacaram a superioridade do *XGBoost* em relação aos outros modelos após o primeiro apresentar os melhores valores das métricas nos cenários estudados.

Raihan *et al.* (2023) reuniram 24 características de 400 pacientes de uma amostra do repositório de ML da Universidade da Califórnia e aplicaram *XGBoost* para prever se um paciente possui doença renal crônica ou não. O modelo entregou, segundo os autores, altíssimos níveis de precisão e de acurácia e, através dos valores SHAP, foi possível determinar que os níveis de hemoglobina e de albumina no organismo são os que mais contribuem para a detecção da doença. Os pesquisadores ainda destacaram a importância da ferramenta SHAP para complementar o *XGBoost*, pois investigou, de maneira mais profunda, como algumas características específicas estavam contribuindo de maneira mais proeminente na previsão do modelo, interpretação que o *XGBoost* por si só não conseguiria entregar aos pesquisadores.

Observando as pesquisas dos autores que empregam a equação gravitacional e a modelagem *XGBoost* na análise da pendularidade, a comparação entre ambos os modelos parece ser algo pertinente.

Os resultados de Masucci *et al.* (2013) mostram que o seu modelo gravitacional, apesar de não descrever bem os dados de transporte coletivo ingleses, modelou de modo satisfatório os fluxos pendulares, bem como, apesar de não ser efetivo para deslocamentos intramunicipais, teve bom desempenho em se tratando de distâncias mais curtas ao nível intermunicipal, superestimando os fluxos somente para distâncias maiores que cem quilômetros. Outra afirmação relevante dos autores é que a especialização dos espaços entre áreas residenciais e áreas de negócios são aspectos determinantes para o padrão de fluxos pendulares e esses aspectos não são captados pelo modelo gravitacional tradicional.

Já Robinson e Dilkina (2018) compararam o *XGBoost* com modelos gravitacionais que incorporam funções de decaimento de distância, tanto exponenciais quanto de potência, para estudar a migração nos Estados Unidos e globalmente. Apesar da tentativa de controlar os efeitos espaciais, eles também chegaram à conclusão de que o modelo de *machine learning* é mais eficaz, ainda que por meio de uma metodologia diferente da proposta nesta pesquisa.

Spadon et al. (2019) conduziram uma análise comparativa entre a modelagem *XGBoost* e um modelo gravitacional estimado por mínimos quadrados ordinários (MQO), sem considerar qualquer tipo de dependência espacial e concluíram que o primeiro apresenta um desempenho superior. Os autores ainda testam a capacidade preditiva de vários tipos de algoritmos na pendularidade, utilizando a base de dados do Censo 2010 e 22 variáveis diferentes.

Vários pesquisadores que identificam as limitações inerentes ao modelo gravitacional também propõem a utilização de modelagens baseadas em *XGBoost* (Liang *et al.*, 2021; Morton; Piburn; Nagle, 2018; Spadon *et al.*, 2019), as quais demonstram superar tais restrições ao apresentar uma capacidade superior de explicar e de prever a dinâmica pendular. No entanto,

é perceptível que a comparação entre o desempenho desses modelos mais recentes e os modelos gravitacionais tradicionais é conduzida com base na aplicação deste último em sua configuração convencional, isto é, através da equação gravitacional estimada por meio de mínimos quadrados ordinários e desprovida de controle de efeitos espaciais. Tal abordagem deixa uma lacuna na literatura e suscita a necessidade de comparar a ferramenta de aprendizado de máquina com uma versão mais refinada da equação gravitacional, considerando deslocamentos entre pares de municípios e incorporando controles para os efeitos econômicos da proximidade geográfica.

Aproveitando esta oportunidade para propor novas abordagens tanto para a equação gravitacional quanto para o modelo *XGBoost* na análise da pendularidade, é crucial definir como essas modelagens serão especificadas em consonância com as características dos dados disponíveis. Esta tarefa será empreendida no próximo capítulo.

# 3. PROCEDIMENTOS METODOLÓGICOS

Como evidenciado na revisão de literatura, a maioria dos estudos sobre a pendularidade emprega o modelo gravitacional, dada a sua eficácia histórica em explicar padrões de mobilidade baseados nas populações e na distância entre os locais. Entretanto, avanços recentes na análise de dados têm visto uma crescente adoção de algoritmos de ML, que têm demonstrado melhorias significativas em relação ao modelo gravitacional tradicional. Estes ganhos podem ser atribuídos à capacidade do *XGBoost* de capturar relações não-lineares e interações complexas entre variáveis, que possivelmente são desconsideradas em modelos gravitacionais tradicionais.

Contudo, é importante notar que muitos desses estudos que utilizam o *XGBoost* comparam seus resultados diretamente com os do modelo gravitacional tradicional. A inclusão do efeito espacial pode ser importante para a pendularidade e, dessa forma, aprimorar ambos os modelos, podendo aumentar seus poderes explicativos.

Dada essa conjuntura, este estudo justifica a aplicação de ambos os modelos propostos para análise da pendularidade. O diferencial desta pesquisa reside na estimativa de um modelo gravitacional não-linear, incorporando a dimensão espacial através de uma variável de interação e de um *XGBoost* que considera tal dimensão. Esta abordagem busca não apenas comparar a eficácia dos modelos, mas também entender como a inclusão de fatores espaciais pode aprimorar os modelos, oferecendo uma nova interpretação dos padrões de pendularidade na RM de Londrina.

Dessa forma, a análise econométrica proposta apresentará algumas interpretações como: a análise do efeito de cada variável sobre a pendularidade; a comparação do desempenho de cada modelo com e sem efeitos espaciais isoladamente, ou seja, determinar se o efeito espacial constituiu algum diferencial para explicar a pendularidade em cada modelo e; determinar se, após a sofisticação do modelo gravitacional, este ainda seria útil para explicar o fenômeno frente à modelagem de ML.

O estudo adotado possui metodologia quantitativa em função das ferramentas estatísticas empregadas, natureza descritiva ao identificar as relações entre as variáveis propostas na modelagem e a abordagem experimental decorre do experimento realizado para o *cluster* paranaense.

A metodologia desta pesquisa propõe investigar como interagem os aspectos espacial e econômico dos municípios na determinação dos fluxos pendulares dentro da RM de Londrina, lançando mão de dois tipos de estimação: *XGBoost* e Modelo Gravitacional.

Este capítulo está dividido em cinco partes: na primeira o modelo gravitacional é descrito com suas respectivas características, críticas e propostas de modelagem específica para a base de dados; na segunda expõem-se os motivos para a inserção de efeitos espaciais em modelos que analisam a pendularidade e; na terceira parte do capítulo é detalhado o modelo *XGBoost* com sua ferramenta SHAP; na sequência, são apresentadas as métricas para comparação da performance e da capacidade preditiva entre o modelo gravitacional e o modelo *XGBoost* e; por fim, se encontram a base de dados e as variáveis que são apresentadas com suas respectivas justificativas.

## 3.1 O modelo gravitacional

A variável de interesse desta pesquisa passa por uma transformação que será explicada com mais detalhes posteriormente, porém, é importante abordar como a sua natureza e a sua transformação afetam a escolha da estimação da equação gravitacional. Na amostra criada para este estudo, cada observação possui informação sobre quantos *commuters* se deslocam da cidade $i$ para a cidade $j$, isto é um exemplo de dado de contagem. De acordo com Wooldridge (2002), dados de contagem são aqueles cujas variáveis assumem valores inteiros não negativos e, para modelá-los, emprega-se algumas estimações específicas como o PPML (Poisson Pseudo-Máxima Verossimilhança). Alguns exemplos de variáveis de contagem são o número de vezes que alguém foi preso durante um certo ano, o número de casos de overdose em uma sala de emergência numa dada semana e o número de cigarros fumados por dia (Wooldridge, 2002).

No contexto do tratamento de dados de contagem em corte transversal, os modelos mais populares são os *Poisson* e, de acordo com Cameron e Trivedi (2013), o modelo *Poisson* máxima-verossimilhança (PML) é o mais simples utilizado para a tarefa de regressão destes dados, com as seguintes premissas: i) É preciso especificar corretamente a média condicional para que os estimadores sejam consistentes; ii) Qualquer inferência estatística significante requer equidispersão, ou seja, a variância condicional deve ser proporcional à média condicional; iii) Mesmo que os dados não sejam equidispersos, ainda é possível realizar inferência desde que pelo menos a média condicional seja bem especificada; iv) A variável dependente não precisa ter distribuição *Poisson* para que a estimação seja consistente, basta que os item imediatamente anterior seja válido; v) Se os dados não forem equidispersos, então existe estimadores melhores do que PML.

De acordo com Cameron e Trivedi (2013), a restrição de equidispersão da estimação PML é análoga à restrição da homoscedasticidade em modelos lineares, ou seja, se os erros forem heterocedásticos em um modelo linear, a estimação por MQO deixa de ser preferível. O mesmo ocorre na situação de violação do pressuposto de equidispersão, algo frequente em dados de contagem; por isso, estimação PPML se torna uma maneira melhor de modelar esses dados.

Segundo Wooldridge (2002), a equidispersão é um dos pressupostos mais importantes de uma distribuição Poisson e pode ser definida matematicamente ao observar-se a proporção entre a variância condicional e a média condicional:

$$V(Y_i|x_i) \equiv hE(Y_i|x_i) \tag{3.1.1}$$

onde $h > 0$ representa a proporção variância-média (Wooldridge, 2002). Esta proporção assume alguns valores que representam situações específicas, como $h = 1$, que representa a situação de equidispersão, $h > 1$ significa que a variância condicional (lado esquerdo da equação) é maior do que a média condicional, ou seja, sobredispersão e, por último, $h < 1$ significa que a média é maior, ou seja, subdispersão. O problema causado por dados sobredispersos, por exemplo, é que as estatísticas t calculadas dos modelos PML ficam infladas, o que leva a conclusões otimistas demais a respeito da significância estatística das variáveis explicativas (Cameron; Trivedi, 2013).

O modelo PML especifica que a variável dependente $Y_i$, dada a variável explicativa $x_i$, possui distribuição Poisson com a seguinte função densidade (Wooldridge, 2002):

$$f(Y_i|x_i) = \frac{e^{-\mu_i}{\mu_i}^{Y_i}}{Y_i!} \quad , Y_i = 0,1,2,3, \ldots \tag{3.1.2}$$

determinada pela média condicional:

$$E(Y_i|x_i) = \mu_i = \exp(x_i'\beta) \tag{3.1.3}$$

Essa especificação (3.1.3) é chamada função de média exponencial com modelo de regressão Poisson, sendo também um modelo log-linear pois o logaritmo da média condicional é linear nos parâmetros, ou seja, $\ln E(Y_i|x_i) = x_i'\beta$ .

Então, supondo que as observações sejam independentes, a função de log-verossimilhança é:

$$\ln L(\beta) = \sum_{i=1}^{n}\{Y_i x_i'\beta - exp(x_i'\beta) - lnY_i!\} \quad (3.1.4)$$

O estimador PML $\hat{\beta}_{PML}$ é a solução da condição de primeira ordem da equação (3.1.4):

$$\sum_{i=1}^{n}[Y_i - exp(x_i'\beta)]x_i = 0 \quad (3.1.5)$$

O pressuposto de distribuição, de acordo com Cameron e Trivedi (2013), é mais forte do que o necessário para realizar inferência nos parâmetros de $\hat{\beta}_{PML}$ e, para os estimadores serem consistentes, basta que o lado esquerdo da equação (3.1.5) tenha valor esperado igual a zero, ou seja:

$$E(Y_i|x_i) = \exp(x_i'\beta) \;\rightarrow\; E\{[Y_i - exp(x_i'\beta)x_i]\} = 0 \quad (3.1.6)$$

Dessa forma, de acordo com essa condição de primeira ordem, $Y$ não precisa nem ser uma variável de contagem, basta apenas que $Y \geq 0$ (Cameron; Trivedi, 2013). Assim, dada a robustez ao pressuposto de distribuição, é possível utilizar $\hat{\beta}_{PML}$ mesmo que o processo gerador de dados para $Y_i$não seja Poisson, fazendo com que o estimador se torne em PPML e assim, as condições de primeira ordem de (3.1.5) são também capazes de gerar este estimador (Cameron; Trivedi, 2013). Ou seja, assume-se a média condicional Poisson, mas, relaxa-se a restrição de equidispersão e o estimador $\hat{\beta}_{PPML}$ é definido com a solução de (3.1.5) desde que a função da média exponencial (3.1.3) seja válida:

$$\hat{\beta}_{PPML} \sim [\beta, V_{PPML}[\hat{\beta}_{PML}] \quad (3.1.7)$$

onde:

$$V_{PPML}[\hat{\beta}_{PML}] = (\sum_{i=1}^{n}\mu_i x_i x_i')^{-1}(\sum_{i=1}^{n}\omega_i x_i x_i')(\sum_{i=1}^{n}\mu_i x_i x_i')^{-1} \quad (3.1.8)$$

e:

$$\omega_i = V(Y_i|x_i) \tag{3.1.9}$$

é a variância condicional de $Y_i$. A implementação de (3.1.8) depende da forma funcional assumida por (3.1.9) (Cameron; Trivedi, 2013).

De acordo com Santos Silva e Tenreyro (2006), a forma do estimador PPML em (3.1.5) deixa claro que tudo o que ele precisa para ser consistente é especificar a média condicional $E(Y_i|x_i) = \exp(x_i'\beta)$, além disso, os dados sequer precisam ter distribuição Poisson e nem precisam ser inteiros, no caso da variável dependente.

Na pesquisa econômica, o modelo gravitacional postula que uma força atrativa de interação entre duas regiões onde há atividade humana é gerada pelas massas populacionais dessas mesmas duas áreas e uma força repulsiva dessa mesma interação é causada pelo tamanho do espaço entre essas duas massas, onde tal fenômeno ocorre (Carrothers, 1956). Ou seja, esse modelo se baseia na premissa de que a magnitude dos fluxos de deslocamento entre duas áreas é diretamente proporcional à massa populacional das áreas envolvidas e inversamente proporcional à distância entre elas.

De acordo com Simini *et al* (2012), esta é a forma matemática do modelo:

$$T_{i,j} = \frac{m_i^{\alpha} n_j^{\beta}}{f(r_{ij})} \tag{3.1.10}$$

Onde $T_{i,j}$ representa o número de indivíduos que se deslocam entre as áreas *i* e *j*, $m_i$ é a população de origem *i*, $n_j$ é a população da área de destino *j* e $r_{ij}$ é a distância entre as áreas *i* e *j*. $\alpha$, $\beta$ são parâmetros ajustáveis e $f(r_{ij})$ é a função de dissuasão escolhida para ajustar os dados.

Wölver, Burgard e Breßlein (2018) afirmam que os modelos gravitacionais são divididos em dois grupos de acordo com a estimação empregada: Modelos estimados na forma log-log, como MQO, mínimos quadrados iterados estruturais e *Tobit's*. O outro grupo estima modelos gravitacionais na forma multiplicativa com modelos lineares generalizados, como PPML, gamma pseudo máxima-verossimilhança (GPML) e mínimos quadrados não-lineares (NLS). Segundo os autores, as vantagens das estimações do último grupo são: a facilidade em entendê-los e programá-los; seus resultados são facilmente comparáveis entre si e; com exceção dos métodos *Tobit,* formam o único grupo capaz de lidar com valores zero na variável dependente.

De acordo com Head e Mayer (2014), uma vantagem da equação gravitacional é que, depois de aplicar logaritmo natural na equação, os parâmetros podem ser estimados por MQO, porém, alertam para duas questões fundamentais para a estimação correta da equação gravitacional que, se negligenciadas, podem gerar resultados muito diferentes dos valores observados. 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, a equação gravitacional é estimada por MQO na maioria dos estudos e, de acordo com Santos Silva e Tenrenyo (2006), mesmo que controlados por efeitos fixos, estes erros serão heterocedásticos, o que pode gerar um viés muito forte. O modelo gravitacional estimado por MQO na forma log-linear, da equação (3.1.10) seria:

$$\ln\left(T_{i,j}\right) = \alpha \ln(m_i) + \beta \ln\left(n_j\right) - \ln\left(f\left(r_{ij}\right)\right) \qquad (3.1.11)$$

Esta transformação log-linear tradicional da equação gravitacional, sob condições de heteroscedasticidade, resulta em estimativas que não são confiáveis. Isso acontece pois o valor esperado do logaritmo de uma variável aleatória depende de momentos de ordem mais alta da sua distribuição. Dessa forma, caso haja erros com heteroscedasticidade, é provável que exista uma correlação entre estes erros transformados e as variáveis independentes do modelo (Santos Silva; Tenrenyo, 2006). Em outras palavras, pode haver endogeneidade no modelo e os estimadores de MQO, que se comportariam como elasticidades, podem ser viesados e inconsistentes.

Quanto às observações com valor nulo na variável dependente, é muito comum durante a coleta de dados sobre pendularidade entre municípios se observar fluxos zero. Por exemplo, entre dois municípios-satélite dentro de alguma região metropolitana que não recebem e nem doam trabalhadores entre si, possivelmente uma consequência das assimetrias de informação, do custo alto de deslocamento pendular, dentre outros fatores (Brito; Ramalho, 2019). Isto acaba se tornando um problema para a modelagem pois a aplicação do logaritmo de zero, com a finalidade de resolver o problema da independência dos fluxos, não é definida matematicamente.

Os autores também reforçam a atenção nos fluxos nulos em modelos gravitacionais e como os dados de fluxos pendulares são mais afetados:

> Pode-se notar que um dos problemas mais tratados na literatura – especialmente a literatura voltada às trocas comerciais – está diretamente relacionado aos potenciais vieses provenientes do tratamento inadequado dos fluxos com valor zero. No geral, os

> modelos gravitacionais expressam as relações com múltiplas variáveis estritamente positivas, isto é, consideram-se variáveis que não geram valores nulos. Contudo, ao se explorar o deslocamento pendular, naturalmente surgirão fluxos nulos entre municípios. Tem-se, portanto, um problema relacionado aos fluxos zeros (Brito; Ramalho, 2019, p. 108).

De acordo com Santos Silva e Tenrenyo (2006), muitos pesquisadores excluem essas observações de suas bases de dados e estimam os modelos por MQO, para não lidarem com esse problema, porém, este processo de linearização da equação gravitacional e posterior estimação por MQO, ao excluir os fluxos zero pode causar viés de seleção, posto que as informações contidas, mesmo que sejam zero, são importantes demais para serem descartadas sem algum critério teórico, então esta não é uma opção viável (Head; Mayer, 2014).

Outra maneira bastante difundida de tratar os valores zero na variável dependente é aplicar uma constante igual a um, transformando $(T_{i,j} + 1)$ ou estimar utilizando *Tobit*, porém, isto também leva a estimadores inconsistentes (Santos Silva; Tenrenyo, 2006). Head e Mayer (2014) acrescentam que este método de adição de constante deve ser evitado porque os resultados dependem das unidades de medida, ou seja, a interpretação das elasticidades é prejudicada. Os autores ainda citam um exemplo de como esse procedimento é prejudicial com os dados da sua própria estimação: o impacto estimado de um dos coeficientes mudou de negativo e significativo para positivo e insignificante apenas ao mudar a unidade de medida da observação de milhões para bilhões.

Diante da delineação dos efeitos da heterocedasticidade e da manifestação de fluxos nulos na estrutura do modelo, optou-se pela técnica de estimação do modelo gravitacional por meio da Poisson Pseudo-Máxima Verossimilhança (PPML) na tentativa de estimar a pendularidade corretamente dados esses dois problemas, seguindo Santos Silva e Tenreyro (2006). Tal abordagem propicia a renúncia à log-linearização da equação conferindo assim um tratamento apropriado aos fluxos nulos (Head; Mayer, 2014). Uma matriz de covariância robusta pode ser utilizada para lidar com a heteroscedasticidade. Dessa maneira as estimativas se tornam consistentes independentemente da distribuição dos dados, se são ou não Poisson, ou até mesmo se são dados de contagem (Arvis, Shepherd, 2013). Assim, o modelo PPML é utilizado em detrimento do uso do modelo PML que considera que os dados sejam variáveis independentes com distribuição Poisson, pois relaxa a restrição deste de equidispersão (Cameron; Trivedi, 1998), ou seja, para que o estimador seja consistente, só é preciso a especificação correta da média condicional (Santos Silva; Tenreyro, 2006) dado que não é necessário que os dados possuam tal distribuição.

O fato de a variável dependente ter uma alta proporção de zeros não afeta o desempenho do estimador, pelo contrário, a presença de zeros é um motivo adicional para usar a abordagem PPML porque, nesse caso, todos os estimadores baseados na log-linearização da equação gravitacional precisam usar métodos não razoáveis, como dois destacados acima, para lidar com observações zero (Santos Silva; Tenreyro, 2011).

Head e Mayer (2014) afirmam que a estimação PPML se mantém consistente e eficiente mesmo sob sobredispersão, ou seja, mesmo quando a variância condicional é maior do que a média condicional. Além de que, mesmo na presença de uma grande quantidade de zeros na variável dependente, a performance dos estimadores PPML permanece adequada.

Santos Silva e Tenreyro (2006) acrescentam que os estimadores PPML são boas alternativas aos modelos log-lineares estimados por MQO, representando uma forma natural de lidar com valores zero na variável dependente, após compararem as estimativas obtidas por PPML e por MQO na forma log-linear, tanto na forma tradicional quanto por efeitos fixos. Head e Mayer (2014) ainda sugerem que, mesmo sendo preferível às outras estimações, é necessário compará-la com estimação por MQO e por GPML destacando que: i) Se os três modelos produzem estimativas parecidas, então a modelagem está bem especificada; ii) Se as estimativas de PPML e GPML são idênticas, porém, diferentes das estimativas de MQO, então há um problema de heterocedasticidade e este modelo não é confiável.

Santos Silva e Tenreyro (2011) afirmam que, apesar de ambos os estimadores PPML e GPML serem consistentes e geralmente bem-comportados, o estimador PPML parece ser mais robusto à desvios dos pressupostos implícitos de heterocedasticidade, além disso, este estimador é bem-comportado mesmo quando não há equidispersão. Os autores ainda afirmam que estimadores *Tobit* e MQO têm vieses muito altos que não desaparecem com o aumento da amostra, confirmando que são inconsistentes.

Agora que se sabe que a maioria dos dados nos problemas estatísticos não são equidispersos, que a estimação PPML não requer que as observações representem inteiros e que esta é adequada para tratar observações zero na variável dependente, é fácil explicar a transformação pela qual a variável de interesse desta pesquisa é submetida. Como dito anteriormente cada observação de $Y_i$ possui uma informação sobre a quantidade de *commuters* que se deslocam da cidade $i$ para a cidade $j$ e, após a reunião destes dados, divide-se cada valor do deslocamento pela população do município de origem, fazendo com que os dados (quando não forem zero) deixem de ser inteiros. Isto representaria um problema para a estimação caso não se aplicasse estimação PPML na equação gravitacional.

Dessa forma, nesta pesquisa, a taxa de deslocamento pendular entre dois municípios é modelada como uma função exponencial dos Índices IFDM-ER (Índice Firjan de Desenvolvimento Municipal – Emprego e Renda) de cada município, da distância entre eles, e de outras variáveis relevantes, com os parâmetros estimados utilizando a técnica PPML para contornar problemas de heteroscedasticidade e excesso de zeros nos dados de pendularidade. A forma do modelo gravitacional PPML é a seguinte:

$$Tx.Pendulares_{ij} = \exp\left(\beta_0 + \beta_1 IFDM_i + \beta_2 IFDM_j + \beta_3 Distância_{ij} + \gamma Z_{ij} + \varepsilon_{ij}\right) \quad (3.1.12)$$

Onde $Tx.Pendulares_{ij}$ é a taxa de deslocamento pendular entre os municípios i e j; $IFDM_i$ e $IFDM_j$ são os índices IFDM-ER dos pares de municípios; $Distância_{ij}$ é a distância entre as duas cidades, $Z_{ij}$ é o termo que engloba regressores adicionais que podem influenciar a pendularidade, como, por exemplo, *dummies* de interação entre ser um município vizinho e sua dinâmica econômica, presença de algum tipo específico de indústria, política municipal de mobilidade, entre outros; $\varepsilon_{ij}$ é o termo do erro do modelo e $\beta_0,\ \beta_1, \beta_2,\ \beta_3, \gamma$ são os parâmetros a serem estimados.

Mesmo após apresentar a base teórica que sustenta a escolha pela estimação PPML, um teste de log-normalidade dos resíduos, chamado teste MaMu (Head; Mayer, 2014), é aplicado para se certificar de que a estimação PPML é a mais adequada. Apesar de estimadores PPML serem consistentes mesmo quando a variância é mal especificada, pode-se usar o teste MaMu (Santos Silva; Tenreyro, 2011). Este teste serve para escolher o melhor modelo ao analisar a distribuição do termo do erro, tomando o logaritmo da equação $V(Y_i|x_i) \equiv hE(Y_i|x_i)^{\lambda}$, que é uma generalização da equação (3.1.1) onde $\lambda = 1$ ou seja, uma situação de equidispersão. Este procedimento, após a aplicação do logaritmo, substitui a variância e o valor esperado por seus equivalentes na amostra para obter a equação abaixo estimada por MQO (Head; Mayer, 2014):

$$\ln \varepsilon_i^2 = \ln h + \lambda \ln E(Y_i|x_i) \quad (3.1.13)$$

$$\ln \varepsilon_i^2 = constante + \lambda \ln \widehat{Y}_i \quad (3.1.14)$$

Portanto, os resíduos foram calculados pela diferença entre os valores observados da variável dependente e os valores ajustados pelo modelo. Após este passo, tanto os quadrados dos resíduos quanto os logaritmos desses resíduos quadrados foram computados. Por último,

os logaritmos dos valores ajustados foram obtidos e utilizados como variável independente em uma regressão estimada por MQO contra os logaritmos dos resíduos quadrados (Head; Mayer, 2014)[5]. Dessa forma, se o coeficiente estimado $\lambda$ em (3.1.14) for significativo e apresentar valor menor que 2, significa que a estimação PPML é a mais apropriada e se $\lambda = 2$, a estimação GPML se torna mais adequada (Head; Mayer, 2014).

Na próxima seção, a importância dos efeitos espaciais na pesquisa sobre pendularidade, bem como a inserção desses efeitos nos modelos através da criação de novas variáveis, serão apresentadas.

## 3.2 Os efeitos espaciais na pendularidade.

Os métodos de identificação e incorporação de efeitos espaciais nos modelos de mobilidade pendular buscam descobrir novas relações entre as variáveis, a fim de proporcionar novas perspectivas na interpretação do fenômeno da pendularidade sob a ótica espacial. Por exemplo, esses procedimentos possibilitam compreender as discrepâncias nos fluxos pendulares quando um município específico possui um, dois ou mais vizinhos com níveis de desenvolvimento mais elevados.

Portanto, o espaço e os seus potenciais efeitos não podem ser ignorados quando se estuda o deslocamento dos trabalhadores, não somente para compreender o alcance da influência de variáveis econômicas e espaciais, mas também para produzir modelos mais eficientes, dado que a ausência de variáveis que representem estes efeitos, quando relevantes, provoca viés de omissão de variável nas estimativas (Almeida, 2012).

Segundo Le Sage e Fischer (2010) há algumas justificativas econométricas para o uso de defasagens espaciais da variável dependente em modelos de interação espacial, porém, não há justificativas teóricas para tal. Esta constatação os motivou a abordar a utilização dessas defasagens em modelos econométricos para o comércio internacional, a migração e a pendularidade. A ideia dos autores era produzir modelagem com efeitos espaciais com foco nos padrões de interação de maneira agregada ao invés de focar nas relações bilaterais entre os espaços. Em um exemplo, os autores afirmam que, no caso de uma variável dependente de mercado de trabalho, uma mudança, *ceteris paribus*, em um dado município provocaria efeitos de transbordamento nas oportunidades de mercado de trabalho nos outros municípios.

[5] Mais detalhes sobre a definição matemática do teste MaMu em Head e Mayer (2014).

Possivelmente, isso não proporciona uma base teórica robusta para a inclusão desses efeitos na variável dependente do modelo de pendularidade. Qual seria o impacto da variação na taxa de deslocamento pendular de uma cidade específica sobre as taxas das cidades vizinhas? Não é necessariamente a magnitude dos fluxos que afeta uns aos outros, mas, sim as variáveis econômicas subjacentes, como um aumento na atividade do mercado de trabalho de um município, a inauguração de um sistema de transporte de alta velocidade que reduziria drasticamente os custos de deslocamento, ou qualquer fator que influencie os fluxos.

Os exemplos mencionados estão contemplados pelas variáveis explicativas do modelo. Portanto, um possível transbordamento de efeitos espaciais teria origem principalmente nessas variáveis explicativas. A variável dependente, por esse motivo, não apresentaria tais transbordamentos, sendo apenas uma consequência do que ocorre no lado oposto da equação. Criou-se então uma matriz de contiguidade, que informa, para cada par de municípios se eles são vizinhos (fronteiriços) ou não, atribuindo valor igual a 1 para o caso de serem contíguos e 0 caso contrário.

A ideia por trás dessa nova variável é abordar a possibilidade de que as variáveis explicativas do modelo apresentem algum efeito de transbordamento direcionado para as regiões vizinhas através da criação de uma variável de interação entre a matriz de contiguidade e o Índice Firjan-ER. Então, o produto da *dummy* de contiguidade com a variável de mercado de trabalho do município de destino dá origem a uma *dummy* de interação entre ser vizinho *versus* o desenvolvimento do município de destino. Em outras palavras, busca-se entender se o nível de desenvolvimento de um município afeta o fluxo pendular do seu vizinho, ou seja, a diferença é que agora assume-se que o município de destino é necessariamente um vizinho fronteiriço e, só então após esta assunção, analisa-se o efeito da variável de dinâmica de mercado de trabalho do município sobre a cidade de origem do deslocamento. Adicionalmente, a própria *dummy* de contiguidade foi utilizada como variável explicativa na tentativa de modelar o efeito espacial.

Um ponto crucial que justifica a introdução de uma variável *dummy* de interação para capturar os efeitos espaciais do desenvolvimento econômico na pendularidade, em oposição ao uso da defasagem espacial do desenvolvimento econômico, é a complexidade na modelagem desta última. Esse desafio é particularmente evidente ao considerar o fluxo pendular específico entre pares de municípios, conforme é o foco desta pesquisa, em contraposição ao fluxo total de entrada e/ou saída de cada município. Vale ressaltar que trabalhos que adotam a modelagem econométrica espacial convencional para controlar esse efeito espacial nos modelos pendulares limitam-se a considerar fluxos totais, negligenciando a especificidade dos pares de municípios.

Este é o caso da pesquisa de Brito e Silva (2020), na qual a variável dependente consiste no fluxo total de saída ou de entrada de pendulares em cada município, sem uma análise detalhada do fluxo específico entre pares de municípios. Consequentemente, a quantidade de trabalhadores que se deslocam diretamente do município A para o município B, por exemplo, não pode ser estimada.

Sellner, Fischer e Koch (2010) comparam a qualidade do ajuste do seu modelo PML gravitacional com efeitos espaciais na variável dependente, através do erro quadrático médio (RMSE) e do $R^2$ ajustado, contra outros modelos gravitacionais sem efeitos espaciais e estes se provaram piores neste aspecto. Isto posto, os modelos estimados nesta pesquisa procuraram incorporar os efeitos da dinâmica econômica dos municípios vizinhos na determinação da pendularidade do município "i", controlando potenciais impactos do "Índice Firjan ER" do vizinho "j". Desse modo, é realizada uma estimação do tipo SLX, na qual o efeito espacial é incorporado na variável explicativa.

O modelo SLX, por não ter um multiplicador espacial, apresenta os impactos de transbordamento das regiões próximas de forma local e não global (Almeida, 2012), ou seja, os impactos são interpretados apenas para os fluxos dos municípios vizinhos. Outra questão alertada por Almeida (2012) é que se espera uma correlação alta entre a variável explicativa e a sua própria defasagem espacial, gerando multicolinearidade.

Portanto, a *dummy* de interação utilizada nesta pesquisa possibilita a estimação do modelo gravitacional sem considerar efeitos espaciais, permitindo analisar, por exemplo, como o desenvolvimento econômico de um município receptor influencia o fluxo de saída de pendulares de outro município. Em seguida, ao introduzir efeitos espaciais, a estimação do mesmo modelo gravitacional possibilita examinar a importância do desenvolvimento econômico desse município (agora assumido como vizinho) sobre o mencionado fluxo de saída do outro município. Essas conclusões são valiosas ao investigar, por exemplo, em que medida a instalação de uma indústria específica em uma região atrai trabalhadores de áreas vizinhas.

Outro argumento em favor da inserção de efeitos espaciais na modelagem é a constatação de Moura (2010) sobre a concentração da movimentação pendular nos municípios que são vizinhos ou muito próximos uns dos outros e de que fluxos de saída se concentram de cidades bem próximas ao polo da região metropolitana, como os municípios vizinhos da RM de Londrina em direção ao próprio polo, ou seja, não somente ser próximo, mas ser vizinho ou suficientemente próximo de um município com maior dinâmica econômica pode fazer a diferença no deslocamento pendular.

### 3.3 *XGBoost*

Lundberg et al (2020) afirmam que à medida em que vai se experimentando não-linearidade nos dados, modelos lineares vão perdendo *performance* em termos de acurácia e de interpretabilidade em relação aos modelos de *gradient boosted trees* por causa do viés cada vez maior, mas, apesar de o modelo gravitacional proposto ser um exemplar de modelo não-linear, o que, teoricamente, não o faria sofrer de tal problema da mesma maneira que um modelo linear, faz-se necessário comparar seu desempenho com uma técnica de ML como o *XGBoost*.

Este modelo de ML é, na verdade, uma biblioteca com um conjunto de técnicas de modelagem baseados na pesquisa de Friedman (2001) sobre *Gradient Boosting* e implementado na forma do *XGBoost* por Cheng e Guestrin (2016). Em modelos como este, as variáveis explicativas são comumente chamadas de características enquanto a dependente é chamada de alvo. Isto ocorre porque o modelo busca prever ou explicar uma variável alvo dadas as suas características, porém, apesar da nomenclatura, a relação variável explicada-explicativa permanece. O modelo é aplicado com e sem uma variável que representa efeitos espaciais, seguido de análise dos valores SHAP (SHapley Additive exPlanations) para interpretar as estimativas, teste de capacidade preditiva e, por último, é feita a análise das métricas de comparação entre os modelos gravitacionais com e sem a variável supracitada de efeitos espaciais e os modelos *XGBoost* propostos aqui.

Para explicar o funcionamento dessa ferramenta é interessante apresentar conceitos mais básicos de outras ferramentas que a compõem. Primeiramente, uma árvore de decisão classifica dados ao impor uma série de perguntas a respeito de características associadas. Cada pergunta está contida em um nó (ramificação) que atribui um valor a cada resposta a respeito dessa característica e leva a um próximo nó com uma nova pergunta, atribuindo um novo valor e isto se repete quantas vezes se deseja, formando uma árvore (Kingsford; Salzberg, 2008). A forma da árvore se dá ao se imaginar uma raíz como uma pergunta inicial, seguida de ramificações e a soma dos valores da trajetória até o resultado, que se chamaria uma folha, a qual contém todas as características específicas atribuídas àquele valor.

Árvores de decisão são criadas adicionando progressivamente nós que fazem perguntas, com base em exemplos de treino já classificados para orientar qual pergunta será feita. O ideal seria que uma pergunta simples e única conseguisse separar todos os exemplos de treinamento nas categorias corretas. Quando não é possível encontrar uma pergunta que faça uma separação perfeita, escolhe-se uma que faça a melhor separação possível. Uma pergunta eficaz é aquela que consegue dividir um grupo de itens com diferentes categorias em grupos

menores, onde os itens de cada grupo são quase todos da mesma categoria, organizando os dados de forma que cada grupo tenha uma variação mínima nas suas categorias (Kingsford; Salzberg, 2008).

Agora que se definiu uma árvore de decisão é mais fácil explicar a modelagem do *XGBoost:* um conjunto de árvores de regressão e classificação ou CART[6] (Chen; Guestrin, 2016) que, ao contrário das árvores de decisão onde cada folha contém um valor de decisão, cada CART contém um valor real que permite trabalhar com regressão além de apenas classificação. Chen e Guestrin (2016) definem uma função objetivo, que pode tomar qualquer forma que se deseje, representando esse conjunto de $K$ árvores, dada uma amostra com $n$ exemplos e $m$ características para prever um resultado $\hat{Y}_i$.

$$\hat{Y}_i = \phi(x_i) = \sum_{k=1}^{K} f_k(x_i)\,,\ f_k \in \mathcal{F}, \tag{3.3.1}$$

$$\text{onde } \mathcal{F} = \left\{f(x) = w_{q(x)}\right\}\ ,\ (q\colon \mathbb{R}^m \to \mathrm{T}, w \in \mathbb{R}^{\mathrm{T}}) \tag{3.3.2}$$

onde T é a quantidade de folhas na árvore, q(x) é uma função que representa a correspondência de cada observação ao valor da sua respectiva folha, ou seja, é o caminho de cada árvore até o valor da 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 todas as árvores de regressão possíveis, ou seja, todos os CART's. Além disso, como dito anteriormente, diferentemente das árvores de decisão, cada CART pode ser otimizado e tem um valor contínuo em cada folha, então, a estrutura de decisão das árvores, dada por q, é utilizada para prever o valor final somando os valores das suas folhas dadas por w (Chen; Guestrin, 2016).

De acordo com os autores, para o aprendizado do modelo de conjunto de árvores, define-se e minimiza-se a função objetivo regularizada $\mathcal{L}(\phi)$ (Chen; Guestrin, 2016):

$$\mathcal{L}(\phi) = \sum_i l(y_i, \hat{y}_i) \sum_k \Omega(f_k)\,, \tag{3.3.3}$$

$$\Omega(f_k) = \gamma \mathrm{T} + \frac{1}{2}\lambda ||w||^2 \tag{3.3.4}$$

[6] Classification and regression trees

onde $l$ é uma função de perda convexa e diferenciável e mede a diferença entre a previsão $\hat{y}_i$ e o alvo $y_i$, $\Omega$ é uma função de regularização que penaliza a complexidade do modelo e ajuda a evitar sobreajuste (Chen; Guestrin, 2016), $\gamma$ representa a complexidade de cada folha, $\lambda$ é um parâmetro de comprometimento aplicado para mensurar a penalidade sobre o valor das folhas, T é o total de folhas em todo o modelo e *w* representa o vetor com o peso das folhas (Ma *et al*, 2021). Dessa forma, percebe-se que uma boa função objetivo é acurada, graças à função de perda que busca aproximar o valor previsto do valor observado, mas, também é simples, graças à função de regularização. Ao se observar os operadores de soma na função objetivo (3.3.3), pode-se perceber que se trata de várias funções adicionadas, ou seja, de um conjunto de árvores e chama-se isto de um modelo de *tree boosting*:

$$\mathcal{L}^{(t)} = \sum_n l(y_i, \hat{y}_i^{(t)}) \sum_k \Omega(f_k) \tag{3.3.5}$$

O modelo de *tree boosting* acima tem parâmetros nas árvores que precisam ser encontrados através do aprendizado das funções $f$ que contêm as estruturas das árvores T e o valor de cada folha $w$. Encontrar esses parâmetros é uma tarefa mais complexa do que seria em problemas de otimização tradicionais onde se poderia tomar o gradiente da função, então, a abordagem de *boosting* é utilizada no treinamento do modelo para minimizar a função objetivo da seguinte forma: Adiciona-se uma nova árvore ou função $f_t$ na t-ésima iteração enquanto o modelo prossegue com o treinamento, ou seja, o treinamento se dá de maneira aditiva (Osman *et al*, 2021).

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

$$\hat{y}_i^{(0)} = 0 \tag{3.3.6}$$

$$\hat{y}_i^{(1)} = f_1(x_i) = \hat{y}_i^{(0)} + f_1(x_i) \tag{3.3.7}$$

$$\hat{y}_i^{(2)} = f_1(x_i) + f_2(x_i) = \hat{y}_i^{(1)} + f_2(x_i) \tag{3.3.8}$$

$$\hat{y}_i^{(3)} = f_1(x_i) + f_2(x_i) + f_3(x_i) = \hat{y}_i^{(2)} + f_3(x_i) \tag{3.3.9}$$

$$\hat{y}_i^{(t)} = \sum_{k=1}^{t} f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i) \tag{3.3.10}$$

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

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_k) \tag{3.3.11}$$

Isso significa que, de maneira míope[7], adiciona-se a árvore $f_t$ que melhora o modelo de forma mais acentuada de acordo com a equação de *tree boosting* (3.3.5). Então, o processo de otimização ocorre por expansão de Taylor até a segunda ordem e, na t-ésima iteração, com as condições de primeira e de segunda ordem abaixo, a função objetivo será:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l[g_i f_t(x_i) + \frac{1}{2} h_i \partial_t^2(x_i)] + \Omega(f_k) \tag{3.3.12}$$

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

Substitui-se (3.3.4) em (3.3.12) e, dado que $I_j = \{i|q(x_i) = j\}$ é o conjunto dos dados observados relacionados à j-ésima folha, tem-se a nova função objetivo do *XGBoost*:

$$\hat{\mathcal{L}}^{(t)} = \sum_{j=1}^{T} [(\sum_{i \in I_j} g_i) w_j + \frac{1}{2} (\sum_{i \in I_j} h_i + \lambda) w_j^2] + \gamma T \tag{3.3.13}$$

Da equação (3.3.13), para uma dada estrutura $q(x)$, pode-se calcular o valor ótimo $w_j^*$ da folha $j$:

$$w_j^* = -\frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda} \tag{3.3.14}$$

e calcular também o valor ótimo da função que define a qualidade da estrutura da árvore $q$:

$$\hat{\mathcal{L}}^{(t)}(q) = -\frac{1}{2} \sum_{j=1}^{T} \frac{(\sum_{i \in I_j} g_i)^2}{\sum_{i \in I_j} h_i + \lambda} + \gamma T) \tag{3.3.15}$$

De acordo com Chen e Guestrin (2016), para encontrar esse valor $\hat{\mathcal{L}}^{(t)}(q)$ é necessário apenas as estatísticas dos gradientes de primeira e de segunda ordem dados por cada folha e

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

então aplicar a fórmula da equação (3.3.15) 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 algorítmo míope começa a otimização a partir de uma folha e vai adicionando ramificações à árvore de maneira iterativa (Ma *et al*, 2021).

Então, 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:

$$\mathcal{L}_{split} = \frac{1}{2}\left[\frac{(\sum_{i\in I_L} g_i)^2}{\sum_{i\in I_L} h_i+\lambda} + \frac{(\sum_{i\in I_R} g_i)^2}{\sum_{i\in I_R} h_i+\lambda} - \frac{(\sum_{i\in I} g_i)^2}{\sum_{i\in I} h_i+\lambda}\right] - \gamma \qquad (3.3.16)$$

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 $\gamma$ representa a regularização sobre a folha adicional, então se o ganho com a divisão for menor do que $\gamma$, 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.

Então, como 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 diferenciável (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).

A técnica de *gradient boosting* proposta incialmente por Friedman (2001) é uma modelagem de aprendizado de máquina que usa um conjunto de modelos de árvores de decisão para prever um resultado. Cada modelo de árvore é treinado para prever o erro residual do modelo anterior e o resultado é uma combinação ponderada dos modelos anteriores de árvore de decisão. Chen e Guestrin (2016) se baseiam na pesquisa do autor supracitado para produzir o modelo *XGBoost* e explicam que se trata de uma biblioteca de ferramentas de código aberto, disponível para qualquer um em diversos aplicativos estatísticos como *Python*, *Julia* e *R*. Os autores ainda apontam que a biblioteca *XGBoost*, diferentemente do *Gradient Boosting* convencional, incorpora um modelo regularizado para evitar a sobredispersão, ou seja, evitar que o modelo se ajuste de maneira excessiva aos dados de treinamento e isto afete a capacidade de estimação de novos dados fora do treino. A biblioteca, diferentemente de outros métodos de ML, também é capaz de lidar com dados esparsos, ou seja, dados em que a maioria dos seus elementos é zero (é o caso desta pesquisa) ou não possui informação e o faz adicionando uma direção padrão em cada nó, o que permite classificar essas observações no mesmo padrão do restante da parte da amostra que contém alguma informação (Chen; Guestrin, 2016).

*XGBoost* é uma ferramenta de *gradient tree boosting* escalável de ML que apresenta uma das melhores atuações em termos de soluções para vários problemas estatísticos em aprendizado supervisionado utilizando uma quantidade reduzida de recursos, além de ser muito utilizado por cientistas de dados devido a sua rapidez em execução de tarefas (Osman *et al*, 2021) e, de acordo com Morton, Piburn e Nagle (2018), especificamente no caso da pendularidade, é um modelo que performa melhor do que os modelos gravitacional e de radiação tradicionalmente aplicados no tema. Ma *et al* (2021) afirmam que o *XGBoost* é atualmente o mais rápido e mais bem integrado dentre os algoritmos de árvores de decisão, além de ser altamente versátil, flexível e capaz de resolver a maioria dos problemas de regressão e de classificação, bem como funções objetivo criadas pelos próprios usuários.

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 econômico como variáveis-chave.

É essencial, após a definição do modelo e a sua aplicação, conseguir extrair alguma informação sobre as variáveis estudadas interpretando seus resultados, então, 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. Destaca-se então a necessidade de ferramentas como o SHAP, que proporciona uma maior transparência e compreensão dos modelos e isto é crucial nesta pesquisa pois entender as previsões é tão importante quanto a acurácia destas.

SHAP é uma ferramenta baseada nos valores *Shapley* da Teoria dos Jogos utilizada para entender como um modelo de ML toma suas decisões ao tentar explicar o desempenho deste separando a informação de como cada característica de entrada, ou seja, cada variável explicativa, contribui para a decisão final e isto é feito adicionando a contribuição de cada característica para chegar ao resultado (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.

Spadon et al (2019) apresentam a forma matemática do valor *Shapley* para uma variável dada uma previsão $x$ para criar uma escala de importância de características dada pela soma dos valores *Shapley* $\phi_i$de toda a amostra:

$$\phi_i(f,x) = \sum_{S \subseteq M \setminus i} \frac{|S|!(M-|S|-1)!}{M} [f_x(S \cup\ i) - f_x(S)] \tag{3.3.17}$$

Onde: M é o total de variáveis (características); S é o conjunto de características utilizadas para induzir o modelo e $f_x$ é a previsão dadas as características selecionadas. Os autores ainda recomendam o uso da biblioteca SHAP para computar o valor *Shapley* devido ao tamanho do cálculo que envolve muitas combinações.

Lundberg et al (2020) afirmam que os valores *Shapley* representam o único método dentre outros do tipo de características aditivas que satisfazem três importantes propriedades extraídas da teoria dos jogos e que são desejáveis para gerar boas explicações do modelo, sejam:

i) Acurácia local (Eficiência, em Teoria dos Jogos):

As somas dos valores *Shapley* $\phi_i(f,x)$ para cada característica $i$ se igualam a $f(x)$ quando se aproximam do modelo original $f$ para uma dada variável $x$:

$$f(x) = \phi_0(x) + \sum_{i=1}^{M} \phi_i(f,x) \tag{3.3.18}$$

Ou seja, a saída da função é a soma dos valores das características e estas são equivalentes à saída do modelo de ML original.

ii) Consistência (Monotonicidade, em Teoria dos Jogos):

Se a importância de alguma característica aumenta ou permanece, com ou sem mudança na importância das outras, a entrada dessa característica não pode diminuir, ou seja, uma variação positiva em uma variável importante não vai reduzir o valor *Shapley* atribuído a ela:

Dados dois modelos $f$ e $f'$

$$f'_x(S) - f'_x(S\backslash i) \geq f_x(S) - f_x(S\backslash i) \tag{3.3.19}$$

Para todos os subconjuntos de características $S \in M$, então

$$\phi_i(f',x) \geq \phi_i(f,x) \tag{3.3.20}$$

iii) Efeitos nulos (*missingness*):

Características que não possuem efeitos sobre o conjunto $f_x$ também não terão importância, ou seja, não terão um valor *Shapley* designado:

Supondo um modelo em que

$$f_x(S \cup i) = f_x(S) \tag{3.3.21}$$

Para todos os subconjuntos de características $S \in M$, então

$$\phi_i(f,x) = 0 \tag{3.3.22}$$

Isto significa que dados sem informação não terão importância atribuída.

O cumprimento dessas três propriedades faz com que a ferramenta SHAP seja consistente e forneça uma boa capacidade interpretativa de modelos *XGBoost* tanto de maneira

local quanto de maneira global, sendo essa uma grande vantagem da sua aplicação, ou seja, 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).

Portanto, ao interpretar os valores SHAP, é crucial lembrar que eles oferecem uma explicação local (Lundberg; Lee, 2017), isto é, eles explicam a previsão para uma única observação de dados, mas, quando se agrega todas as previsões é possível observar a contribuição da variável na previsão da variável dependente. 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.

Agora que estão definidos ambos os modelos gravitacional e *XGBoost*, além do panorama dos efeitos espaciais dentro da pesquisa sobre pendularidade e da sua modelagem, deve-se, após a apresentação dos resultados, realizar uma avaliação do desempenho de ambos e descobrir se os efeitos espaciais são relevantes para melhorar os modelos e se a equação gravitacional consegue ser tão boa quanto a outra técnica. O próximo item elucidará as métricas utilizadas para mensurar e padronizar tal desempenho.

### 3.4 As métricas de comparação

Para realizar a comparação entre o desempenho do modelo gravitacional com o *XGBoost*, é crucial adotar uma métrica padronizada. Métricas comumente utilizadas para essa tarefa, como critérios de informação e coeficiente de determinação ($R^2$), por exemplo, foram preteridas nesta pesquisa pois, apesar dos critérios de informação serem suficientes para modelos gravitacionais, ainda que na forma não-linear estimados por PPML e GPML (Kareem *et al*, 2016), não são utilizados em modelos de ML.

Nesse sentido, 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 mais próximos de 1 possível, com este valor indicando que as matrizes estimada e observada são idênticas. Esta métrica foca na comparação do número de pendulares entre pares de locais que são comuns tanto na rede real quanto na rede prevista. 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:

$$CPC\left(T,\tilde{T}\right)=\frac{2\sum_{i,j=1}^{n}\min\left(T_{ij},\tilde{T}_{ij}\right)}{\sum_{i,j=1}^{n}T_{ij}+\sum_{i,j=1}^{n}\tilde{T}_{ij}} \quad (3.4.1)$$

- ✓ *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:

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

Resumindo, enquanto o CPC dá ênfase à precisão dos volumes de fluxo entre os nós (municípios) da rede, o CPL dá ênfase à precisão na estrutura da rede de conexões. Um modelo pode ter uma alta CPC se prever corretamente o número de pendulares, mesmo que não acerte exatamente quais pares de cidades estão conectados, enquanto um alto CPL indica que o modelo previu corretamente quais pares de cidades estão conectados, independentemente de quantos pendulares foram previstos corretamente entre eles. Ou seja, o CPC mede a correspondência dos volumes de fluxo entre as redes reais e previstas, enquanto o CPL mede a correspondência

na existência de conexões entre as redes. Ambas as métricas são importantes para avaliar diferentes aspectos da qualidade de modelos de previsão de redes de pendularidade.

- ✓ *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.

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

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. Então, após a determinação das métricas de desempenho, é necessário abordar o espaço compreendido pela região de interesse da pesquisa.

## 3.5 O espaço analisado

O espaço estudado consiste na região metropolitana que apresenta a segunda maior dinâmica de desenvolvimento econômico do Estado do Paraná: a RM de Londrina. A base de dados utilizada será retirada dos microdados do Censo Demográfico 2010 do IBGE. A área terá os quatro modelos (com e sem efeitos espaciais) descritos a priori para efeito de análise e de comparação: modelo gravitacional tradicional; modelo gravitacional com controle da dinâmica econômica da vizinhança; *XGBoost* e; *XGBoost* com inclusão de uma *dummy* visando o controle da dinâmica econômica da vizinhança.

A RM de Londrina possui 25 municípios, totalizando 600 fluxos, ou seja, fez-se um empilhamento dos municípios para o ano de 2010, pois, como se trata de fluxos entre pares de

municípios, estes dados são empilhados. Espera-se que, após a análise exploratória dos dados, encontre-se uma quantidade significativa de fluxos pendulares dentro da faixa dos cem quilômetros percorridos. Optou-se por analisar essa região em vez de abranger todo o território estadual ou nacional, considerando os resultados apresentados por diversos autores na fundamentação teórica sobre a limitação do modelo gravitacional em estimar fluxos que envolvem distâncias muito extensas. Dessa maneira, considera-se mais apropriado examinar clusters de desenvolvimento individualmente ao lidar com essa modelagem.

Delgado, Moura e Cintra (2013) observam que, com relação aos dados demográficos, a RM de Londrina é a segunda mais populosa do estado, com cerca de 1,1 milhão de habitantes, atrás apenas da RM de Curitiba. Já com relação à pendularidade, a RM de Londrina em 2010 possuia 12,4% do seu total de ocupados trabalhando em outro município sendo que 5,7% residiam nos municípios periféricos e trabalhavam no polo (Londrina), ou seja, Londrina se apresenta como a grande receptora dos pendulares da sua região.

Uma razão adicional para escolher a Região Metropolitana de Londrina em vez de todos os municípios paranaenses é que, ao considerar todos os fluxos entre todos os municípios, entrariam na base de dados fluxos como “Matinhos-Querência do Norte”, os quais provavelmente seriam fluxos zero devido à distância excessiva e à falta de dinamismo econômico entre eles. Isso resultaria em uma quantidade expressiva e predominante de fluxos zero, representando municípios praticamente sem nenhuma relação econômica, prejudicando a estimação. Portanto, a justificativa para a escolha da Região Metropolitana de Londrina baseia-se nas considerações teóricas de que as cidades dentro dessa região estão inseridas em um mesmo *cluster* de desenvolvimento e que, para distâncias superiores a cem quilômetros, os trabalhadores estariam bastante desincentivados a se deslocar.

Após a definição do espaço estudado e das razões para a sua escolha, a próxima seção deve elencar as variáveis utilizadas, bem como o motivo de terem sido selecionadas e a maneira como serão tratadas nas modelagens posteriores.

### 3.6 As variáveis Selecionadas e Fonte dos dados

Nesta seção, serão detalhadas as justificativas para a escolha de cada variável. No entanto, é importante ressaltar inicialmente que alguns pressupostos cruciais no processo de seleção dessas variáveis incluem a facilidade de acesso aos dados e a possibilidade de replicação em diversos espaços. Isso se deve ao fato de que, com a possibilidade de analisar tantos municípios, é imperativo ser parcimonioso na escolha dos dados. Tavares e Monteiro (2019)

destacam que a grande fonte de dados que alimenta a literatura nacional sobre pendularidade é o IBGE através da realização dos Censos Demográficos que são realizados a cada dez anos no território nacional. Alguns autores citados nesta pesquisa (Brito; Silva, 2020; Brito; Ramalho, 2019; Cintra; Delgado; Moura, 2012; Spadon *et al*, 2019; Barbosa, 2015) também utilizam a base de dados do Censo.

### 3.6.1 Taxa de deslocamento pendular

A taxa de deslocamento dos pendulares entre as cidades é a variável de interesse desta pesquisa. É através da relação entre as variáveis subsequentes que se pretende compreender o comportamento dos deslocamentos nas áreas de desenvolvimento estabelecidas aqui. O questionamento sobre onde o indivíduo trabalha é um dado que faz parte do Censo, especificamente nas informações dos microdados das pessoas entrevistadas é possível determinar a quantidade de *commuters* que se deslocam entre os pares de municípios analisados.

É importante salientar que não se trata de uma troca de trabalhadores entre cidades, ou seja, a informação sobre a quantidade de trabalhadores que reside na cidade *x* e que trabalha na cidade *y* não é necessariamente a mesma que a quantidade de trabalhadores que reside em *y* e trabalha em *x* e ambas constam como observações diferentes na base de dados produzida, ou seja, o fluxo de pendulares ao qual esta pesquisa se refere significa apenas o dado sobre a quantidade de trabalhadores que sai de um determinado município até outro e não o dado como trânsito do total de *commuters* entre essas duas cidades.

A maioria dos autores que investiga o Brasil, alguns dos quais foram apresentados *a priori*, utiliza a entrada ou a saída total de trabalhadores pendulares como variável dependente, ou seja, só se sabe a quantidade de trabalhadores que entra ou sai de uma cidade, mas não se sabe a origem ou o destino de cada um deles. Esta pesquisa se diferencia destas pois, após tratamento dos dados colhidos nos microdados do Censo 2010, tem-se as quantidades de trabalhadores pendulares que saem de um mesmo município para todos os outros possíveis, isoladamente. Além disso, em outros países com mais dados disponíveis sobre a pendularidade observa-se o uso das variáveis de distância percorrida e de tempo gasto pelo *commuter*, como evidenciado na pesquisa de Li et al. (2022), que combina essas duas últimas variáveis. Essa combinação é possível para os autores devido à disponibilidade de dados na China. No entanto, nos últimos Censos realizados no Brasil, não foram incluídas perguntas sobre distância percorrida de deslocamento pendular para trabalho. Dessa forma, a variável dependente deste estudo é a taxa de deslocamento pendular entre os pares de municípios. Essa taxa consiste na

razão entre o deslocamento observado de uma cidade a outra, informação retirada dos microdados, e a população ocupada do município de origem.

$$T_{ij} = \frac{t_{ij}}{P_i} \tag{3.6.1}$$

Com $T_{ij}$ a taxa de deslocamento pendular, $t_{ij}$ o fluxo absoluto observado entre o município i e o município j e $P_i$ a população ocupada do município de origem observado, todos em 2010. A população ocupada de um dado município consiste na observação retirada do Censo 2010 de pessoas ocupadas na semana de referência que trabalhavam fora do domicílio e retornavam para seu domicílio diariamente (IBGE, 2010).

A ideia da criação dessa variável é normalizar os dados em relação ao tamanho do mercado de trabalho das cidades em termos de população ocupada, excluindo quem não se desloca de maneira alguma para trabalhar. Isso permite comparar os deslocamentos entre municípios de maneira mais precisa. Por exemplo, uma dada quantidade de movimentos pendulares a partir de uma cidade com 100 mil trabalhadores pode não ser tão relevante quanto um fluxo semelhante originário de uma cidade com 30 mil trabalhadores; porém, caso não houvesse tal tratamento na variável, os modelos considerariam que ambos os municípios sofrem com a saída de trabalhadores de maneira idêntica. A taxa de deslocamento fornece uma melhor ideia da intensidade relativa do movimento pendular. Além disso, as taxas podem fornecer uma interpretação mais direta. Por exemplo, uma taxa de deslocamento de 0,02 pode ser interpretada como "2% da população ocupada do município de origem se desloca para outro município". Isso é mais fácil de entender do que um número absoluto, que precisa ser contextualizado com informações adicionais sobre a população ocupada de origem.

A justificativa para a utilização dos dados de 2010 é que, até o presente momento, os microdados do Censo 2022 não foram publicados, com previsão para divulgação somente em setembro de 2024 (IBGE, 2024). Vale ressaltar que os últimos três censos vêm apresentando uma atenção crescente com a pendularidade ao aplicarem, a cada nova edição, questionários com cada vez mais detalhes a respeito deste tema. Novos dados devem ser gerados para a comunidade científica e, com a chegada do Censo 2022, esta tendência deve se manter, pois consta no seu questionário de amostra a informação adicional sobre o tipo de meio de transporte utilizado para se deslocar (IBGE, 2022), possibilitando, no futuro, a reprodução das ferramentas desta pesquisa com informações fidedignas em relação ao panorama da pendularidade atual da RM de Londrina ou de quaisquer outros municípios brasileiros.

### 3.6.2 Índice Firjan – Emprego e Renda (IFDM-ER)

Neste estudo, uma das variáveis explicativas se refere ao IFDM-ER para os municípios da Região Metropolitana de Londrina, utilizado como proxy para o dinamismo de mercado de trabalho. Este dado é fornecido pela Fundação das Indústrias do Rio de Janeiro (FIRJAN), utilizando estatísticas públicas oficiais, e monitora três dimensões: emprego e renda, educação e saúde (Firjan, 2018). Entretanto, para atender às expectativas do estudo, somente o índice IFDM – Emprego e Renda (IFDM-ER) foi utilizado, excluindo as dimensões sociais.

O aspecto econômico do IFDM, conforme a Firjan (2018), abrange a geração de emprego formal, a taxa de formalização do mercado de trabalho, a geração de renda, a massa salarial real no mercado de trabalho formal e o índice de *Gini* de desigualdade de renda no trabalho formal. Ou seja, capta a dinâmica, especialmente do mercado de trabalho formal. O índice varia de 0 a 1, sendo que valores próximos de 1 indicam maior desenvolvimento econômico do município, enquanto valores próximos de 0 indicam menor desenvolvimento. A escala determinada para o IFDM – Emprego e Renda apresenta quatro conceitos (Firjan, 2018):

i) Cidades com baixo desenvolvimento econômico: IFDM entre 0,0 e 0,4

ii) Cidades com desenvolvimento econômico regular: IFDM entre 0,4 e 0,6

iii) Cidades com desenvolvimento econômico moderado: IFDM entre 0,6 e 0,8

iv) Cidades com alto desenvolvimento econômico: IFDM entre 0,8 e 1,0

Considerando que o IFDM abrange aspectos econômicos mais realistas do mercado de trabalho de um município, a hipótese subjacente a essa desagregação é que os trabalhadores, ao se deslocarem para outro município, não consideram amenidades sociais, como lazer, saúde e educação. Em vez disso, focalizam especialmente no ganho de renda proporcionado pelo deslocamento, como exemplificado por Stam (2013) que, ao analisar os determinantes que influenciam o *commuter*, chega à conclusão de que o fator primário e dominante da pendularidade tem natureza econômica. Caso esses fatores sociais fossem relevantes para os trabalhadores, a decisão provavelmente seria migrar. Portanto, apenas o aspecto econômico do município é crucial para determinar o deslocamento de pendulares nesta pesquisa, dado que a

amostra da pesquisa contém somente os entrevistados que informaram realizar o movimento pendular.

O IFDM-ER utilizado é referente ao ano de 2009, o que implica que a variável está defasada temporalmente em relação aos fluxos, que são do ano de 2010. Essa escolha procede por razões teóricas e econométricas: teoricamente, presume-se que um indivíduo que decide se deslocar para trabalhar em outro município já residia naquela cidade e, portanto, utiliza informações passadas sobre o nível de desenvolvimento entre os pares de municípios para tomar sua decisão. A justificativa econométrica visa minimizar problemas de endogeneidade, ou seja, ao utilizar uma variável independente defasada, assume-se que o valor passado da variável não é afetado pela variável dependente contemporânea, rompendo uma possível causalidade recíproca entre ambas e auxiliando a isolar a direção da causalidade desejada no modelo.

Um indicador com características semelhantes ao IFDM – ER seria o Índice de Desenvolvimento Humano Municipal desagregado para a renda (IDHM-R), disponibilizado pelo Programa das Nações Unidas para o Desenvolvimento (PNUD). Brito e Silva (2020) utilizam a dimensão da renda do IDHM, além de outros indicadores socioeconômicos, para captar as amenidades urbanas no seu modelo e Brito e Ramalho (2019) também incluem o IDHM; porém, com todas suas dimensões como variável explicativa e defasado temporalmente pois buscam analisar os determinantes socioeconômicos da pendularidade. Apesar de ser bastante utilizada por outros pesquisadores, esta variável foi preterida porque foi computada no Censo de 2000 e no Censo de 2010 e, como é necessário defasar temporalmente a variável, a utilização de uma observação do ano 2000 representaria uma defasagem muito alta.

Isto posto, se o problema é encontrar uma variável que esteja disponível em algum ano mais recente e que represente aspectos econômicos dos municípios, por que não usar o PIB (Produto Interno Bruto) municipal? A resposta a essa indagação reside na própria composição do IFDM-ER mencionado anteriormente. Este índice engloba dados que capturam de maneira mais eficaz os efeitos relacionados ao mercado de trabalho, como geração de emprego e desigualdade salarial. Dado que a pendularidade está intrinsecamente ligada ao mercado de trabalho, o IFDM-ER se revela mais apropriado do que o Produto Interno Bruto (PIB) para tal análise. O PIB, por sua vez, é menos eficiente na tarefa de revelar os dinamismos do mercado de trabalho, uma vez que representa a soma de todos os bens e serviços finais produzidos pelo município, podendo este estar concentrado, sem oferecer insights sobre a distribuição dessa riqueza.

Para ilustrar, um município com um PIB elevado, mas com elevada concentração de renda e baixa utilização de mão-de-obra, pode ser equivocadamente rotulado como receptor de

pendulares. Em contrapartida, outra cidade com um PIB menor, mas com um dinamismo mais significativo em termos de renda e emprego, pode ser mais atrativa para os trabalhadores de outros municípios.

Assim, o IFDM-ER se mostra mais adequado, pois, além de estar disponível para qualquer município brasileiro, incorpora os aspectos econômicos pertinentes ao mercado de trabalho, que são considerados verdadeiramente relevantes. Isso é feito sem a necessidade de introduzir uma quantidade excessiva de variáveis, o que contribui para mitigar o problema da heterogeneidade não observada.

É importante destacar que a utilização dessa variável, mesmo que aparentemente mais eficiente do que o PIB, o IDH e até mesmo o próprio índice Firjan com todas as suas dimensões, ainda assim apresenta limitações uma vez que não consegue absorver totalmente a desigualdade de cada município. Por exemplo, um município com alto IFDM-ER pode possuir determinadas áreas ou grupos populacionais que destoam do valor observado em termos de dinamismo do mercado de trabalho. Para amenizar este problema, mais variáveis poderiam ser incluídas, porém, dada a característica do modelo gravitacional, isto só poderia ser feito no *XGBoost*, então, como os dois modelos devem ser estimados com as mesmas variáveis, constitui-se assim uma limitação não somente do IFDM-ER, mas, também da própria modelagem proposta.

### 3.6.3 Distância entre os pares de municípios.

A conclusão mais recorrente entre os diversos estudos que modelam o fenômeno da pendularidade é a evidente influência da distância entre duas regiões quaisquer em seus fluxos pendulares. Essa constatação é tão consistente em várias pesquisas, abrangendo diversas localidades e utilizando diversas abordagens metodológicas, que se consolidou como um consenso: a distância apresenta uma relação inversa com o nível de pendularidade.

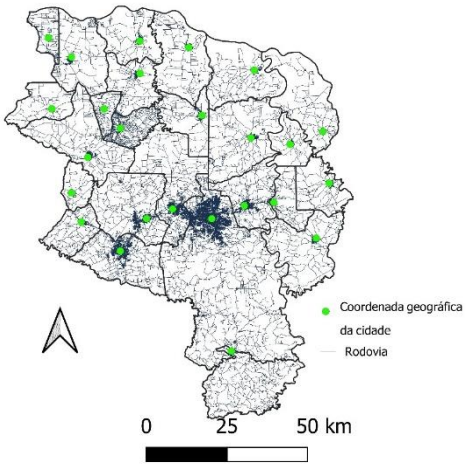
A utilização dessa variável decorre da necessidade de quantificar os custos associados ao deslocamento pendular no modelo. Esses custos dependem de uma série de fatores que representam características heterogêneas, tornando desafiadora a tarefa de mensurar e definir uma base de dados adequada. Por exemplo, entre alguns municípios, pode haver a opção de deslocamento por trem, enquanto em outras pode haver cobrança de pedágio. Além disso, a existência de cobertura por aplicativos de carona, as características das rodovias (simples ou duplicadas) são outros elementos a serem considerados. Há também trabalhadores cujo custo de transporte é pago pela empresa, por exemplo, no caso de ônibus fretados diariamente para transportar os funcionários de uma determinada empresa dentro da RM. Coletar dados para

todas essas particularidades da mobilidade intermunicipal seria uma tarefa inviável, considerando a premissa de que o modelo deve ser facilmente replicado para outras regiões. Portanto, como há diversas variáveis que afetam o custo de deslocamento, a distância em quilômetros entre os pares de municípios é utilizada como uma *proxy* para capturar o potencial efeito dos custos de deslocamento.

Diversos estudos corroboram essa abordagem. Brito e Silva (2022) empregam a distância até a capital estadual como *proxy* para o nível de acessibilidade e os custos de transporte em direção às regiões centrais. Brito e Ramalho (2019) utilizam a distância euclidiana entre os centróides dos municípios como *proxy* para os custos de movimentação pendular em seu modelo gravitacional. Spadon et al. (2019) apontam que a distância é o fator mais importante para prever o fluxo de *commuters* entre pares de municípios, enquanto Abouelhamd (2021) afirma que a proximidade das regiões centrais de negócios reduz o custo e o tempo do deslocamento pendular.

No contexto desta pesquisa, as coordenadas geográficas de cada município foram coletadas através das malhas territoriais disponibilizadas pelo IBGE. Importante destacar que a localização dessas coordenadas não coincide com os centroides dos municípios, mas sim com as regiões de maior urbanização de cada território municipal. Por exemplo, no município de Londrina, a coordenada fornecida pelo IBGE refere-se à região mais urbanizada da cidade, o que pode constituir uma vantagem em relação ao uso de centroides, uma vez que estes podem se localizar muito distantes da região de moradia e/ou trabalho dos *commuters* e fazer com que se encontre resultados ainda mais distantes da realidade. A localização dos pontos fornecidos pelo IBGE sobre a urbanização de cada município está ilustrada na figura 1.

Figura 1: Mapa de rodovias e de coordenadas geográficas da RM de Londrina.




Fonte: Elaborado pelo autor.

É possível perceber que os pontos estão localizados sobre a região com maior concentração de malha rodoviária de cada cidade, o que leva à conclusão anterior de que os pontos se localizam nas regiões urbanas e mais desenvolvidas dos municípios.

A métrica utilizada para calcular as distâncias é a fórmula de *Haversine*, que leva em consideração o raio da Terra e utiliza trigonometria esférica para calcular a distância mais curta entre dois pontos na superfície de uma esfera (P. Dauni *et al.*, 2019).

$$Distância = 2\pi * arcsen(\sqrt{sen^2\left(\frac{\Delta\varphi}{2}\right) + \cos(\varphi_1) * \cos(\varphi_2) * sen^2\left(\frac{\Delta\gamma}{2}\right)}) \quad (3.6.2)$$

Onde: $\varphi_1$ e $\varphi_2$ são as latitudes dois pontos em radianos, $\Delta\varphi$ é a diferença entre as latitudes, $\Delta\gamma$ é a diferença entre as longitudes e $\pi$ é o raio da Terra.

Como dito anteriormente, uma limitação enfrentada no uso desta variável que merece destaque é que não se considera o fato de que há trabalhadores que não ponderam o custo do deslocamento para decidir realizar o deslocamento pendular, uma vez que as firmas que o empregam já arcam com essas despesas, como, por exemplo, por meio de veículos fretados por elas. Dessa forma, estes trabalhadores seriam indiferentes entre realizar ou não o deslocamento pendular, tendo como único fator de decisão o ganho de renda. Para eles, a distância não representa o custo de deslocamento. Como não há dados no período para modelar essa assimetria e os modelos também não conseguem capturá-la, este elemento constitui mais um limitante desta pesquisa.

### 3.6.4 Resumo da Estratégia Empírica utilizada nas Estimações

Após a descrição de todos os procedimentos metodológicos, apresenta-se abaixo a Tabela 2, na qual tem-se um resumo das variáveis explicativas selecionadas e suas nomenclaturas (em parênteses). Ressalta-se que todas as estimativas foram efetuadas no software R.

Tabela 2: Descrição das variáveis explicativas 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 pelo autor.

Produziu-se um total de seis modelos com estas variáveis, cujos resultados são apresentados no próximo capítulo, a saber: três experimentos com modelo gravitacional e três com modelo *XGBoost*, alterando apenas a dimensão do efeito espacial entre eles. A tabela 3 resume as modelagens:

Tabela 3: Descrição dos experimentos econométricos.

| Modelos | Variável dependente (alvo) | Variáveis explicativas (características) |
|---|---|---|
| | 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 pelo autor.

Dessa maneira, foi possível comparar os modelos gravitacionais entre si em termos de influência das novas variáveis de efeito espacial. A mesma comparação foi realizada para o modelo *XGBoost*, havendo também uma comparação cruzada entre ambas as modelagens.

Foram realizadas estimações MQO, GPML e PPML na equação gravitacional para se certificar de que os modelos PPML são bem especificados[8]. Além disso, o teste MaMu de log-normalidade dos resíduos foi utilizado para verificar se a estimação PPML era realmente a mais adequada (Head; Mayer, 2014). Aplicou-se logaritmo nos dados, a priori, para a estimação

[8] Se os modelos produzem coeficientes idênticos, não haverá problemas de heterocedasticidade nem de especificação (Head; Mayer, 2014). Ver apêndice com as estimações MQO e GPML.

MQO, porém, para as estimações PPML e GPML além do modelo *XGBoost*, este procedimento não é necessário dada as características dos modelos, então foram utilizadas as variáveis originais da base de dados. 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 acima 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*, esses parâmetros 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. O próximo capítulo apresenta a discussão dos resultados dos experimentos detalhados.

# 4. PENDULARIDADE NA RML: EVIDÊNCIA EMPÍRICA DOS SEUS DETERMINANTES

Este capítulo apresenta a análise dos resultados. Inicialmente, por meio da exploração dos dados da amostra, analisou a dinâmica dos fluxos pendulares entre os municípios e como as variáveis selecionadas se relacionam para influenciar esses deslocamentos. Já na segunda seção, tem-se a análise dos resultados das análises econométricas propostas, observando como cada variável age sobre os deslocamentos pendulares, o desempenho dos modelos nas métricas estabelecidas e a relevância dos efeitos espaciais para a análise.

## 4.1 Os movimentos pendulares na RM de Londrina

O Estatuto da Metrópole, instituído através da Lei nº. 13.089/2015, determina que as regiões metropolitanas brasileiras devem produzir seus Planos de Desenvolvimento Urbano Integrados (PDUI's). No caso da RML, este foi produzido utilizando os microdados do Censo 2010, os Planos Diretores municipais e diversos outros estudos de apoio para abordar, dentre outros assuntos metropolitanos, a pendularidade (SEDU, 2022). Dessa forma, algumas informações a respeito dos Planos Diretores das cidades da RM de Londrina (RML) e alguns diagnósticos ratificam os resultados encontrados durante a análise descritiva dos dados desta pesquisa.

A RML foi criada em 1998 na 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 agora possui 25 municípios incluindo os supracitados, com população combinada de 1.000.062 habitantes e, destes, 378.216 pessoas declararam que trabalham fora dos seus domicílios e retornam diariamente, de acordo com os microdados coletados no Censo 2010.

Na tabela 4 se encontra o panorama da RML em relação às populações ocupadas de cada cidade, o índice Firjan na dimensão de emprego e renda, além da entrada e da saída total de trabalhadores pendulares de cada município. Através desta tabela é possível ter uma ideia inicial do fluxo líquido de cada cidade, ou seja, quais, após a realização dos deslocamentos, são majoritariamente doadoras ou receptoras de trabalhadores.

Tabela 4: Características selecionadas dos municípios da RML - 2010

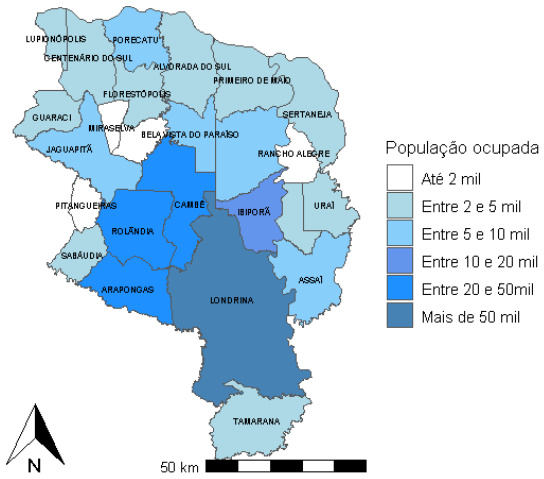
| Municípios | População ocupada | IFDM-ER | Entrada de *commuters* | Saída de *commuters* | Taxa de entrada | Taxa de saída |
|---|---|---|---|---|---|---|
| Alvorada do Sul | 3368 | 0,43 | 13 | 58 | 0.38% | 1.72% |
| Arapongas | 42088 | 0,86 | 231 | 60 | 0.54% | 0.14% |
| Assaí | 5786 | 0,63 | 15 | 52 | 0.25% | 0.89% |
| Bela Vista do Paraíso | 6062 | 0,56 | 17 | 235 | 0.28% | 3.87% |
| Cambé | 41617 | 0,76 | 304 | 1614 | 0.73% | 3.87% |
| Centenário do Sul | 4162 | 0,49 | 18 | 199 | 0.43% | 4.78% |
| Florestópolis | 4482 | 0,79 | 176 | 200 | 3.92% | 4.46% |
| Guaraci | 2408 | 0,53 | 7 | 63 | 0.29% | 2.61% |
| Ibiporã | 16486 | 0,72 | 202 | 505 | 1.22% | 3.06% |
| Jaguapitã | 5051 | 0,81 | 266 | 48 | 5.26% | 0.95% |
| Jataizinho | 4193 | 0,52 | 32 | 348 | 0.76% | 8.29% |
| Londrina | 184480 | 0,83 | 2656 | 324 | 1.43% | 0.17% |
| Lupionópolis | 2080 | 0,57 | 3 | 39 | 0.14% | 1.87% |
| Miraselva | 713 | 0,52 | 8 | 127 | 1.12% | 17.81% |
| Pitangueiras | 1401 | 0,56 | 11 | 87 | 0.78% | 6.20% |
| Porecatu | 5482 | 0,60 | 240 | 125 | 4.37% | 2.28% |
| Prado Ferreira | 1285 | 0,66 | 12 | 102 | 0.93% | 7.93% |
| Primeiro de Maio | 4183 | 0,41 | 12 | 47 | 0.28% | 1.12% |
| Rancho Alegre | 1348 | 0,44 | 9 | 64 | 0.66% | 4.74% |
| Rolândia | 22160 | 0,91 | 526 | 230 | 2.37% | 1.03% |
| Sabáudia | 2927 | 0,79 | 30 | 177 | 1.02% | 6.04% |
| Sertaneja | 2263 | 0,48 | 15 | 29 | 0.66% | 1.28% |
| Sertanópolis | 6527 | 0,69 | 49 | 69 | 0.75% | 1.05% |
| Tamarana | 3375 | 0,56 | 17 | 48 | 0.50% | 1.42% |
| Uraí | 4389 | 0,59 | 18 | 37 | 0.41% | 0.84% |

Fonte: Elaborado pelo autor.

Inicialmente, pode-se perceber que se trata de uma RM com municípios, em sua maioria, de pequeno porte, o que afeta quantidade observada de pessoas ocupadas. São 20 municípios, dos 25, com menos de 20 mil habitantes. Londrina surge como um *outlier* nas observações e os municípios que possuem mais de 20 mil habitantes (Arapongas, Cambé, Ibiporã e Rolândia) fazem necessariamente fronteira com a cidade-polo.

A figura 2 ilustra toda a dinâmica das populações ocupadas de cada município em termos de localização geográfica:

Figura 2: Disposição espacial dos municípios da RML por quantidade de trabalhadores ocupados




Fonte: Elaborado pelo autor.

Grande parte da população ocupada da RML se concentra ao redor da cidade-polo, sobretudo, na região de conurbação[9] entre o polo Londrina e as cidades de Cambé e Ibiporã. Porecatu apresenta uma quantidade maior de população ocupada em relação aos outros municípios do noroeste da região. Além disso, outras duas cidades com maior quantidade de população ocupada da região, Rolândia e Arapongas, se encontram em processo de conurbação, além de Jataizinho (SEDU,2022). Este último processo pode ser observado tanto na figura 2 quanto na figura 3 e evidencia o importante papel da Rodovia Federal BR-369 como um eixo de ocupação que perpassa esta região.

Figura 3: Mapa do processo de conurbação na RML

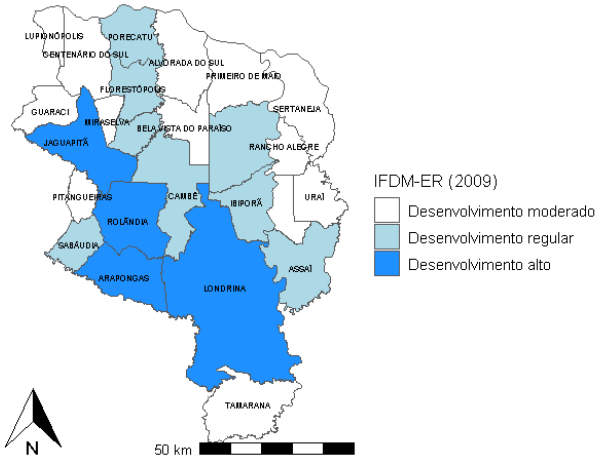
Fonte: SEDU, 2022 (Adaptado pelo autor).

[9] 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.

A mancha escura no mapa representa os núcleos urbanos das cidades, que coincidem com as coordenadas geográficas utilizadas na pesquisa e, observando a localização dos núcleos de Londrina, Cambé, Ibiporã, Rolândia, Arapongas e Jataizinho, é possível notar o processo de conurbação em andamento.

Sobre o índice Firjan – Emprego e Renda - dos municípios, percebe-se que, de acordo com a métrica estabelecida pela Firjan, não há municípios com baixo nível de desenvolvimento em termos de dinamismo econômico (com enfoque no mercado de trabalho), ou seja, todos possuem um nível, pelo menos, moderado, como se pode observar na figura 4.

Figura 4: Mapa dos índices Firjan - Emprego e Renda (IFDM-ER) da RML




Fonte: Elaborado pelo autor.

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, já possuem indicadores melhores em relação aos seus pares de mesma característica populacional, principalmente no município de Porecatu, que vai dando indícios de possuir uma dinâmica de fluxo própria.

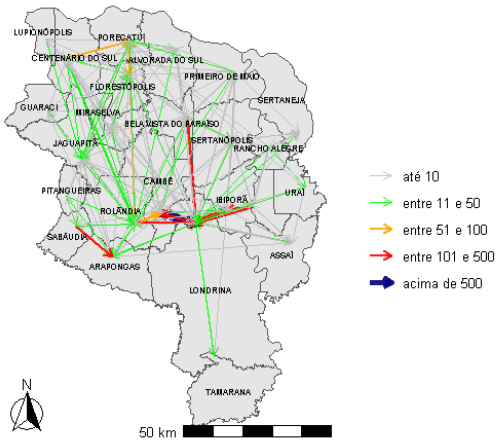
A cidade com o menor Índice Firjan de Desenvolvimento Municipal com Enfoque em Renda (IFDM-ER) é Primeiro de Maio, situada ao norte da Região Metropolitana de Londrina (RML). Mais uma vez, constata-se uma tendência de indicadores mais elevados na área central onde estão localizadas as cidades-polo e suas adjacências. No entanto, destaca-se que Rolândia apresenta o índice mais alto da região, superando inclusive Londrina e a até mesmo a capital paranaense, o que pode ser explicado, de certa maneira, pelo processo de conurbação em que este município está inserido, bem como pelo fluxo alto fluxo observado de pendulares em relação aos outros municípios da região, seja de entrada ou de saída, que remetem a um maior

nível de desenvolvimento em potencial. Ademais, verifica-se um patamar superior de IFDM-ER nas cidades a oeste, em comparação com aquelas situadas a leste, o que pode sugerir uma dinâmica mais intensa entre a RML e as Regiões Metropolitanas de Maringá e de Apucarana[10].

Esta suposição é reforçada pela SEDU (2022) ao sugerir a inserção do município de Apucarana na RML e alertar para a intensificação de movimentos pendulares (nesse caso, não somente por motivo de trabalho) fora do recorte regional de Londrina. Além disso, de acordo com o Plano Diretor de Pitangueiras, a cidade, apesar de integrar a RML, possui mais dinamismo com Apucarana do que com Londrina (SEDU, 2022). Isto corrobora com a hipótese de que há um efeito que não está sendo captado, em decorrência desta proximidade com outras Regiões Metropolitanas do Estado.

Ao observar as colunas de entrada e de saída de trabalhadores pendulares da tabela 4 é possível perceber que das 25 cidades, apenas cinco delas são receptoras líquidas de trabalhadores: Arapongas, Jaguapitã, Londrina, Porecatu e Rolândia. Isso indica, novamente, a força de atração exercida pela cidade-polo e pelas cidades que possuem dinâmica própria dentro da RML, como pode ser percebido também através da figura 5, que ilustra os fluxos de *commuters* entre todas as cidades.

Figura 5: Mapa de deslocamentos pendulares da RML




Fonte: Elaborado pelo autor.

A magnitude dos deslocamentos parece convergir com as populações e as dinâmicas de mercado de trabalho vistas nas outras figuras. É evidente a grande concentração de

[10] O Paraná possui 8 RM's, porém, no norte do estado se encontram, além da RML, as RM's de Maringá e de Apucarana. As três cidades-polo dessas regiões estão a, no máximo, 100 quilômetros uma da outra. Essa proximidade pode ser um indicativo de transbordamento do desenvolvimento regional de forma que os fluxos pendulares da RML podem ser afetados pelas outras RM's.

deslocamentos que ocorrem na região de conurbação e como a cidade-polo recebe trabalhadores pendulares de praticamente todos os municípios, deixando mais clara a sua força de atração de *commuters* logo em um primeiro momento da análise.

As cidades de Rolândia, Bela Vista do Paraíso, Jataizinho, Ibiporã e Cambé são as maiores responsáveis pelos 2656 trabalhadores que se deslocam até Londrina, com destaque para Cambé, a maior doadora de *commuters* da amostra, doando 1497 de seus trabalhadores somente para a cidade-polo. Esta cidade provavelmente envia tantos trabalhadores para lá pois sua conurbação já torna uma tarefa difícil a distinção da fronteira entre ambos os municípios (SEDU,2022), o que torna ainda mais forte o poder de atração de pendulares. Já Londrina, enquanto é a maior receptora de *commuters*, é apenas a quarta maior doadora em termos absolutos e envia quase 70% de seus trabalhadores pendulares às duas cidades conurbadas, Cambé e Ibiporã.

É interessante observar outras dinâmicas de pendularidade mais afastadas e que não envolvem a cidade-polo. O noroeste da RML possui dinâmica própria, com Porecatu, Centenário do Sul e Florestópolis trocando trabalhadores pendulares entre si, mas, com destaque para Porecatu que se apresenta como a maior receptora de pendulares dessa área do mapa. Outras dinâmicas que merecem destaque são os movimentos pendulares de Jataizinho a Ibiporã, Sabáudia a Arapongas e Rolândia a Cambé. Por último, Rolândia merece atenção por ser a segunda maior receptora de pendulares da RML, com fluxos mais intensos vindo de municípios como Cambé, Florestópolis e Centenário do Sul. Esses movimentos pendulares que não envolvem Londrina possivelmente ocorrem dado algum tipo de demanda por mão de obra específica desses municípios através, por exemplo, da presença de algum tipo de indústria ou de negócio que faça com que aquela cidade exerça uma força de atração de pendulares suficientemente maior do que a força de atração proveniente da cidade-polo, bem como pela redução da força desta dada a distância maior, uma vez que os trabalhadores que residem, por exemplo, próximo à Porecatu, terão maior custo de deslocamento até Londrina e, assim, criem uma nova dinâmica nesses municípios do norte da RML.

Dessa forma, percebe-se que municípios mais afastados e com características mais modestas, em termos de população ocupada e de mercado de trabalho, também apresentam poucos fluxos, seja de entrada ou de saída, enquanto municípios próximos com maior mercado de trabalho e população ocupada tendem a gerar mais fluxos. Isso sugere que a distância desempenha um papel de repulsão para os deslocamentos dos trabalhadores, enquanto as dinâmicas do mercado de trabalho exercem influência sobre esses movimentos. Assim, resta

determinar de que maneira essas dinâmicas afetam os deslocamentos e se, além da distância geográfica, há algum efeito espacial que necessita ser considerado na análise.

Ainda sobre a distância, esta parece se apresentar novamente como uma forte influência para o sentido e a magnitude dos deslocamentos pendulares, uma vez que se observa sempre maiores fluxos entre cidades próximas, mesmo que não estejam próximas da cidade-polo, como no caso de Porecatu-Florestópolis, ou Sabáudia-Arapongas. De acordo com os Planos Diretores catalogados na pesquisa da SEDU (2022), a distância, quando mais reduzida, é o principal fator responsável pelo aumento da dinâmica entre os municípios da RML e isso pode ser observado através desta análise descritiva.

É possível perceber que cidades não somente próximas, mas, fronteiriças, tendem a apresentar os maiores movimentos de trabalhadores entre si, seja de entrada ou de saída. Então, além da distância geográfica, o efeito espacial pode estar inserido neste aspecto da localização das cidades bem como pelo fato de não somente ser um vizinho, mas por ter uma característica de mercado de trabalho mais desenvolvida, algo que se supôs também ser um fator de atração de trabalhadores pendulares. Na seção seguinte, os resultados da estimação dos modelos e a comparação das métricas de desempenho são apresentadas.

## 4.2 Resultados dos modelos gravitacional e *XGBoost*: Determinantes da Pendularidade da RML

Antes de iniciar a apresentação dos resultados, vale destacar que em todos os modelos propostos a amostra permaneceu a mesma, com separação de dados para treino e teste, com 70% da amostra, ou seja, 420 observações, dedicada para treino dos modelos gravitacional e *XGBoost* e o restante para testar a capacidade preditiva. Toda a análise econométrica foi desenvolvida em linguagem *R*, no software *Rstudio*.

Na Tabela 5, encontram-se três modelos gravitacionais: o modelo (G1), sem variáveis de efeitos espaciais estimado por PPML; o modelo (G2) do tipo SLX, sendo um gravitacional que inclui a variável *dummy* de contiguidade para captar efeitos espaciais e; o modelo (G3), que inclui uma variável explicativa para capturar os efeitos da dinâmica econômica da vizinhança, ou seja, a variável de interação que multiplica a *dummy* de contiguidade pelas observações do IFDM-ER do município vizinho. Além disso, o valor do teste MaMu dos três modelos também é apresentado nessa tabela.

É importante destacar que as estimações do modelo gravitacional por MQO e por GPML também foram realizadas[11] e, em todas elas, o sentido e a significância das estimativas são idênticos, o que indica que os dados e os modelos não sofrem de problemas de heterocedasticidade. Além disso, de acordo com toda a base teórica detalhada no capítulo 3 e no fato de que o estimador PPML não requer nenhuma suposição adicional (além da especificação em 3.1) a respeito da distribuição da variável dependente para ser consistente (Wooldridge, 2013), o modelo parece estar bem especificado.

Tabela 5: Coeficientes do modelo gravitacional PPML – modelos (G1), (G2) e (G3)

| 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* | - |
| Teste MaMu ($\lambda$) | 1,95* | 1,97* | 1,96* |

Fonte: Elaborado pelo autor.
(*) estatisticamente significativo a 1%

Observa-se inicialmente na Tabela 5 que o teste MaMu para os três experimentos indicou que a estimação PPML é de fato a mais adequada. Além disso, os três modelos gravitacionais apresentam estimativas significativas e convergentes em suas conclusões. Por exemplo, 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, isto significa,

[11] Tabelas 7 e 8 no apêndice apresentam as estimações do modelo gravitacional por MQO e por GPML.

primeiramente, que existe um efeito espacial de vizinhança. Ou seja, ser um município fronteiriço a outro eleva a força de atração de trabalhadores pendulares entre essas cidades. Esse resultado condiz não somente com os fluxos observados entre as cidades conurbadas, mas com as outras cidades vizinhas no noroeste e no oeste da RML, como Porecatu-Florestópolis e Sabáudia-Arapongas.

Sobre o último modelo gravitacional, a variável de interação, do tipo SLX, também foi positiva e significativa, um resultado esperado. Nesse contexto, o efeito espacial que influencia a atração entre dois municípios não se limita apenas à proximidade geográfica, mas também está associado à qualidade do mercado de trabalho do envoltório. De fato, ser um município vizinho com características laborais mais favoráveis aumenta 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.

Agora sobre o modelo *XGBoost*, sua aplicação também produziu três modelos: 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.

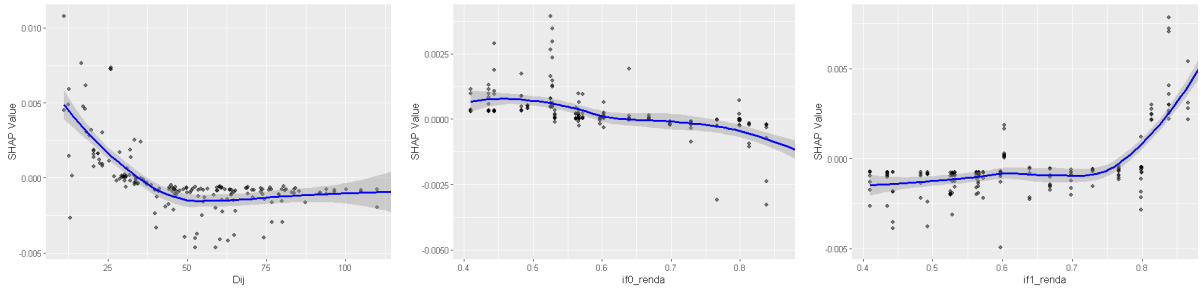
Dessa forma, nos três modelos, o método padrão do software com escolha manual de iterações foi superior: no caso do modelo sem efeitos espaciais (X1) foram escolhidas 22 iterações; no modelo com *dummy* de contiguidade (X2) foram escolhidas 23 iterações e; no

modelo com variável de interação (X3), 27 iterações. Vale ressaltar que, assim como o procedimento adotado para os modelos gravitacionais, os gráficos gerados para todos os modelos *XGBoost* daqui em diante são resultado da previsão realizada nos dados de teste referentes aos 30% da amostra inicial (180 observações).

Modelos de aprendizado de máquina como o *XGBoost* possuem a desvantagem da dificuldade de interpretação e, para contornar este obstáculo, utilizou-se a ferramenta SHAP, que estimou as contribuições das características à previsão da variável alvo, resultando na figura 6 para o modelo sem efeitos espaciais (X1):

Figura 6: Valores SHAP - modelo (X1)

(a) (b) (c)




Fonte: Elaborado pelo autor

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, é fácil perceber que se trata de relações não lineares. Na figura 6(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 6(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 a cidade em que vive apresenta baixa dinâmica de mercado de trabalho em relação aos outros.

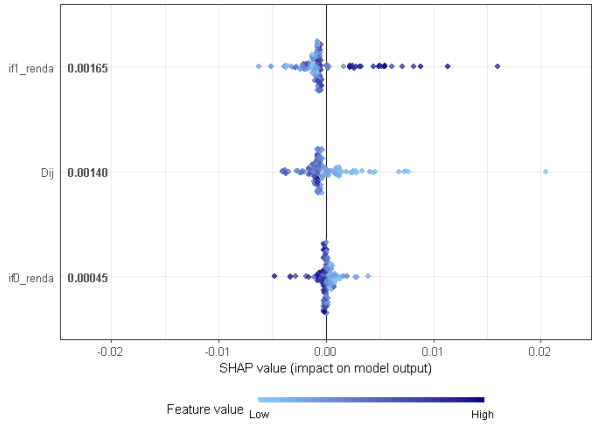
Se considerar que a Região Metropolitana de Londrina tem 15 municípios com um IFDM-ER inferior a 0,65, é possível inferir que a maioria desses municípios, correspondendo a 60% do total, demonstra indícios de um fenômeno de força centrífuga, no qual trabalhadores são repelidos para outras localidades. 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.

Sobre a Figura 6(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[12], ele opta por realizar o movimento. Em outras palavras, a cidade 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.

Com efeito, ao analisar a Tabela 4, observa-se que apenas seis municípios da RML apresentaram IFDM-ER superior a 0,77. Curiosamente, esses seis municípios foram os que registraram as maiores entradas de trabalhadores pendulares no ano de 2010.

Na figura 7, ainda sobre os valores SHAP do modelo (X1), é possível inferir mais algumas conclusões a respeito do modelo.

Figura 7: Médias dos valores SHAP absolutos e impacto na variável explicada - modelo (X1).




Fonte: Elaborado pelo autor.

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

Na Figura 6, obtêm-se explicações tanto locais quanto globais dos valores SHAP. Na Figura 7, além dessas interpretações, 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. Essa abordagem permite uma compreensão mais aprofundada das contribuições individuais de cada variável para as previsões do modelo.

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.

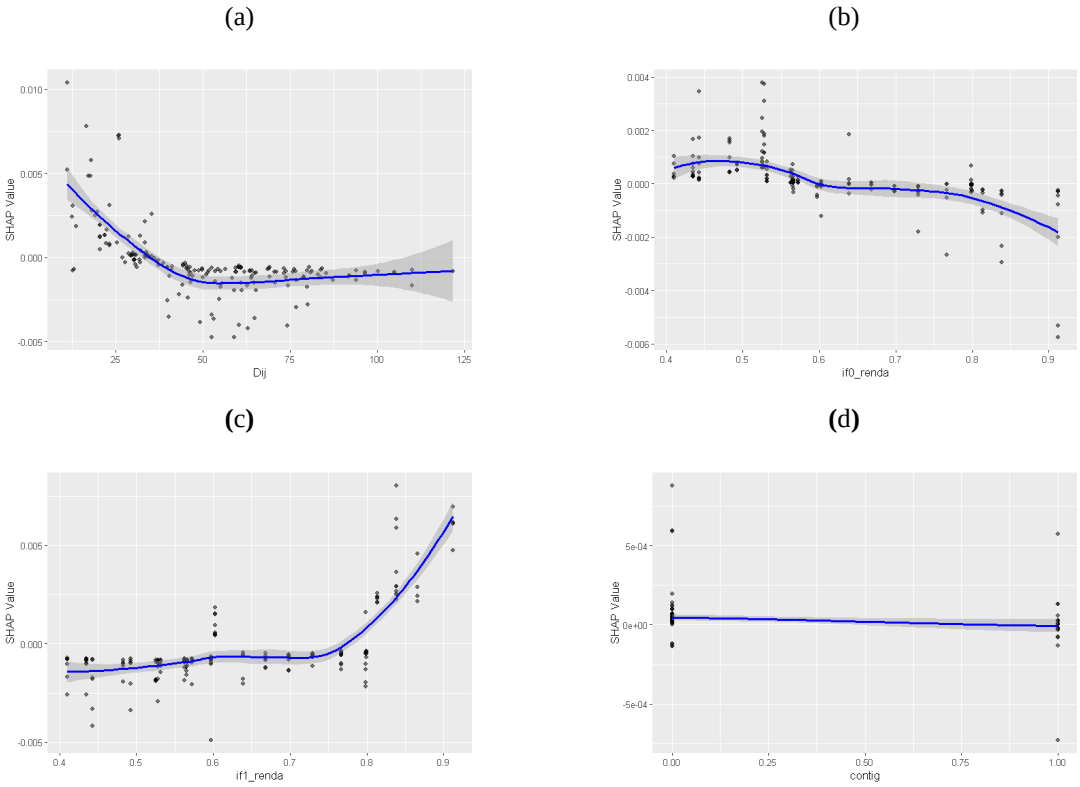
Os resultados do modelo (X1) apontam o seguinte:

- ✓ If1_renda: Primeiramente, 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, à 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 dessa vez 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 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.

Agora, no modelo (X2) a variável de contiguidade foi incorporada e a figura 8 mostra os valores SHAP correspondentes:

Figura 8: Valores SHAP - modelo (X2)




Fonte: Elaborado pelo autor.

Novamente as curvas representam relações não lineares. A figura 8(a) mostra que uma distância de aproximadamente 35km representa o ponto de inflexão entre o aumento e a redução dos deslocamentos pendulares. Para distâncias maiores, os movimentos se reduzem. Este gráfico do modelo (X1) apresenta efeitos idênticos ao mesmo gráfico referente ao modelo (X2).

Na Figura 8(b), os deslocamentos pendulares, diferentemente do primeiro modelo, passam cair a partir um IFDM-ER do município de origem de aproximadamente 0,6. Já na figura 8(c), o ponto de mudança entre aumento e redução de deslocamentos pendulares volta a ser idêntico ao seu equivalente do modelo (X1).

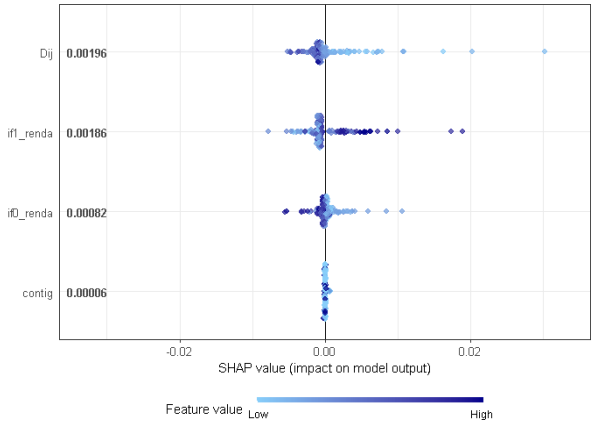
Por último, a variável de contiguidade na figura 8(d) merece um pouco mais de atenção. Como se trata de uma *dummy* os valores das observações só podem assumir 0 ou 1. Pode-se perceber que, dentre os valores 0 das observações, a maioria possui valor SHAP positivo, isto significa que municípios que não 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 do gráfico, apesar de a 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 de 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 a cidade-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.

Apesar disso, o efeito da variável de contiguidade não parece ser tão claro e não se sabe ao certo se a média do valor SHAP dela representa um efeito majoritariamente positivo ou negativo, além do seu grau de importância para a previsão média. Portanto, a figura 9 surge para elucidar essas e outras questões.

Figura 9: Médias dos valores SHAP absolutos e impacto na variável explicada - modelo (X2).




Fonte: Elaborado pelo autor.

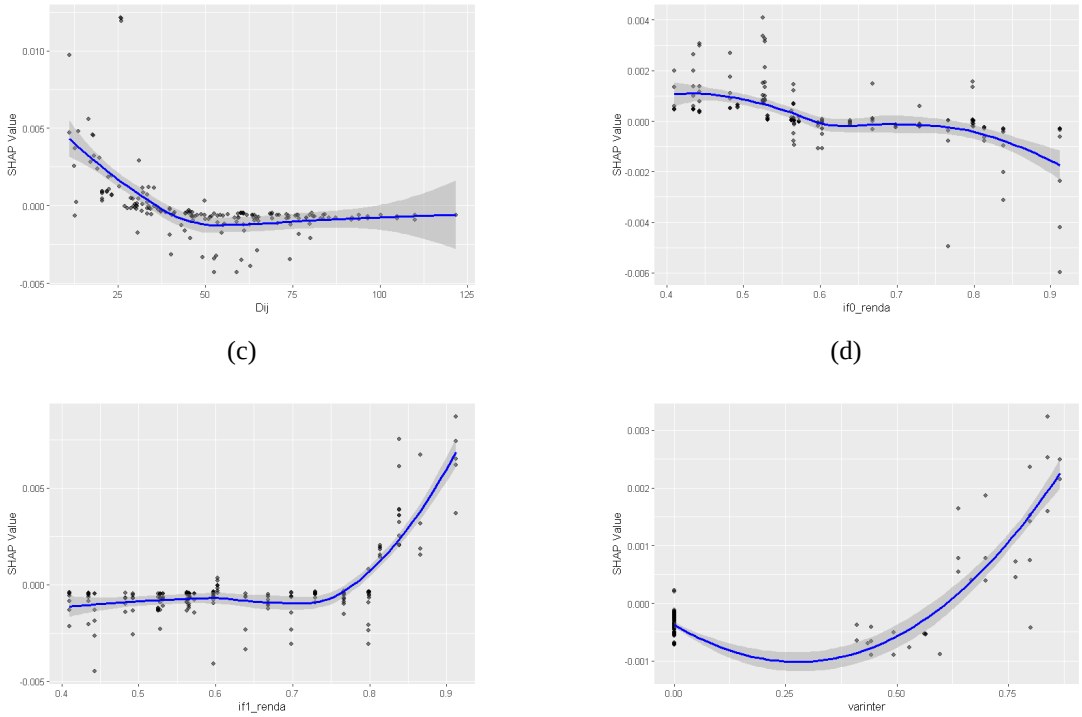
No modelo (X1) a variável de IFDM-ER do município de destino do deslocamento era a mais influente sobre a previsão; agora no modelo (X2), a distância entre os pares de cidades toma o seu lugar. Percebe-se, no que diz respeito às primeiras variáveis, efeitos idênticos aos das características do modelo (X1) sobre a previsão, ou seja: distâncias maiores reduzem a previsão média dos deslocamentos pendulares; IFDM-ER da cidade de destino do fluxo, quando alto, tende a elevar a previsão e; o efeito do IFDM-ER da cidade de origem é inverso.

Sobre a média do valor SHAP para a contiguidade, percebe-se que as observações se encontram muito próximas de zero bem como a sua média de 0.00006. Isto é uma evidência da heterogeneidade do efeito da vizinhança na previsão que, além de ser a variável menos influente na previsão, implica que ser um município vizinho não necessariamente é um fator de atração de pendulares dado que outros municípios mais distantes, ou pelo menos não fronteiriços, com maior dinâmica de mercado de trabalho, exerceriam uma força de atração maior.

A figura 10 apresenta os valores SHAP do modelo *XGBoost*, dessa vez com uso da variável de interação. Essa nova característica, criada como um produto da coluna de contiguidade com a coluna de IFDM-ER do município de destino dentro da base de dados, vai gerar conclusões a respeito do efeito espacial causado pelo fato de um município de destino ter alto desenvolvimento e ser vizinho do município de origem.

Figura 10: Valores SHAP - modelo (X3)

(a) (b)



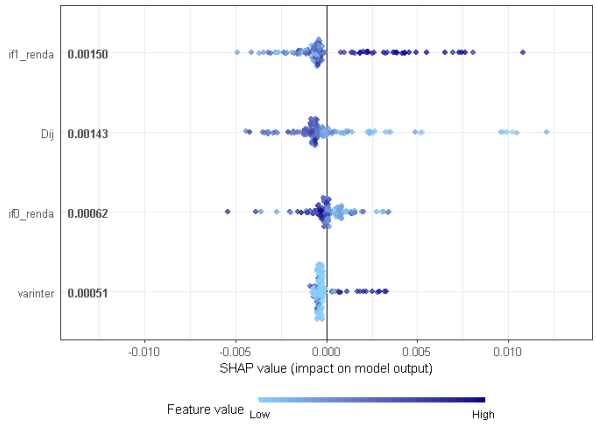

(c) (d)

Fonte: Elaborado pelo autor.

Mais uma vez se observa relações não lineares. 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, com destaque para o comportamento de if0_renda que se aproxima mais da sua variável equivalente no modelo (X2). É interessante analisar a variável de interação com mais detalhes. Observa-se 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.

Pode-se observar ainda durante a análise da curva da Figura 10(d) que existe somente uma observação da amostra que representa um vizinho com alto índice Firjan, mas está associada a uma redução na previsão média dos deslocamentos pendulares. Isso indica que o efeito é bem contundente. Para completar a análise dos modelos, a Figura 11 ilustra a média do valor SHAP para cada característica.

Figura 11: Médias dos valores SHAP absolutos e impacto na variável explicada - modelo (X3).




Fonte: Elaborado pelo autor.

Neste modelo (X3) a característica if1_renda volta a ser a variável mais relevante na previsão dos deslocamentos, assim como no modelo (X1). Ademais, dentre as três primeiras características, a interpretação é análoga aos outros modelos *XGBoost*, uma vez que os efeitos possuem sentidos idênticos.

Sobre a variável de interação, destaca-se que, mesmo quando não é nula, 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.

Os resultados encontrados nas três propostas de modelos gravitacionais estão de acordo com os resultados encontrados por outros autores que usaram métodos estatísticos através da equação gravitacional na análise de movimentos pendulares, como em Renkow e Hoover (2000), Brito e Ramalho (2019) e Zhao *et al.* (2023), além de também convergirem com as conclusões de outras pesquisas que utilizaram outros métodos, como em Polyzos, Tsiotas e Minetos (2013) e Brito e Silva (2020).

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 RM de Londrina. Essa comparação se encontra na tabela 5.

Tabela 6: 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 pelo autor.

Da tabela 5, 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. Se o espaço analisado compreendesse, por exemplo, todos os municípios paranaenses, o efeito espacial proposto nesta pesquisa provavelmente seria, além de significativo, determinante para um ganho de capacidade preditiva dos modelos, uma vez que o estado conta com várias áreas com dinâmicas próprias de desenvolvimento e, consequentemente, de pendularidade. Dessa maneira, um modelo que não incluísse a variável de interação proposta aqui, poderia estimar de maneira equivocada um deslocamento pendular entre duas cidades somente porque uma delas possui alto índice IFDM-ER e a outra possui baixo índice e ambas estão à uma distância razoável, porém, não se considerou o efeito da proximidade dentro de um dado cluster.

## CONSIDERAÇÕES FINAIS

Esta pesquisa analisou o efeito da distância geográfica e da característica de mercado de trabalho entre os pares de cidades 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.

Dessa forma, faz sentido realizar a comparação entre os dois modelos dado que alguns dos autores supracitados já declaravam a superioridade do modelo de ML e, após toda a análise, sugere-se que o modelo gravitacional estimado por PPML considerando efeitos espaciais é não somente adequado, mas, recomendável para analisar a pendularidade. Além disso, a inserção de variáveis que consideram efeitos espaciais traz uma nova abordagem dentro do modelo

gravitacional que, até então, sofria de críticas sobre a quantidade restringida de variáveis a serem analisadas (Simini *et al*, 2012) dada a estrutura limitante da equação.

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 entre os municípios vizinhos dos observados, mas que estão fora da base de dados foram ignorados. 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.

Delgado, Moura e Cintra (2013) ainda apontam que esse tipo de deslocamento, de dentro da RML para fora, como, por exemplo, em direção a Apucarana ou a Maringá, é até mais evidente proporcionalmente do que o observado na RM de Curitiba, indicando claramente que um efeito relevante foi excluído da pesquisa em nome da análise de um *cluster* mais específico.

Além disso, o efeito observado nesta pesquisa possivelmente não é representativo da realidade atual, uma vez que, pela escassez de dados, a análise leva a conclusões com defasagem de mais de uma década e, dessa forma, a divulgação dos microdados do Censo 2022 tornará possível uma análise adequada à conjuntura recente.

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.

Para isso, é necessário um investimento coordenado em várias frentes. Primeiramente, programas de desenvolvimento econômico local devem ser implementados para fortalecer as oportunidades de emprego nos municípios de origem, reduzindo assim a necessidade de deslocamento pendular. Isso pode incluir incentivos ao empreendedorismo, apoio à criação de novas empresas e qualificação da mão de obra local para atender às demandas do mercado de trabalho.

Além disso, políticas de infraestrutura de transporte devem ser priorizadas para melhorar a conectividade entre os municípios, tornando o deslocamento mais rápido, eficiente e acessível. Isso inclui investimentos em estradas, transporte público e outras formas de mobilidade. Ao mesmo tempo, políticas de desenvolvimento regional equilibrado são essenciais para distribuir de forma mais justa as oportunidades econômicas em toda a região, reduzindo as disparidades que levam ao deslocamento pendular excessivo.

Por fim, programas de capacitação e educação profissional devem ser expandidos para preparar os trabalhadores locais para as oportunidades de emprego disponíveis em suas próprias cidades, reduzindo assim a dependência de empregos em outras localidades. Em conjunto, essas políticas podem ajudar a promover um desenvolvimento econômico mais equitativo e sustentável na RML, reduzindo a necessidade e os impactos negativos do deslocamento pendular de trabalhadores.

# 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. Disponível em: https://doi.org/10.21608/jesaun.2021.57245.1026. Acesso em: 1 jul. 2023.

ALMEIDA, Eduardo. **Econometria espacial**. Campinas: Alínea, 2012.

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. Disponível em: https://doi.org/10.1080/13504851.2012.718052. Acesso em: 9 nov. 2023.

BARBOSA, Marina Rogério de Melo. **Condicionantes da mobilidade urbana: uma análise empírica para a Região Metropolitana do Recife**. Dissertação (Mestrado) – Universidade Federal de Pernambuco, Recife, 2 mar. 2015. Disponível em: https://repositorio.ufpe.br/handle/123456789/18662. Acesso em: 1 jul. 2023.

BHATT, V.; CHANDRASEKHAR, S.; SHARMA, A. Regional Patterns and Determinants of Commuting Between Rural and Urban India. **The Indian Journal of Labour Economics**, v. 63, n. 4, p. 1041-1063, dezembro 2020. Disponível em: https://doi.org/10.1007/s41027-020-00276-9. Acesso em: 1 jul. 2023.

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. Disponível em: https://www.bnb.gov.br/revista/index.php/ren/article/view/875. Acesso em: 1 jul. 2023.

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. Disponível em: https://doi.org/10.52292/j.estudecon.2021.1916. Acesso em: 1 de julho de 2023.

CAMERON, A. Colin; TRIVEDI, Pravin K. **Regression Analysis of Count Data**. Cambridge: Cambridge University Press, 1998. ISBN 978-0-521-63201-0 (capa dura); ISBN 978-0-521-63567-7 (brochura).

CAMERON, A. C.; TRIVEDI, P. K. **Regression Analysis of Count Data**. 2ª ed. Cambridge: Cambridge University Press, 2013. (Econometric Society Monographs, Book 53).

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. Disponível em: https://doi.org/10.1080/01944365608979229. Acesso em: 20 set. 2023.

CHEN, Hui et al. "Data-Driven Analysis on Inter-City Commuting Decisions in Germany". **Sustainability**, vol. 13, no 11, janeiro de 2021, p. 6320. Disponível em: https://doi.org/10.3390/su13116320. Acesso em: 1 de julho de 2023.

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. Disponível em: https://doi.org/10.1145/2939672.2939785. Acesso em: 24 set. 2023.

CHRISTALLER, Walter. Central Places in Southern Germany. Englewood Cliffs, N.J.: **Prentice-Hall**, 1966. Disponível em: https://cir.nii.ac.jp/crid/1130000798225693952. Acesso em: 5 out. 2023.

CINTRA, Anael; DELGADO, Paulo; MOURA, Rosa. "Deslocamentos intermunicipais para trabalho e estudo - Curitiba". Comunicado nº 21, junho 2012**. IPARDES - Instituto Paranaense de Desenvolvimento Econômico e Social**. Disponível em: https://www.ipardes.pr.gov.br/Pagina/Comunicados-2012. Acesso em: 1 de julho de 2023.

CIOMMI, Mariateresa, et al. "Integrating Parametric and Non-Parametric Multivariate Analysis of Urban Growth and Commuting Patterns in a European Metropolitan Area". **Quality & Quantity**, vol. 53, no 2, março de 2019, p. 957–79. Springer Link, https://doi.org/10.1007/s11135-018-0798-2. Acesso em: 1 de julho de 2023.

COLLA, Crislaine. **Migração e pendularidade na região metropolitana de Curitiba entre 2000 e 2010: complementaridade ou substituição?** 2018. Tese (Doutorado em Demografia) – Universidade Federal de Minas Gerais, Belo Horizonte, 2018. Disponível em: http://hdl.handle.net/1843/30112. Acesso em: 19 set. 2023.

DAUNI, P. et al. Implementation of Haversine formula for school location tracking. **Journal of Physics**: Conference Series, v. 1402, n. 7, 2019. Disponível em: https://doi.org/10.1088/1742-6596/1402/7/077028. Acesso em: 13 out. 2023.

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. Disponível em: https://ipardes.emnuvens.com.br/cadernoipardes/article/view/616. Acesso em: 1 de julho de 2023. Acesso em: 1 de julho de 2023.

ESCAMILLA, Juan Bacilio Guerrero et al. "Gravitational Model to Predict the Megalopolis Mobility of the Center of Mexico". **International Journal of Advanced Engineering, Management and Science**, vol. 4, n. 1, p. 239961, jan. 2018. Disponível em: https://www.neliti.com/publications/239961/. Acesso em: 1 jul. 2023.

FEDERAÇÃO DAS INDÚSTRIAS DO ESTADO DO RIO DE JANEIRO – FIRJAN. **Índice FIRJAN de Desenvolvimento Municipal (IFDM) 2018**: Ano-base 2016. Rio de Janeiro: FIRJAN, jun. 2018. Disponível em: https://www.firjan.org.br/ifdm. Acesso em: 27 set. 2023.

FRIEDMAN, Jerome H. Greedy function approximation: A gradient boosting machine. **The Annals of Statistics**, v. 29, n. 5, p. 1189-1232, out. 2001. Disponível em: https://doi.org/10.1214/aos/1013203451. Acesso em: 17 set. 2023.

HAAS, Anette; OSLAND, Liv. "Commuting, Migration, Housing and Labour Markets: Complex Interactions". **Urban Studies**, vol. 51, n. 3, p. 463-476, fev. 2014. DOI: 10.1177/0042098013498285. Acesso em: 1 de julho de 2023.

HAZANS, Mihails. **Commuting in the Baltic States**: Patterns, determinants and gains. ZEI working paper, No. B 02-2003, 2003. Disponível em: EconStor: Commuting in the Baltic States: Patterns, determinants and gains. Acesso em: 22 set. 2023.

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. Disponível em: https://doi.org/10.1016/B978-0-444-54314-1.00003-3. Acesso em: 16 out. 2023.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA – IBGE. **Questionários - Censo 2022**. Disponível em: https://censo2022.ibge.gov.br/sobre/questionarios.html. Acesso em: 26 fev. 2024.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA - IBGE. **Diálogos para o Futuro: Plano de Trabalho 2024**. Rio de Janeiro, 2024. Disponível em: https://www.ibge.gov.br/media/com_mediaibge/arquivos/f1f2bc78c86dadd599fcb48a83dae8e7.pdf. Acesso em: 29 jan. 2024.

KADEM, Mason et al. "**XGBoost for Interpretable Alzheimer's Decision Support**". In: Proceedings of the AAAI Symposium Series, vol. 1, nº 1, out. 2023, p. 135-141. Disponível em: https://doi.org/10.1609/aaaiss.v1i1.27491. Acesso em: 15 dez. 2023.

KAREEM, Fatima Olanike et al. **Fitting the Gravity Model When Zero Trade Flows Are Frequent**: A Comparison of Estimation Techniques Using Africa's Trade Data. 2016. Disponível em: https://ageconsearch.umn.edu/record/230588/. Acesso em: 23 nov. 2023.

KINGSFORD, Carl; SALZBERG, Steven L. What are Decision Trees? **Nature Biotechnology**, v. 26, n. 9, p. 1011-1013, setembro de 2008. Disponível em: https://www.nature.com. DOI: 10.1038/nbt0908-1011. Acesso em: 21 de dezembro de 2023.

LESAGE, James P.; FISCHER, Manfred M. Spatial Econometric Methods for Modeling Origin-Destination Flows. In: FISCHER, Manfred M.; GETIS, Arthur (Org.). **Handbook of Applied Spatial Analysis**: Software Tools, Methods and Applications. Springer, 2010, p. 409-433. Disponível em: https://doi.org/10.1007/978-3-642-03647-7_20. Acesso em: 14 out. 2023.

LESAGE, J. P.; PACE, R. K. Spatial econometric modeling of origin-destination flows**. Journal of Regional Science**, v. 48, n. 5, p. 941-967, dez. 2008. Disponível em: https://doi.org/10.1111/j.1467-9787.2008.00573.x. Acesso em: 10 out. 2023.

LI, Kaiming et al. Spatial Variations of Commuting Behavior and Their Impact Factors in Shanghai Metropolitan Area. **Frontiers in Built Environment**, v. 8, 2022. Disponível em: https://www.frontiersin.org/articles/10.3389/fbuil.2022.789024. Acesso em: 1 de julho de 2023.

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. Disponível em: https://www.mdpi.com/2220-9964/11/6/335. DOI: 10.3390/ijgi11060335. Acesso em: 1 de julho de 2023.

LI, Ziqi. Extracting spatial effects from machine learning model using local interpretation method: An example of SHAP and XGBoost. **Computers, Environment and Urban Systems**, v. 96, artigo 101845, set. 2022. Disponível em: https://doi.org/10.1016/j.compenvurbsys.2022.101845. Acesso em: 10 out. 2023.

LIANG, Quan et al. Travel Destination Prediction of Public Transport Commuters by Integrating XGBoost Algorithm and Graph Adjustment Method**.** 交通信息与安全, v. 39, n. 4, p. 68-76, 2021. Disponível em: https://doi.org/10.3963/j.jssn.1674-4861.2021.04.009. Acesso em: 22 out. 2023.

LUNDBERG, Scott; LEE, Su-In. **A Unified Approach to Interpreting Model Predictions**. arXiv:1705.07874v2 [cs.AI], 25 de novembro de 2017. Disponível em: https://doi.org/10.48550/arXiv.1705.07874. Acesso em: 05 de dezembro de 2023.

LUNDBERG, Scott M., et al. "From Local Explanations to Global Understanding with Explainable AI for Trees". **Nature Machine Intelligence**, vol. 2, no 1, janeiro de 2020, p. 56–67. Disponível em: https://doi.org/10.1038/s42256-019-0138-9. DOI: 10.1038/s42256-019-0138-9. Acesso em: 22 de dezembro de 2023.

MA, Meihong, et al. XGBoost-based method for flash flood risk assessment. **Journal of Hydrology**, vol. 598, julho de 2021, p. 126382. Disponível em: https://www.sciencedirect.com/science/article/pii/S002216942100379X. Acesso em: 15 jan. 2024. DOI: 10.1016/j.jhydrol.2021.126382.

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. Disponível em: https://doi.org/10.1016/j.engstruct.2020.110927. DOI: 10.1016/j.engstruct.2020.110927. Acesso em: 17 de dezembro de 2023.

MASUCCI, A. Paolo et al. Gravity versus radiation models: On the importance of scale and heterogeneity in commuting flows**. Physical Review E**, v. 88, n. 2, p. 022812, agosto de 2013. Disponível em: https://journals.aps.org/pre/abstract/10.1103/PhysRevE.88.022812. DOI: 10.1103/PhysRevE.88.022812. Acesso em: 1 de julho de 2023.

MCARTHUR, David Philip et al. The Spatial Transferability of Parameters in a Gravity Model of Commuting Flows. **Journal of Transport Geography**, v. 19, n. 4, p. 596–605, julho de 2011. Disponível em: https://www.sciencedirect.com/science/article/pii/S096669231000089X. DOI: 10.1016/j.jtrangeo.2010.06.014. Acesso em: 1 de julho de 2023.

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. Disponível em: https://www.osti.gov/biblio/1468184. Acesso em: 29 set. 2023.

MOURA, Rosa. Movimento pendular da população no Paraná: evidência da desconexão moradia/trabalho. **Cadernos Metrópole**, v. 12, n. 23, p. [páginas do artigo], jan. 2010. Disponível em: https://www.researchgate.net/publication/277092192. Acesso em: 18 set. 2023.

MULALIC, Ismir; ROUWENDAL, Jan. **Public Transport Investments, Commuting and Gentrification: Evidence from Copenhagen**. Preprint: 4404889, 30 mar. 2023. Social Science Research Network. Disponível em: https://doi.org/10.2139/ssrn.4404889. Acesso em: 18 jan. 2024.

NTI, Isaac Kofi; SOMANATHAN, Arjun Remadevi. "A Scalable RF-XGBoost Framework for Financial Fraud Mitigation". In: **IEEE Transactions on Computational Social Systems**, 2022, p. 1-8. IEEE Xplore. Disponível em: https://doi.org/10.1109/TCSS.2022.3209827. Acesso em: 28 dez. 2023.

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. ScienceDirect, https://doi.org/10.1016/j.asej.2020.11.011. Acesso em: 15 jan. 2024.

PARANACIDADE; SERVIÇO SOCIAL AUTÔNOMO PARANACIDADE vinculado à Secretaria de Desenvolvimento Urbano e de Obras Públicas (SEDU). **PDUÍ: Plano de Desenvolvimento Urbano Integrado da Região Metropolitana de Londrina** - Resumo Técnico. Londrina: [s.n.], novembro de 2022. Disponível em: https://www.amep.pr.gov.br/Pagina/PDUI-Londrina. Acesso em: 09 fev. 2024.

PARKER, H. R. Lösch, A.: "The Economics of Location" (Book Review). **The Town Planning Review**, v. 28, n. 2, p. 164, 1957. Disponível em: https://www.jstor.org/stable/40102105. Acesso em: 28 set. 2023.

PNUD/ATLAS DO DESENVOLVIMENTO HUMANO NO BRASIL. **Programa das Nações Unidas para o Desenvolvimento:** Atlas do Desenvolvimento Humano no Brasil - Índice de Desenvolvimento Humano Municipal (IDHM). Rio de Janeiro, 2023. Disponível em: http://www.atlasbrasil.org.br/acervo/atlas. Acesso em: 19 de junho de. 2023.

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. Acesso em: 1 de julho de 2023.

RAIHAN, Md Johir, et al. “Detection of the Chronic Kidney Disease Using XGBoost Classifier and Explaining the Influence of the Attributes on the Model Using SHAP”. **Scientific Reports**, vol. 13, no. 1, abril de 2023, p. 6263. Disponível em: https://www.nature.com/articles/s41598-023-33525-0. Acesso em: 15 jan. 2024. DOI: 10.1038/s41598-023-33525-0.

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. Disponível em: https://www.hindawi.com/journals/cmmm/2021/2577375/. Acesso em: 15 jan. 2024. DOI: 10.1155/2021/2577375.

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. DOI: 10.1111/0022-4146.00174. Acesso em: 1 de julho de 2023.

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. Disponível em: https://doi.org/10.1145/3209811.3209868. Acesso em: 12 out. 2023.

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. Disponível em: https://doi.org/10.1162/rest.88.4.641. Acesso em: 9 set. 2023.

SELLNER, Richard *et al*. A Spatial Autoregressive Poisson Gravity Model: A SAR Poisson Gravity Model. **Geographical Analysis**, v. 45, n. 2, p. 180-201, abril de 2013. DOI: 10.1111/gean.12007. Acesso em: 1 de julho de 2023.

SIMINI, Filippo *et al*. A Universal Model for Mobility and Migration Patterns. **Nature**, v. 484, n. 7392, p. 96-100, abr. 2012. Disponível em: https://doi.org/10.1038/nature10856. Acesso em: 15 out. 2023.

SOHN, Jungyul. Are commuting patterns a good indicator of urban spatial structure? **Journal of Transport Geography**, v. 13, n. 4, p. 306-317, dez. 2005. Disponível em: https://doi.org/10.1016/j.jtrangeo.2004.07.005. Acesso em: 26 out. 2023.

SPADON, Gabriel *et al*. Reconstructing commuters network using machine learning and urban indicators. **Scientific Reports**, v. 9, n. 1, p. 11801, agosto de 2019. Disponível em: https://www.nature.com/articles/s41598-019-48295-x. DOI: 10.1038/s41598-019-48295-x. Acesso em: 1 de julho de 2023.

STAMM, Cristiano. **Determinantes do movimento de trabalhadores pendulares na aglomeração urbana do nordeste do Rio Grande do Sul: uma análise a partir dos transportes coletivos**. 2013. Tese (Doutorado em Planejamento Urbano e Regional) - Universidade Federal do Rio Grande do Sul, Porto Alegre, 2013. Disponível em: https://lume.ufrgs.br/handle/10183/71909. Acesso em: 22 dez. 2023.

STAMM, Cristiano. **O fenômeno dos movimentos pendulares dos trabalhadores intermunicipais entre cidades de porte médio: o caso de Cascavel e Toledo (PR)**. 2005. Dissertação (Mestrado em Desenvolvimento Regional e Agronegócio) – Universidade Estadual do Oeste do Paraná, Cascavel, 2005. Disponível em: https://tede.unioeste.br/handle/tede/2234. Acesso em: 21 out. 2023.

STAMM, Cristiano; STADUTO, Jefferson Andronio Ramundo. Movimentos pendulares das cidades interioranas de porte médio de Cascavel e Toledo, no Paraná. **Revista Brasileira de Estudos de População**, v. 25, p. 131-149, junho de 2008. Disponível em: https://www.scielo.br/scielo.php?script=sci_arttext&pid=S0102-30982008000100008. DOI: 10.1590/S0102-30982008000100008. Acesso em: 1 de julho de 2023.

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. Disponível em: https://www.sciencedirect.com/science/article/pii/S2352146517300697. DOI: 10.1016/j.trpro.2017.05.069. Acesso em: 1 de julho de 2023.

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. Disponível em: https://periodicos.ufsc.br/index.php/geosul/article/view/1982-5153.2019v34n73p33. DOI: 10.5007/1982-5153.2019v34n73p33. Acesso em: 1 de julho de 2023.

TINBERGEN, Jan. Shaping the World Economy. **The International Executive**, v. 5, n. 1, p. 27-30, dez. 1963. DOI: 10.1002/tie.5060050113. Disponível em: https://onlinelibrary.wiley.com/doi/10.1002/tie.5060050113. Acesso em: 23 nov. 2023.

VARGA, Levente *et al*. **An Improved Radiation Model and Its Applicability for Understanding Commuting Patterns in Hungary**. fevereiro de 2017. Disponível em: https://mpra.ub.uni-muenchen.de/76806/#:~:text=Here%2C%20we%20introduce%20a%20new%20version%20of%20the,the%20results%20of%20the%20previous%20radiation%20type%20models. Acesso em: 1 de julho de 2023.

VON THÜNEN, Johann Heinrich. Isolated state: an English edition of Der isolierte Staat. Nova Iorque: **Pergamon Press**, 1966. Disponível em: https://archive.org/details/isolatedstateeng0000thun. Acesso em: 30 set. 2023.

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. Disponível em: https://doi.org/10.1016/j.envres.2021.111660. DOI: 10.1016/j.envres.2021.111660. Acesso em: 02 de dezembro de 2023.

WIESE, T. "Predicting Operating Train Delays into New York City using Random Forest Regression and XGBoost Regression Models." **International Journal of Engineering, Business and Management**, vol. 7, n. 1, p. 34–41, 2023. Disponível em: https://doi.org/10.22161/ijebm.7.1.5. Acesso em: 18 jan. 2024.

WÖLWER, Anna-Lena *et al*. Gravity Models in R. **Austrian Journal of Statistics**, v. 47, n. 4, p. 16-35, jun. 2018. Disponível em: https://doi.org/10.17713/ajs.v47i4.688. Acesso em: 30 set. 2023.

WOOLDRIDGE, Jeffrey M. ***Econometric Analysis of Cross Section and Panel Data***. 2ª ed. Cambridge, MA: MIT Press, 2002. P. 645-670

XGBOOST. **XGBoost Documentation**. Disponível em: https://xgboost.readthedocs.io/en/stable/index.html. Acesso em: 23 nov. 2023.

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. Springer Link, https://doi.org/10.1007/s11430-022-1022-9. Acesso em: 15 jan. 2024.

# APÊNDICE A – ESTIMAÇÕES ADICIONAIS DO MODELO GRAVITACIONAL

As tabelas 7 e 8 apresentam os coeficientes dos modelos gravitacionais estimados por MQO e por GPML. A estimação MQO utilizou a técnica de adição de uma constante na variável dependente para resolver o problema dos fluxos zero.

Tabela 7: Coeficientes do modelo gravitacional MQO (Y+1) – modelos (M1), (M2) e (M3)

| Variáveis | Modelo (M1) Coeficientes | Modelo (M2) Coeficientes | Modelo (M3) Coeficientes |
|---|---|---|---|
| Intercepto | -4,13* | -5,44* | -5,80* |
| Distância entre pares | -1,13* | -0,81* | -0,74* |
| IFDM-ER cidade de origem | -3,11* | -3,10* | -3,13* |
| IFDM-ER cidade de destino | 2,09* | 2,10* | 1,96* |
| Variável de interação | - | - | 1,17* |
| Dummy de contiguidade | - | 0,62* | - |

Fonte:Elaborado pelo autor.
(*) estatisticamente significativo a 1%

Tabela 8: Coeficientes do modelo gravitacional GPML – modelos (GP1), (GP2) e (GP3)

| Variáveis | Modelo (GP1) Coeficientes | Modelo (GP2) Coeficientes | Modelo (GP3) Coeficientes |
|---|---|---|---|
| Intercepto | 0,85 | 0,95 | 0,75 |
| Distância entre pares | -1,93* | -1,95* | -1,90* |
| IFDM-ER cidade de origem | -3,05* | -3,07* | -3,05* |
| IFDM-ER cidade de destino | 5,31* | 5,32* | 5,31* |
| Variável de interação | - | - | 0,05 |
| Dummy de contiguidade | - | 0,03 | - |

Fonte:Elaborado pelo autor.
(*) estatisticamente significativo a 1%