**UNIVERSIDADE DE SÃO PAULO
ESCOLA DE ENGENHARIA DE SÃO CARLOS**

Lucas Assirati

Análise da influência da vizinhança no comportamento individual relativo a viagens através de dados em painel

São Carlos

2018

Lucas Assirati

Análise da influência da vizinhança no comportamento individual relativo a viagens através de dados em painel

Tese apresentada à Escola de Engenharia de São Carlos da Universidade de São Paulo, para obtenção do título de Doutor em Ciências - Programa de Pós-Graduação em Engenharia de Transportes.

Área de concentração: Planejamento e Operação de Sistemas de Transporte

Orientadora: Profa. Dra. Cira Souza Pitombo

**São Carlos
2018**

AUTORIZO A REPRODUÇÃO TOTAL OU PARCIAL DESTE TRABALHO,
POR QUALQUER MEIO CONVENCIONAL OU ELETRÔNICO, PARA FINS
DE ESTUDO E PESQUISA, DESDE QUE CITADA A FONTE.

Ficha catalográfica elaborada pela Biblioteca Prof. Dr. Sérgio Rodrigues Fontes da
EESC/USP com os dados inseridos pelo(a) autor(a).

AA848a Assirati, Lucas
 Análise da influência da vizinhança no
comportamento individual relativo a viagens através de
dados em painel / Lucas Assirati; orientador Cira
Pitombo. São Carlos, 2018.

Tese (Doutorado) - Programa de Pós-Graduação em
Engenharia de Transportes e Área de Concentração em
Planejamento e Operação de Sistemas de Transporte --
Escola de Engenharia de São Carlos da Universidade de
São Paulo, 2018.

1. Comportamento individual relativo a viagens. 2.
Coleta de dados passiva. 3. Smartphones. 4. Análise de
Clusters. 5. K-médias. 6. Análise espacial. 7.
Indicadores globais. 8. Geoestatística. I. Título.

FOLHA DE JULGAMENTO

Candidato: Bacharel **LUCAS ASSIRATI**.

Título da tese: "Análise da influência da vizinhança no comportamento individual relativo a viagens através de dados em painel".

Data da defesa: 04/09/2018.

Comissão Julgadora:

Resultado:

Profa. Associada **Cira Souza Pitombo**
(Orientadora)
(Escola de Engenharia de São Carlos/EESC)

Aprovado

Prof. Titular **Jose Reynaldo Anselmo Setti**
(Escola de Engenharia de São Carlos/EESC)

Aprovado

Prof. Dr. **Cassiano Augusto Isler**
(Universidade Federal de Santa Catarina/UFSC)

APROVADO

Profa. Dra. **Suely da Penha Sanches**
(Universidade Federal de São Carlos/UFSCar)

APROVADO

Profa. Dra. **Ana Margarita Larrañaga Uriarte**
(Universidade Federal do Rio Grande do Sul/UFRGS)

Aprovado

Coordenadora do Programa de Pós-Graduação em Engenharia de Transportes:

Profa. Associada **Ana Paula Camargo Larocca**

Presidente da Comissão de Pós-Graduação:
Prof. Associado **Luís Fernando Costa Alberto**

*Dedico esse doutorado a Deus e aos meus
amados pais José Achilles e Eliana*

AGRADECIMENTOS

A Deus, pelas graças alcançadas e por estar sempre presente em minha vida, me iluminando.

Aos meus amados pais, pelo apoio incondicional.

À minha orientadora, pela sabedoria transmitida.

Aos meus amigos, por me darem forças para seguir em frente.

À Universidade de São Paulo (USP) e à Escola de Engenharia de São Carlos (EESC), pela estrutura disponibilizada.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), pelo apoio financeiro.

*“Sometimes life’s going to hit you in the head with a brick. Don’t lose faith.
I’m convinced that the only thing that kept me going was that I loved what I did.”*

Steve Jobs

RESUMO

ASSIRATI, L. **Análise da influência da vizinhança no comportamento individual relativo a viagens através de dados em painel.** 2018. 118p. Tese (Doutorado) - Escola de Engenharia de São Carlos, Universidade de São Paulo, São Carlos, 2018.

O comportamento individual relativo a viagens sofre a influência de fatores individuais e do meio urbano. Assim, a vizinhança seria uma das variáveis a serem consideradas na análise comportamental relacionada aos deslocamentos. O objetivo principal deste trabalho é analisar a influência da vizinhança no comportamento individual relativo a viagens, através de dados em painel. Dados em painel constituem importante ferramenta em análises comportamentais subjacentes a viagens urbanas, uma vez que propiciam maior quantidade de informações quando comparados aos dados seccionais. Padrões de viagens são mais bem evidenciados, através de dados em painel, caracterizando as habituais rotinas de atividades e viagens, além de melhor identificar comportamentos atípicos. Todavia, a obtenção desses dados comumente não é atividade trivial, demandando recursos monetários e de tempo. Um dos objetivos secundários deste trabalho é apresentar uma maneira prática e pouco onerosa de obtenção de dados em painel através de *Smartphones*. Tais dados, posteriormente, são aplicados à classificação de indivíduos segundo comportamento relacionado às viagens. A potencialidade da proposta sugerida é validada por meio de um estudo de caso relativo aos estudantes universitários do município de São Carlos – SP, Brasil. Através dos dados em painel, fornecidos pelos estudantes, utilizou-se o algoritmo k-médias considerando quatro variáveis relativas aos deslocamentos. As três categorias obtidas apresentam estrutura espacial e, portanto, possibilitam análises espaciais exploratórias e confirmatórias, almejando a compreensão de influências da vizinhança nas dinâmicas cotidianas. Este trabalho atesta a existência de autocorrelação espacial do conjunto de dados por meio de dois indicadores: Moran e SivarG (*Global Spatial Indicator Based on Variogram*). A corroboração da dependência espacial, apontada pelos indicadores globais, é confirmada por meio de dois modelos de escolha discreta. Um contendo apenas variáveis originais da base de dados. Outro, análogo ao primeiro, porém adicionado de covariáveis regionais, obtidas por preceitos da geoestatística. A incorporação das covariáveis regionais aumenta a precisão do modelo e promove um incremento das taxas de acertos em validações cruzadas.

Palavras-chave: Comportamento individual relativo a viagens. Coleta de dados passiva. *Smartphones*. Análise de *Clusters*. K-médias. Análise espacial. Indicadores globais. Geoestatística.

ABSTRACT

ASSIRATI, L. **Analysis of neighborhood influence on travel behaviour through panel data**. 2018. 118p. Thesis (PhD) – University of São Paulo, São Carlos School of Engineering, , São Carlos, 2018.

Individual travel behaviour is influenced by individual factors and the urban environment. Thus, the neighborhood influence would be one of the variables to be considered in travel behavior analysis related to urban displacements. The main objective of this work is to analyze the influence of neighborhood on travel behavior by panel data. Panel data is an important tool in urban travel behavioral analyzes, since they provide a greater amount of information when compared to sectional data. Travel patterns are more evident through panel data, characterizing the usual routines of activities, as well the atypical behaviors. However, obtaining these data is not a simple task, requiring monetary and time resources. Secondary goals of this work aim to present a practical and inexpensive way to obtain panel data through Smartphones. These data are then applied to the classification of individuals according to travel behavior. The potential of the proposal is validated by a case of study concerning undergraduate and PhD students from São Carlos - SP, Brazil. Using the data provided by the students, a k-means algorithm was used considering four variables regarding displacements. These three categories have spatial structure and allow exploratory and confirmatory spatial data analyzes aiming the comprehension of the nearby influence of data at daily dynamics. This work attests to the existence of spatial autocorrelation of the data set by two indicators: Moran and SivarG (*Global Spatial Indicator Based on Variogram*). Corroboration of spatial dependence, pointed by the global indicators, is confirmed by two discrete choice models. The first one includes just the original database variables. The second one, analogous to the first, but added of regional covariates obtained by geostatistical concepts. The addition of regional variables leads to a more accurate model, increasing cross-validations hit rates.

Keywords: Travel Behaviour. Passive Data Collection. Smartphones. Cluster Analysis. K-means. Spatial analysis. Global indicators. Geostatistics.

LISTA DE FIGURAS

Figura 1 – Esquema semivariograma teórico/experimental. Adaptado de Burrough (1986).	42
Figura 2 – Localização do município de São Carlos, SP – Brasil.	44
Figura 3 – Mapa do município de São Carlos, SP – Brasil, apresentando a localização dos campi da Universidade de São Paulo e o campus da Universidade Federal de São Carlos.	45
Figura 4 – Representação espacial do binário de escolha modal na cidade sintética.	46
Figura 5 – Fluxograma das etapas metodológicas.	47
Figura 6 – Exemplo de clusterização de tempos e distâncias para identificação dos locais de atividades dos indivíduos: (A) dados brutos; e (B) resultado do pré-processamento, com o local de atividade identificado.	49
Figura 7 – Exemplo de dados originais submetidos pelos participantes.	50
Figura 8 – Exemplo de obtenção de tempo e distância de deslocamento.	50
Figura 9 – Exemplo de estimação da velocidade média do deslocamento.	51
Figura 10 – Exemplo de obtenção de recorrência das localidades.	51
Figura 11 – Exemplo de obtenção da variável " <i>Caráter de motorização individual</i> ".	52
Figura 12 – Exemplo de formação da cadeia de viagens.	54
Figura 13 – Acréscimo de informações quantitativas à cadeia de viagens.	56
Figura 14 – Etapas do algoritmo k -médias.	59
Figura 15 – Variograma omnidirecional para a base de dados principal.	61
Figura 16 – Exemplo de cálculo de probabilidades, por categorias, para o ponto X, segundo quatro vizinhanças de interesse.	62
Figura 17 – Aplicativo <i>Google Maps</i> com as configurações necessárias para coleta de dados.	69
Figura 18 – Interface <i>web</i> para submissão dos dados coletados.	70
Figura 19 – Percentual da diferença absoluta entre valores declarados e computados frente ao universo para cada variável.	75
Figura 20 – Particionamento estabelecido pelo algoritmo k -médias.	77
Figura 21 – Mapa construído a partir das coordenadas da residência dos indivíduos estudados.	85
Figura 22 – Valores das métricas comparativas para os modelos não-espacial e espacial, usando-se a base de dados auxiliar.	92
Figura 23 – Valores das métricas comparativas para os modelos não-espacial e espacial - Base de dados principal.	98

LISTA DE TABELAS

Tabela 1 – Medidas descritivas da base de dados auxiliar – Cidade Fictícia	46
Tabela 2 – Sumarização das características dos dados coletados por <i>Smartphones</i> .	58
Tabela 3 – Sumarização da utilização de ferramentas durante as etapas do projeto	71
Tabela 4 – Comparativo entre dados declarados e computados - Amostra piloto (parte1).	73
Tabela 5 – Comparativo entre dados declarados e computados - Amostra piloto (parte2).	74
Tabela 6 – Resultados dos testes não paramétricos para amostras independentes. .	75
Tabela 7 – Sumarização das características do Grupo 1	79
Tabela 8 – Sumarização das características do Grupo 2	80
Tabela 9 – Sumarização das características do Grupo 3	81
Tabela 10 – Resultado dos indicadores Moran e SivarG aplicados ao banco de dados sintético.	89
Tabela 11 – Calibração pelo método da máxima verossimilhança considerando as variáveis originais significativas da base de dados sintética	90
Tabela 12 – Calibração pelo método da máxima verossimilhança considerando as covariáveis espaciais acrescidas significativas, para base de dados sintética.	91
Tabela 13 – Resultado dos indicadores Moran e SivarG aplicados ao banco de dados principal.	93
Tabela 14 – Calibração pelo método da máxima verossimilhança considerando as variáveis originais significativas da base de dados principal	95
Tabela 15 – Calibração pelo método da máxima verossimilhança considerando as covariáveis espaciais acrescidas significativas, para base de dados principal.	97

LISTA DE ABREVIATURAS E SIGLAS

CSS	<i>Cascading Style Sheets</i>
ESDA	<i>Exploratory Spatial Data Analysis</i>
GPS	<i>Global Positioning System</i>
HTML	<i>Hypertext Markup Language</i>
IBGE	Instituto Brasileiro de Geografia e Estatística
Km	Quilômetro
Km/h	Quilômetro por hora
logit	<i>Logistic Model</i>
m	Metros
min	Minutos
mySQL	<i>My Structured Query Language</i>
PHP	<i>Hypertext Preprocessor</i>
PIB	Produto Interno Bruto
SivarG	<i>Global Spatial Indicator Based on Variogram</i>
SP	Estado de São Paulo
SSL	<i>Secure Sockets Layer</i>
UFSCar	Universidade Federal de São Carlos
USP	Universidade de São Paulo
WebDev	<i>Web Development</i>

SUMÁRIO

1	INTRODUÇÃO	23
1.1	Objetivos	25
1.1.1	Objetivo Principal	25
1.1.2	Objetivos específicos	25
1.2	Estrutura do documento	25
2	OBTENÇÃO DE DADOS E CARACTERIZAÇÃO COMPORTAMEN- TAL	27
2.1	Dados em Painel <i>versus</i> dados seccionais	27
2.2	Caracterização e análise do comportamento individual relativo a viagens	31
2.2.1	Estudos utilizando dados de um único dia	31
2.2.2	Estudos utilizando dados de múltiplos dias	32
2.2.3	Estudos utilizando encadeamento de viagens	33
3	ANÁLISE ESPACIAL DE DADOS	35
3.1	Definição e o estado da arte	35
3.2	Análise espacial exploratória	36
3.2.1	Análise espacial exploratória de eventos ou padrões pontuais	36
3.2.2	Análise espacial exploratória de superfícies contínuas	37
3.2.3	Análise espacial exploratória de áreas com taxas agregadas	37
3.3	Análise espacial confirmatória	38
3.3.1	Análise espacial confirmatória de eventos ou padrões pontuais	38
3.3.2	Análise espacial confirmatória de superfícies contínuas	39
3.3.3	Análise espacial confirmatória de áreas com taxas agregadas	39
3.3.4	Análise espacial confirmatória através de modelos de escolha discreta	39
3.4	Análise espacial exploratória e confirmatória aplicadas ao doutorado	41
4	MATERIAIS E MÉTODO	43
4.1	Materiais	44
4.1.1	Base de dados principal: Dados em painel provenientes de <i>Smartphones</i>	44
4.1.2	Base de dados auxiliar: A cidade fictícia	46
4.2	Método	47
4.2.1	Coleta de dados via <i>smartphones</i>	47
4.2.1.1	Pré-processamento dos dados brutos	48
4.2.1.2	Informação original	50

4.2.1.3	Tempo e distância de deslocamento	50
4.2.1.4	Estimação de velocidade média de deslocamento	51
4.2.1.5	Análise de recorrência das localidades	51
4.2.1.6	Caráter de motorização individual	52
4.2.1.7	Validação da coleta de dados através de diário de atividades tradicional . . .	53
4.2.2	Construção das cadeias de viagens	54
4.2.3	Algoritmo de agrupamento k-médias	59
4.2.4	Análise espacial exploratória	60
4.2.5	Proposta de extração de covariáveis espaciais	61
4.2.6	Calibração de funções utilidade apenas com variáveis originais	64
4.2.7	Calibração de funções utilidade com variáveis originais e covariáveis espaciais propostas	65
4.2.8	Análise comparativa entre modelos de escolha discreta não espacial e espacial	67
4.3	Aplicativos e ferramentas	69
4.3.1	Aplicativo para telefones móveis <i>Google Maps</i>	69
4.3.2	Ferramentas de desenvolvimento <i>web</i> (<i>WebDev</i>)	69
4.3.3	Linguagem de programação <i>Python</i>	70
4.3.4	<i>Biogeme</i>	71
4.3.5	Alocação de ferramentas por etapas	71
5	RESULTADOS DO PROCESSO DE CLASSIFICAÇÃO E AGRUPAMENTO	73
5.1	Validação do método de coleta de dados	73
5.2	Resultados do processo de agrupamento	77
6	RESULTADOS DA MODELAGEM PARAMÉTRICA	87
6.1	Pré-requisitos da modelagem na base de dados auxiliar	87
6.2	Resultados da modelagem paramétrica para a cidade fictícia	89
6.3	Resultados da modelagem paramétrica para a base de dados principal	93
7	CONSIDERAÇÕES FINAIS	101
7.1	Síntese do trabalho	101
7.2	Conclusões e contribuições	103
7.3	Recomendações	104
	REFERÊNCIAS	107

1 INTRODUÇÃO

Diariamente, pessoas realizam múltiplos padrões de deslocamentos visando cumprir suas atividades dentro de um ambiente urbano. Bowman e Ben-Akiva (2001) afirmam que a demanda por viagens, e sua consequente relação com transportes, é derivada da demanda por essas atividades. O entendimento da demanda por transportes está, portanto, condicionado a um relevante aspecto denominado análise comportamental.

Os fatores que determinam o comportamento individual relativo a viagens ainda não são completamente compreendidos, havendo diversas circunstâncias que interferem nesse contexto tais como os atributos pessoais e condições socioeconômicas do indivíduo (CARVALHO; PEREIRA, 2011; CARVALHO et al., 2013; SHARABY; SHIFTAN, 2012), motivações para a realização de determinados deslocamentos (HOWARD; BURNS, 2001; TAKANO, 2011; PITOMBO; GOMES, 2014), facilidade de acesso a determinados modos de transporte (BAMBERG; RÖLLE; WEBER, 2003; GIULIANO; DARGAY, 2006), tempo de deslocamento (LISCO, 1967; FRANK et al., 2008), influência exercida pela sua vizinhança (CAO; MOKHTARIAN; HANDY, 2007; KIM; ULFARSSON; HENNESSY, 2007; PITOMBO; KAWAMOTO; SOUSA, 2011), dentre outros.

A influência da vizinhança, já foi tema de estudo em diversos trabalhos, considerando classificações de uso do solo (PAN; SHEN; ZHANG, 2009; ENGBRETSSEN; CHRISTIANSEN, 2011; SILVA, 2014) e/ou dependência espacial das variáveis relativas a transportes, através de uso de técnicas de estatística espacial (HAMMADOU et al., 2008; PRITCHARD et al., 2014). No presente trabalho, o termo *vizinhança* é caracterizado como as escolhas comportamentais dos vizinhos, situados a determinadas faixas de distâncias.

Outras lacunas importantes que este trabalho pretende preencher seriam a obtenção de dados em painel e a classificação de indivíduos, segundo escolhas de deslocamentos de múltiplos dias. Para o estabelecimento de modelos que promovam descrições detalhadas, acerca de uma determinada população, da qual se almeja compreender as atividades cotidianas, faz-se necessário um conjunto de informações pertinente a esse grupo de pessoas e à região onde estão inseridas. Em posse dessas informações, deve-se trabalhar esses dados de forma a agrupar os indivíduos estudados em categorias relativas aos seus padrões comportamentais, pois uma correta categorização embasa as primeiras análises acerca daquele conjunto populacional e serve de subsídio para modelos paramétricos mais complexos.

Assim, este trabalho, inicialmente, destina-se a propor uma maneira prática de obter dados em painel, que possibilitem posterior classificação dos indivíduos em grupos, baseados na análise comportamental de seus deslocamentos, considerando três dias úteis.

A opção por dados em painel ocorreu considerando que esses apresentam maiores benefícios frente à clássica utilização de dados seccionais que, por terem um único recorte temporal, não privilegiam a visualização de heterogeneidades de ações, característica extremamente almejada em análises comportamentais.

Sua obtenção foi viabilizada por meio de *smartphones*, um meio viável e pouco oneroso de se obter um conjunto de dados em painel para a pesquisa. Isso é possível dado que os aparelhos atuais são dotados de sensores GPS e conectividade com a *internet* móvel, além da grande penetração desse tipo de aparelho junto à população. Pesquisas dão conta que no Brasil, 94% das pessoas declaram ter um aparelho móvel (PIZZOL, 2018), sendo que 62% é proprietário de, ao menos, um aparelho do tipo *smartphone* (NOGUEIRA, 2018).

Assim, por meio do uso de aplicativos adequados (aplicações de mapas, aplicações voltadas para atividades físicas, registros de deslocamentos urbanos, etc.), é possível que o telefone colete informações de deslocamentos e as armazene, tornando-se assim, uma ferramenta em potencial para a descrição e estudos de viagens diárias. Seu uso justifica-se por duas razões: a primeira é a de que há considerável presença desse tipo de dispositivo na população; a segunda é a de que as pessoas realizam seus deslocamentos diários em posse de seus *smartphones*, fazendo com que a coleta de dados seja passiva, constante e precisa.

Esses dados coletados apresentam estrutura espacial e, por conseguinte, variáveis derivadas dessas informações, como por exemplo, escolha modal, número total de viagens realizadas por indivíduo, média de distância percorrida, também possuem dependência espacial (PÁEZ, 2007; PÁEZ et al., 2013; SILVA, 2008).

Logo, compreender como é a distribuição geográfica do local de moradia dos indivíduos bem como dos locais de realização de atividades (trabalho, lazer, estudos, etc.), num determinado espaço, é relevante ao estudo de deslocamentos urbanos (ORTÚZAR; WILLUMSEN, 2011), propiciando, posteriormente, a extração de variáveis regionalizadas ou locais.

Entretanto, o uso de variáveis regionalizadas, para incremento de modelos para previsão de demanda por transportes, requer uma investigação inicial acerca da dependência espacial desses dados. Uma forma de avaliação da associação espacial se dá por meio de indicadores que se destinam a averiguar a existência de dependência espacial das informações estudadas, por meio de análises exploratórias. Confirmando-se eventuais associações, têm-se a garantia que tais dados podem compor modelos espaciais confirmatórios para previsão de viagens (ANSELIN, 1995; GETIS; ORD, 1992; NAIZER, 2018).

Valendo-se dos indicadores espaciais Moran (MORAN, 1950; ANSELIN, 1995) e SivarG (NAIZER, 2018) este trabalho busca atestar, de maneira sistemática, as regiões

geográficas do conjunto de dados cujas informações apresentem associação espacial – Análise espacial exploratória. A seguir, promove-se a comparação de dois modelos de escolha discreta: o primeiro apenas com covariáveis tradicionais e o segundo com a incorporação de covariáveis espaciais, propostas a partir de pressupostos da geoestatística – Análise espacial confirmatória. As ações de análise espacial exploratória e confirmatória são aplicadas, primeiramente, em um conjunto de dados sintéticos, para efeitos de testes. Uma vez obtido sucesso neste protótipo, o método é empregado aos dados em painel, alvo do estudo desse doutorado.

1.1 Objetivos

1.1.1 Objetivo Principal

O objetivo principal deste trabalho é investigar a influência da vizinhança (caracterizada a partir de pressupostos da geoestatística) no comportamento individual de viagens, a partir de dados em painel. Assim, a hipótese que pretende-se corroborar é a de que a vizinhança exerce influência no comportamento individual relativo a viagens.

1.1.2 Objetivos específicos

Para alcançar a meta proposta, foram estabelecidos os seguintes objetivos específicos:

- Compor um painel de dados de múltiplos dias através de coleta de dados via *smartphones*;
- Validar a coleta de dados, que compõem o painel, através de uma pesquisa por diário de atividades tradicional;
- Classificar os indivíduos baseados em seus padrões comportamentais (encadeamentos) de múltiplos dias;
- Comparar resultados entre modelos espaciais e não espaciais e verificar implementos a partir da incorporação da vizinhança.

1.2 Estrutura do documento

Esta tese é formada por seis capítulos, além desta introdução. O Capítulo 2 descreve o referencial teórico sobre a obtenção de dados e caracterização comportamental. O Capítulo 3 apresenta referencial teórico sobre análise espacial. O Capítulo 4 é destinado à descrição dos materiais (bancos de dados, ferramentas e aplicativos utilizados) e método. O Capítulo 5 detalha os resultados relativos ao processo de classificação e agrupamento comportamental, enquanto que o Capítulo 6 apresenta os resultados referentes às modelagens paramétricas. Finalmente, o Capítulo 7 discorre sobre as conclusões deste trabalho de doutorado.

2 OBTENÇÃO DE DADOS E CARACTERIZAÇÃO COMPORTAMENTAL

Esse capítulo destina-se ao referencial teórico acerca da obtenção de dados para pesquisas em transportes e da caracterização e análise comportamental dos indivíduos, relativa aos seus deslocamentos cotidianos.

No âmbito da obtenção de dados, são apresentadas as vantagens de se trabalhar com dados em painel frente à abordagem via dados seccionais, no que se refere à identificação de heterogeneidades e diferenças individuais. Além disso, são descritos os cuidados adicionais, tais como a escolha do período de coleta, frisando-se a importância de se estabelecer métodos robustos de obtenção de dados.

Quanto à caracterização e análise comportamental relativa a viagens, discorre-se a respeito de algumas alternativas de caracterização: estudos utilizando dados de um único dia, estudos utilizando dados de múltiplos dias e estudos utilizando encadeamento de viagens. Para todas as alternativas são descritos os meios utilizados para as análises das viagens e categorização dos indivíduos.

2.1 Dados em Painel *versus* dados seccionais

Dados em painel (longitudinais) consistem em uma série cronológica de informações observadas num determinado intervalo de tempo. Diferem-se dos dados seccionais uma vez que esses são coletados em um particular instante de tempo e, portanto, podem transmitir uma equivocada ideia de estabilidade, não considerando uma eventual dinâmica das informações em questão (MARQUES et al., 2000).

Como os dados em painel possuem uma intrínseca cronologia, associada à sua aquisição, trazem benefícios para pesquisas, pois são capazes de revelar heterogeneidades das variáveis analisadas, evidenciadas apenas quando a análise abrange um intervalo de tempo contínuo. Além disso, conta-se com uma maior quantidade de informações contemplando assim uma maior variabilidade dos dados (MARQUES et al., 2000), havendo inclusive esforços para agrupar dados seccionais consecutivos de forma a torná-los o equivalente a uma coleta de dados em painel, quando esses não estão disponíveis (DEATON, 1985).

Dados de corte seccional são uma coleção de observáveis, provenientes de várias entidades, em um mesmo corte temporal. Por exemplo, múltiplos indivíduos e seus deslocamentos em um único dia.

Se para um indivíduo desses fossem coletados dados cobrindo múltiplos dias de deslocamentos, teria-se uma série temporal para este indivíduo. Isto é, uma coleção de observáveis, provenientes de uma única entidade, para um período contínuo de tempo.

Finalmente, ao se combinar ambas as dimensões (dados de corte seccional e séries temporais) obtém-se os dados em painel, que consistem de uma coleção de observáveis, provenientes de várias entidades, em um período contínuo de tempo. No caso deste trabalho, vários indivíduos provendo múltiplos dias de deslocamentos.

Hsiao (2014) lista uma série de vantagens, relativas ao uso de dados em painel, frente os dados de corte seccional. Menciona-se o fato de que dados em painel normalmente provém um número maior de registros e aumentam os graus de liberdade. Propiciam maior precisão nas inferências de modelos paramétricos; possibilitam maior capacidade de construção de hipóteses comportamentais mais realistas; e evidenciam ações contempladas apenas ao se considerar dinâmicas temporais.

É necessário, no entanto, estar atento ao tamanho da amostra pertinente aos objetivos de um determinado estudo. Há diversas maneiras de se realizar processos de amostragem (de maneira aleatória, de maneira sistemática, através de voluntários, por oportunidade, dentre outras).

Para o caso de pesquisas domiciliares que levam em conta diários de atividades de um único dia, é predominante o uso de amostragem estratificada (divisão da população em um certo número de subpopulações que não se sobrepõem para extração de um conjunto de amostras de cada estrato). Bruton e Lee (1978) sugerem que se tenha um tamanho mínimo da amostra de 1 indivíduo para cada 35 habitantes, mas que idealmente almeja-se uma proporção de 1 para 10, sendo que valores intermediários também são utilizados (BIELENKI; RAIA; SILVA, 2008), dependendo da precisão que deseja-se para o estudo.

É sabido que o aumento do número de dias na coleta contempla maior variabilidade (PAS; SUNDAR, 1995; PENDYALA, 1999), o que é desejável. Entretanto, há implicações desse fato no tamanho da amostra uma vez que dados em painel têm alto grau de dependência. Assim, Stopher et al. (2008) propõem um equacionamento que, para uma variável de interesse, determina a variância dos dados em painel considerando o número de indivíduos, o número de dias de coleta de dados, a variabilidade entre os diversos indivíduos e a variabilidade de um mesmo indivíduo.

A formulação proposta, ao considerar o grau de dependência dos dados em painel (pois inclui no cálculo variabilidades individuais e globais), consegue explicitar a relação entre o tamanho da amostra e o número de dias considerados com relação a variância dos dados. Dessa forma, pode-se ter uma boa referência do número de amostras necessárias para o número de dias empregados na composição do painel e a precisão (valor de variância) que deseja-se para o estudo.

Valendo-se de dados de distância de viagem, baseados em GPS para a cidade de Adelaide - Austrália, Stopher et al. (2008) atestam que, usando o equacionamento proposto, dados de uma pesquisa para 7 dias podem diminuir os tamanhos de amostra necessários

em 65% com relação a uma pesquisa convencional de um único dia. Caso a duração seja de 15 dias, o tamanho da amostra seria reduzido em mais de 70 % em relação a uma pesquisa convencional de um dia.

Os mesmos ainda promovem estimativas monetárias e afirmam que, ao se aplicar uma pesquisa de 15 dias com o uso de tecnologia de GPS e amostragem proporcional, é possível haver uma economia de aproximadamente metade dos recursos monetários que seriam empregados em uma pesquisa realizada por meio de entrevistas domiciliares considerando apenas um único dia.

Todas essas virtudes fazem com que os dados em painel sejam utilizados nos mais variados campos da ciência como economia (CORNWELL; TRUMBULL, 1994; LAWRENCE, 1991; LEITÃO, 2010; ROCKOFF, 2004; SANTOLIN; JR; REIS, 2009), saúde e medicina (MENDONÇA; MOTTA, 2005; PARENTE; MCCULLOUGH, 2009; POWELL, 2009; SOUSA; LEITE, 2008), agricultura (ANTONINI; ARGILÉS-BOSCH, 2017; BLANC; SCHLENKER, 2017; KLOMP; HOOGEZAND, 2018; KOUSER; QAIM, 2011) e engenharia (BERALDI, 2017; REITER, 2015; SILVA; MENEZES; NEDER, 2016; YANG et al., 2017).

Na área de demanda por transportes, são utilizados, mais comumente, dados de Preferência Declarada e Revelada, dados domiciliares de pesquisas origem-destino ou diários de atividades. Geralmente, nessas citadas pesquisas, o comportamento individual relacionado aos deslocamentos urbanos (quantidade de viagens diárias; modos de transporte utilizados; rotas escolhidas; horários do dia; localização de destinos, etc.) geralmente é caracterizado pelo dia anterior à entrevista (STOPHER et al., 1996). Desta forma, a análise comportamental não se torna completa, não sendo possível caracterizar, adequadamente, comportamentos atípicos e/ou habituais. Além disso, tais dados estão suscetíveis a erros de percepções, já que muitas informações são coletadas baseadas no juízo dos indivíduos.

Curtis e Perkins (2006) desenvolveram um trabalho de revisão da literatura onde mais de oitenta trabalhos sobre o comportamento do indivíduo subjacente a viagens foram analisados. Em sua vasta maioria, a fonte de dados de todos esses trabalhos é seccional. Os autores ainda mencionam a carência de critérios acerca da obtenção desses dados.

Alguns trabalhos (BOARNET; SARMIENTO, 1998; GIULIANO; DARGAY, 2006; GOUDIE, 2002; VASCONCELLOS, 2005) levam em conta um diário de atividades composto por apenas um único dia de viagens, isso por si só já implicaria pouca variabilidade dos dados, típica de coletas seccionais.

Entretanto, Curtis e Perkins (2006) ainda alertam que mais grave do que o uso de apenas um único dia de registros é a ausência de um critério para seleção deste dia, sendo que há pesquisas (NEWBOLD et al., 2005) em que os participantes podem descrever suas atividades em qualquer dia do ano (dias úteis, feriados, finais de semana). Em muitos

desses casos, os trabalhos destinam-se à análise comportamental relativa a viagens ao trabalho, e, portanto, registros de dias como sábado, domingo e feriados não são exatamente pertinentes ao estudo.

Salientam, ainda, que ater-se à escolha do dia de registro pode não ser suficiente, sendo que, muitas vezes, se faz necessário ser criterioso quanto à escolha do mês da aquisição de dados. Por exemplo, caso o trabalho se destine a compreensão dos deslocamentos destinados a estudo, não é apropriado obter dados em meses de férias escolares. Dado que diferentes projetos consideram diferentes tipos de comportamento subjacente à viagem, fica claro que a adequada seleção do período de tempo dos registros de dados é fundamental para o sucesso das análises.

Cita-se também trabalhos que utilizam diários de atividades que consideram mais de um dia de viagens (BHAT, 1999; FRUSTI; BHAT; AXHAUSEN, 2002; TRAPP et al., 2011; TRAPP et al., 2012). Apesar de serem mais completos do que a abordagem que leva em conta um único dia, ainda são imprecisos quanto à qualidade das informações coletadas.

Na tentativa de sanar a problemática da não variabilidade de dados e eventual homogeneidade das ações de indivíduos, alguns trabalhos passaram a utilizar dados em painel em suas análises, contemplando a almejada diversidade de informações e o aumento do número de dias de registros. Segundo Baltagi (2014), os dados em painel fornecem maior capacidade de capturar a complexidade das influências de determinadas variáveis relativas a transportes. Uma dessas variáveis é a posse de veículos (especialmente automóveis), que sabidamente exerce grande influencia para modelos de análise de deslocamentos (ACKER; WITLOX, 2010; CERVERO; GOLUB; NEE, 2007; DISSANAYAKE; MORIKAWA, 2002; GIULIANO; DARGAY, 2006). Tal influência já era bem identificada por meio de dados seccionais, mas o uso de dados em múltiplos dias traz à tona novas perspectivas, especialmente comportamentais, como por exemplo quais motivações levariam um indivíduo dotado de veículo motorizado a desempenhar deslocamentos não-motorizados (THØGERSEN, 2006; SIMMA; AXHAUSEN, 2003).

Um entrave, no entanto, é o fato de que são poucos os trabalhos que se propõe a obter seus próprios dados. Sendo assim, a posse de dados em painel torna-se tarefa não trivial, sendo necessário muitas vezes obtê-los de terceiros, esbarrando em empecilhos monetários (CAO; MOKHTARIAN; HANDY, 2007) e burocráticos como, por exemplo, a frequência de estabelecimento desses dados por órgãos estatais (CLARK; CHATTERJEE; MELIA, 2016; YANG et al., 2017).

É relevante frisar que as pesquisas em painel não são úteis somente para revelar diferenças entre os diversos indivíduos, mas também para revelar diferenças comportamentais de um mesmo indivíduo.

2.2 Caracterização e análise do comportamento individual relativo a viagens

Na área de modelagem de demanda por transportes, muitas pesquisas já se dedicaram ao estudo de padrões comportamentais de viagens de indivíduos, sendo caracterizados de diversas maneiras. Destacam-se os estudos utilizando um único dia de coletas de dados, múltiplos dias de coleta de dados e composição de sequência de deslocamentos.

2.2.1 Estudos utilizando dados de um único dia

Kockelman (1997), considerando entrevistas relativas a um único dia útil, caracterizou o comportamento relativo a viagens pela distância percorrida, propósito da viagem (trabalho ou outros) e modo utilizado (veículo motorizado ou não motorizado). Posteriormente, tal comportamento foi definido através de modelos lineares e *logit* e variáveis explicativas, tais como densidade populacional, acessibilidade e índices entrópicos para uso do solo. Utilizando dados da pesquisa domiciliar nacional, realizada por entrevistas telefônicas, Santos (2011) estabelecem, por meio de análises estatísticas, padrões comportamentais relativos aos deslocamentos, considerando um único dia de viagens, distintas faixas etárias, modos de transporte motorizado (veículos próprios, transporte coletivo, caronas) e o trabalho como motivação para o deslocamento.

Estudando viagens na cidade de São Carlos, SP – Brasil, Amancio (2005), aplicando modelos de escolha discreta em dados censitários do município, afirmam que variáveis da forma urbana exercem influência no comportamento dos indivíduos, sendo que setores com características que privilegiem o deslocamento de pedestres (boa largura das calçadas, declividade baixa das ruas, boa iluminação no período noturno) podem encorajar deslocamentos realizados a pé.

Daisy et al. (2018) conduziram uma pesquisa em uma grande universidade canadense considerando funcionários e alunos. Aplicando um questionário *on-line*, elaboraram diários de atividades para um dia de deslocamentos e analisaram frequências de viagens, realizadas por automóvel ou por transporte ativo (caminhadas e bicicleta), através de um modelo binomial.

Ng e Acker (2018) propuseram um estudo avaliando a influência do gênero na análise de comportamento, considerando pesquisas censitárias de 8 cidades em 3 continentes diferentes. A caracterização, avaliada para um dia de deslocamentos, se deu por um modelo de regressão logística multinomial no qual a variável dependente era a escolha modal, que alternava entre automóveis, transporte coletivo, transporte ativo, táxi e motocicletas, dependendo da localidade estudada. As variáveis independentes eram gênero, idade, renda, propósito da viagem, distância média de deslocamento e horário de partida.

Esses trabalhos contribuíram com importantes colaborações para a área de análise comportamental. Assim, o comportamento relativo a viagens em um único dia, foi carac-

terizado pelo tempo e distância média de viagens (KOCKELMAN, 1997; NG; ACKER, 2018; DAISY et al., 2018), sequência de motivos de viagens (SANTOS, 2011) e uso do solo (AMANCIO, 2005). No entanto, a utilização de apenas um único dia de dados não permite aprofundar os estudos quanto a eventuais variabilidades existentes entre os indivíduos e a influência que exercem nos seus comportamentos de atividades e viagens realizadas. Assim, graças ao advento de tecnologias para coleta de dados mais modernas, novos trabalhos passam a utilizar múltiplos dias nas suas análises.

2.2.2 Estudos utilizando dados de múltiplos dias

Huff e Hanson (1986), atuando em uma cidade sueca, constituíram uma base de dados que consistia de um conjunto de informações de dias consecutivos, coletados por um período maior que um mês. Através desses dados, valendo-se da avaliação acerca das repetições das ações realizadas pelas pessoas, conseguiram dar interpretações com respeito à variabilidade intrapessoal dos indivíduos no período considerado no estudo corroborando o estudo prévio (HANSON; HUFF, 1981) que propunha que os padrões comportamentais poderiam ser caracterizados por repetições ou por variabilidade.

Visando a compreensão dos deslocamentos com motivação para trabalho, Pas (1987), compôs uma base de dados com indivíduos que realizavam atividades remuneradas, considerando um diário de atividades de cinco dias úteis. Computando o número de viagens diárias por indivíduos, para cada um dos cinco dias, e aplicando um “*índice de variância da taxa de geração de viagens*”, os autores mensuraram quais elementos apresentavam maiores ou menores variações de geração de viagens de um dia para o outro. Essa informação de variabilidade é relevante, pois sua incorporação em modelos torna-os mais precisos uma vez que passam a contemplar a liberdade individual de variação de ações, sendo que modelos anteriores não levavam esse fator em conta. Concluiu-se, para esse estudo, que as pessoas que exercem atividades remuneradas têm padrões que independem do dia útil, isto é, o trabalho como motivação para deslocamentos, não privilegia alternância de rotas ou locais de desempenho de atividades.

Susilo e Kitamura (2005), valendo-se de uma base de dados de 6 semanas, constituída em duas cidades alemãs (AXHAUSEN et al., 2002), caracterizaram o espaço dos locais de realização de atividades dos indivíduos por meio de um “*índice de variabilidade do espaço de atividades*” e concluíram que a dimensão do espaço de atividades é influenciada pela motivação do deslocamento. Atividades obrigatórias como alimentação, estudo e trabalho confinam o espaço de realização de atividades, enquanto atividades não obrigatórias como lazer aos finais de semana estendem o espaço de realização de atividades.

Ainda utilizando a mesma base de dados, Susilo e Axhausen (2014) avaliaram o grau de repetição das escolhas individuais para realização de atividades por meio do “*índice Herfindahl-Hirschman*”. Esse índice é herdado da Economia e sua aplicação em transportes

revela os indivíduos que realizam escolhas tipicamente recorrentes no intervalo de 6 semanas de análise, frente indivíduos que realizam escolhas atípicas no mesmo intervalo de tempo.

O leitor que se interessar por maior formalismo matemático acerca das métricas “*índice de variância da taxa de geração de viagens*”, “*índice de variabilidade do espaço de atividades*” e “*índice Herfindahl-Hirschman*” deve consultar o trabalho de Pizzol (2018) em que há um maior detalhamento sobre os mesmos.

Fica claro que a utilização de múltiplos dias é benéfica para as pesquisas, especialmente aquelas que se destinam à compreensão de padrões comportamentais. Isso porque as necessidades de cada pessoa podem variar de um dia para o outro, influenciando fortemente as escolhas de rotas e tipos de deslocamentos. Repetições diárias (deslocamentos para residência, trabalho, estudo, alimentação) são extremamente relevantes aos estudos, mas podem ser identificadas mesmo nas pesquisas considerando um único dia. O ganho para múltiplos dias está justamente na identificação das atividades que não são necessariamente repetidas no cotidiano das pessoas (deslocamentos para lazer, atividades sociais, etc).

Essa identificação é facilitada por meio da aplicação das métricas supracitadas que proveem boas análises com respeito à interdependência das escolhas individuais. No entanto, ainda há algumas incertezas, como por exemplo, o “*índice Herfindahl-Hirschman*” que não leva em conta a influência que o encadeamento de viagens pode exercer nas escolhas dos indivíduos.

2.2.3 Estudos utilizando encadeamento de viagens

Dinâmicas urbanas podem também ser clarificadas por meio da construção de sequências de deslocamentos e cadeias de viagens sucedidas por modelagens paramétricas.

Indivíduos têm uma tendência natural de agrupamento de tarefas e determinação de itinerários de acordo com a conveniência, necessidade e tempo hábil para realização de suas atividades diárias (PITOMBO, 2003). Ter ciência da ordem que as atividades são realizadas embasa análises comportamentais mais precisas.

Encadeamento de viagens pode ser entendido como a forma organizada que os indivíduos encontram de realizar diversas atividades dentro do período de um dia (GOLOB, 2000), ou ainda como a cronologia de deslocamentos de um indivíduo, objetivando o desempenho de múltiplas atividades (BHAT; KOPPELMAN, 1993). Essa segunda definição, da cronologia dos deslocamentos, almejando a realização de tarefas, é adotada como premissa desta pesquisa de doutorado.

A análise do encadeamento de viagens traz à luz padrões complexos de atividades que alternam deslocamentos para destinos obrigatórios (trabalho, residência, alimentação) e não obrigatórios (lazer, passeios, compras) (TOTH et al., 1990). A consolidação de viagens alternando motivos obrigatórios e não obrigatórios, ajuda a compreender questões como a

resposta dos indivíduos a mudanças da forma urbana, demografia, influências exercidas pelo meio que estão inseridos, aumento de congestionamentos ao se perseguir várias atividades ao longo de uma única cadeia de viagens, dentre outras (AJAY; LEVINSON, 1995).

Arentze e Timmermans (2005), propõem um modelo no qual cadeias de viagens são formadas, baseadas em valores de atratividade em um mapa simulado, sendo a descrição das viagens fundamentada na quantificação dessas cadeias visando determinar o propósito para o deslocamento (atividades de trabalho, sociais, compras, lazer e recreação).

Ramadurai e Ukkusuri (2010), também atuando em um modelo aplicado a um regime simulado, compõem cadeias de atividades considerando deslocamentos a partir do domicílio para locais de trabalho ou outro propósito. Valendo-se de métricas para redes e funções utilidade, estabelecem qual o padrão de equilíbrio para a demanda de atividades e transportes na rede estudada.

Pitombo, Kawamoto e Sousa (2011), utilizando uma pesquisa Origem/Destino, relativa a um dia de deslocamentos, definem padrões comportamentais através do estabelecimento de cadeias de deslocamentos usando codificações para o modo de deslocamento (privado, coletivo ou não-motorizado), motivação para o deslocamento (residência, trabalho, estudo, outras atividades) e a distância dos deslocamentos. As sequências são agrupadas em padrões distintos por meio de uma árvore de decisão.

Ainda que esses trabalhos apresentem importantes constatações, foram baseados em dados seccionais compostos majoritariamente por questionários, entrevistas origem-destino e dados censitários, não contemplando assim os dados em painel de múltiplos dias nem as cadeias de deslocamentos diários.

Sugere-se, neste trabalho, que a composição de cadeias de viagens seja baseada em dados em painel de múltiplos dias, para que se extraia o caráter de variabilidade (que é uma das vantagens das análises de múltiplos dias frente as de único dia) e associe-o à composição do encadeamento de viagens, a fim de contemplar também eventuais influências advindas da sequência de movimentações do cotidiano das pessoas.

3 ANÁLISE ESPACIAL DE DADOS

Esse capítulo apresenta as definições e o estado da arte acerca de análise espacial de dados. A análise espacial de dados é realizada através de duas etapas, sendo a primeira exploratória e a segunda confirmatória. Ao longo do Capítulo 3, o leitor é apresentado a um conjunto de ferramentas exploratórias e confirmatórias que aplicam-se aos diversos tipos de dados. Ao final, são apontadas as ferramentas exploratórias e confirmatórias escolhidas para o desenvolvimento do procedimento proposto neste trabalho.

3.1 Definição e o estado da arte

Análise espacial de dados consiste em um conjunto de técnicas que se destinam a analisar a localização geográfica de determinadas observações, avaliando as possíveis relações entre os valores das variáveis em diferentes localizações. Diversas questões podem ser respondidas mediante análises espaciais. A compreensão da distribuição espacial de determinadas informações, que ocorrem no espaço, é pertinente aos estudos de diversas áreas tais como saúde e epidemiologia (REIS et al., 2008; REZAEIAN et al., 2007; PFEIFFER et al., 2008), agricultura (CAMBARDELLA; KARLEN, 1999; MORGAN et al., 1978; PEROBELLI et al., 2007), meio-ambiente (JERRETT et al., 2005; BORCARD; GILLET; LEGENDRE, 2011), segurança pública (PAULSEN; ROBINSON, 2004; ANDRESEN, 2006; MURRAY et al., 2001; BATELLA; DINIZ, 2010), dentre outras. Para a análise de demanda por transportes, diversas decisões individuais ocorrem no espaço ou são influenciadas pela localização dos indivíduos.

Segundo Anselin (2010), os primeiros estudos na área foram possíveis graças aos avanços da geografia que possibilitaram a composição de modelos capazes de identificar a estrutura espacial das relações (BERRY; MARBLE, 1968). Entendida a estrutura espacial, o conceito evoluiu na década de 1970 quando Tobler (1970) estabelece a “*Primeira Lei da Geografia*” que diz “ainda que todos os elementos num determinado espaço se relacionem entre si, no entanto, elementos próximos estão mais relacionados quando comparados com elementos distantes”.

Hoje, considera-se que a análise espacial pode ser realizada em três possíveis tipos de dados (CÂMARA et al., 2002):

- Eventos ou padrões pontuais – ocorrências pontuais, localizadas no espaço, como por exemplo, localização de um crime, ocorrência de uma espécie de um animal, ocorrência de elementos com igual característica, etc.
- Superfícies contínuas – conjunto de amostras que se estende por uma determinada

área de ocorrência, como por exemplo, mapas ecológicos, mapas topográficos, etc.

- Áreas com taxas agregadas – dados que foram coletados pontualmente, mas que representam uma determinada área no espaço. A razão para essa ocorrência varia, sendo o clássico exemplo o dos dados censitários em que a agregação é realizada para manutenção da confidencialidade dos indivíduos entrevistados.

A análise espacial de um fenômeno compreende uma etapa preliminar, exploratória, na qual é estimado ou mensurado o grau de dependência espacial, de forma local ou global, de um conjunto de dados. Corroborada a dependência espacial dos dados, é indicada a etapa de análise espacial confirmatória, na qual modelos que incluem a estrutura espacial dos dados são utilizados.

3.2 Análise espacial exploratória

A análise exploratória pode ser realizada de variadas formas, dependendo da natureza de dados em que se aplica. Algumas técnicas são apresentadas para cada um dos três tipos de dados espaciais possíveis.

3.2.1 Análise espacial exploratória de eventos ou padrões pontuais

Eventos ou padrões pontuais são ocorrências, localizadas no espaço, que têm como alvo de investigação a detecção de possíveis agrupamentos, considerando eventos semelhantes demasiadamente próximos, a identificação de desvios comportamentais em uma distribuição teoricamente uniforme e a existência de aleatoriedade na distribuição dos eventos. Em resumo, analisam a distribuição espacial de pontos para classificar os padrões como aleatórios, agrupados ou regulares (GUO et al., 2013).

As análises espaciais exploratórias, para pontos, são conduzidas em um âmbito global, por meio de estimadores de intensidade. Já a identificação de desvios é compreendida em um âmbito local por meio de ferramentas de análise de dependência espacial. Como exemplo de estimador de intensidade, cita-se o estimador de agrupamento (*Kernel Estimator*)(BAILEY; GATTREL, 1995) e como exemplos de análise de dependência espacial cita-se o método do vizinho mais próximo (CLARK; EVANS, 1954) e a função K de Ripley (RIPLEY, 1976).

Sabe-se que a classificação dos padrões espaciais é dependente da escala, isto é, os pontos podem apresentar padrão de agrupamento em uma escala, mas serem considerados dispersos quando vistos em outras escalas. Entre o conjunto de métodos baseados em pontos, a função K de Ripley é amplamente utilizada porque descreve padrões de pontos em escalas de distância múltipla (GUO et al., 2013).

3.2.2 Análise espacial exploratória de superfícies contínuas

Dados de superfícies contínuas podem ser valores que se estendem por uma determinada área de ocorrência mas também podem ser provenientes de ocorrências pontuais que foram interpoladas para gerar tais superfícies. Diversos métodos de interpolação são possíveis destacando-se as interpolações por vizinhos próximos, por média simples e média ponderada. Investigações pertinentes a esses tipos de dados são análises de tendência, correlação e variância.

A avaliação acerca das tendências em uma determinada superfície pode ser feita por ferramentas como a estimação de vetores de médias de valores ponderadas pelo tamanho da superfície e a estimação de vetores de desvios (CÂMARA et al., 2002; LOPES, 2005). Segundo Oden (1984), um descritor poderoso para dados de superfícies são os correlogramas.

A estatística espacial abrange um conjunto de ferramentas matemáticas que provê critérios quantitativos que podem ser usados na etapa exploratória. Um exemplo são os métodos que analisam estatísticas de médias, frequência e contagens de valores para determinar características semelhantes numa superfície (GRIFFITH, 2012).

A Geoestatística apresenta-se como um importante método de estatística espacial para superfícies. Desenvolvida por Krige (1951), tornou-se amplamente utilizada pois baseia-se na variância como métrica para análise estatística espacial, diferenciando-se dos demais métodos por considerar a distância entre pontos analisados nos cálculos. Em caráter exploratório, destacam-se como ferramentas geoestatísticas os variogramas experimentais e suas modelagens teóricas.

3.2.3 Análise espacial exploratória de áreas com taxas agregadas

Dados de áreas com taxas agregadas são aqueles onde uma amostra pontual representa uma determinada área. Dessa forma a localização exata do evento não mais é relevante, mas sim seu valor, que passa a representar a área agregada em que este está contido. São pertinentes a esse tipo de dado investigações acerca de possíveis áreas com características comuns e a presença ou ausência de autocorrelação de valores em áreas adjacentes.

A análise de áreas com características comuns é realizada por métodos que utilizam estatísticas de frequência de ocorrência de eventos e/ou características em determinadas regiões do espaço, como o trabalho de Teixeira (2003) que, utilizando dados censitários, realiza contagens e análise de frequências a fim de identificar regiões com homogeneidade de características. Outro exemplo de análise exploratória almejando a identificação de áreas com atributos semelhantes é a matriz de proximidade espacial, ou matriz de vizinhança, usada para estimação da variabilidade espacial de dados de área (LOPES, 2005; CÂMARA

et al., 2002).

Entretanto, os métodos mais usuais são aqueles voltados à análise acerca da presença ou ausência de autocorrelação de áreas adjacentes, destacando-se nesse contexto os indicadores de autocorrelação espacial: Moran Global (MORAN, 1950), Moran Local (ANSELIN, 1995), SivarG (NAIZER, 2018), Teste C de Geary (GEARY, 1954), etc.

Avaliando a tendência que os pontos no espaço têm de serem similares aos valores dos pontos vizinhos e dissimilares dos valores de pontos distantes, os indicadores globais atuam de maneira a retornar um único valor como medida de associação espacial relativa ao conjunto de dados analisados. Indicadores locais têm papel semelhante, mas são associados a escalas menores pois estas demandam maior detalhamento (ANSELIN, 1995; MORAN, 1950; NAIZER, 2018).

Os índices citados fazem parte da ESDA (*Exploratory Spatial Data Analysis*), além de ferramentas de visualização. Alguns dos métodos de visualização são o diagrama de espalhamento de Moran, Box Map, Lisa Map, Moran Map e Gráfico de Barras. Tais ferramentas foram utilizados em diversos trabalhos (CÂMARA et al., 2002; SERRANO; VALCARCE, 2000; KREMPI, 2004; SPRING, 2006) com o intuito de promover auxílio às análises exploratórias por meio de visualização em mapas.

Este trabalho faz uso de um indicador global tradicional da ESDA, Moran global, e também testa um indicador baseado em ferramentas geoestatísticas, SivarG.

3.3 Análise espacial confirmatória

A análise espacial confirmatória engloba o conjunto de modelos de estimação, além de procedimentos de validação. São realizadas análises multivariadas ou univariadas com componentes espaciais. Sendo a análise confirmatória, a continuidade do processo prévio, é natural que as ferramentas exploratórias tenham associação com ferramentas confirmatórias. Mais uma vez, técnicas são apresentadas para cada um dos três tipos de dados espaciais possíveis.

3.3.1 Análise espacial confirmatória de eventos ou padrões pontuais

Dados pontuais comumente têm um perfil binomial, isto é, expressam a presença ou ausência de um determinado evento em uma localização. Portanto, é sugerido nesses casos o uso de modelos confirmatórios do tipo regressão logística, sendo que a variável dependente deve ser a dicotomia de presença ou ausência de uma determinada ocorrência e as variáveis independentes devem ser provenientes da etapa exploratória (intensidades, desvios, autocorrelação, etc) (HOSMER; LEMESHOW; STURDIVANT, 2013).

Como exemplos de aplicações, pode-se mencionar Câmara et al. (2002), que utilizam o método confirmatório de regressão logística para avaliar a dicotomia de mortalidade

infantil no estado do Rio Grande do Sul, baseado nos dados do trabalho original de Hernandez et al. (2011). Já Milton, Shankar e Mannering (2008) utilizam o mesmo método confirmatório para a dicotomia de acidentes em rodovias.

3.3.2 Análise espacial confirmatória de superfícies contínuas

Para dados em superfícies contínuas, merece destaque a ferramenta geoestatística confirmatória denominada krigagem. Uma vez atestada a estrutura espacial em etapa exploratória (cálculo e modelagem de variogramas), a ferramenta promove um processo de estimação linear como uma função dos valores de variáveis próximas, utilizando parâmetros advindos da análise geoestatística prévia e, portanto, aplica-se somente a esse método de análise espacial. Entretanto, é relevante frisar que os parâmetros podem ser estabelecidos por diversas funções teóricas diferentes (ISAAKS; SRIVASTAVA, 1989).

Existem algumas modalidades de krigagem, sendo que os métodos mais comuns de krigagem univariada são krigagem simples, krigagem universal e krigagem ordinária (WACKERNAGEL, 2014). É utilizado como método de análise confirmatória em uma grande quantidade de trabalhos de análise espacial (GOMES et al., 2016a; GOOVAERTS, 2006; AHMADI; SEDGHAMIZ, 2007; GOOVAERTS, 1992; GOMES et al., 2016b).

3.3.3 Análise espacial confirmatória de áreas com taxas agregadas

Para os dados em áreas, destaca-se a regressão espacial como ferramenta confirmatória. Modelos de regressão, em geral, são muito utilizados em diversos campos da engenharia sendo que, em análise de demanda por transportes, também se fazem presentes, na etapa de geração de viagens. A regressão consiste da estimação de valores de uma variável dependente através de outras variáveis explicativas. No caso da análise espacial, se confirmada a autocorrelação espacial na etapa preliminar, pode-se realizar estimativas acerca dessa autocorrelação incorporando a estrutura espacial dos dados como variáveis explicativas do modelo (FOTHERINGHAM; BRUNSDON; CHARLTON, 2000). Existe uma versão local da regressão espacial que é denominada Regressão Geograficamente Ponderada. Difere-se da abordagem tradicional pois gera parâmetros desagregados pelas unidades espaciais de análise, permitindo avaliar a heterogeneidade espacial nas relações estimadas entre a variável dependente e as variáveis explicativas (FOTHERINGHAM; CHARLTON; BRUNSDON, 1998).

3.3.4 Análise espacial confirmatória através de modelos de escolha discreta

Além das ferramentas citadas, outra ferramenta confirmatória, que tem sido muito utilizada, são os modelos de escolha discreta. Modelos de escolha discreta avaliam a probabilidade de um indivíduo escolher uma opção entre um conjunto de alternativas discretas. Tais modelos contrastam com modelos de consumo padrão nos quais a quantidade

consumida por um indivíduo é assumida como uma variável contínua. No caso contínuo, métodos de cálculo de primeira ordem podem ser utilizados na determinação da quantidade ótima da variável dependente, sendo sempre possível realizar a modelagem empiricamente por meio de análises de regressão. Por outro lado, a análise de escolha discreta examina situações onde o resultado (variável dependente) é discreto, de modo que o ótimo não pode mais ser caracterizado por condições de primeira ordem. Assumindo que os modelos de escolha discreta relacionam estatisticamente a escolha (variável dependente discreta) aos atributos dos indivíduos e das alternativas disponíveis, deve-se recorrer a modelagens realizadas através de métodos de máxima verossimilhança paramétricos. O leitor que buscar maior formalismo matemático acerca da técnica deve consultar o trabalho de Ben-Akiva e Lerman (1985).

Modelos de escolha discreta ganharam popularidade como ferramental útil para o estudo e modelagem de diversos tipos de escolha nas mais variadas áreas tais como escolha residencial (BHAT; GUO, 2004; BEKHOR; PRASHKER, 2008), *marketing* e gestão de recursos (CHAKRAVARTI; SINHA; KIM, 2005), interações interpessoais (BROCK; DURLAUF, 2001; PÁEZ; SCOTT; VOLZ, 2008), dentre outros.

Escolhas entre modos de transporte (BHAT, 2008; ANAS, 1982; HESS, 2005) são outro clássico exemplo de escolha discreta, onde se deve escolher uma alternativa dentro de um conjunto finito de possibilidades (automóvel, transporte público, caminhada, etc.). Os atributos das alternativas modais (como por exemplo, tempo de viagem, custo e tarifação, etc.) e os atributos dos indivíduos (renda, idade, sexo, etc.) são utilizados para cálculo de probabilidades de escolha dos modos de transporte possíveis.

Entretanto, para o caso de demanda por transportes, vale ressaltar que, na prática, é muito difícil conhecer todos os fatores que afetam as decisões individuais, uma vez que existe uma heterogeneidade subjacente ao comportamento dos indivíduos. Além disso, as localizações domiciliares e das atividades de destino podem influenciar as escolhas modais, bem como os atributos individuais e das alternativas. Desta forma, a incorporação de variáveis relacionadas à localização geográfica aos estudos de previsão de demanda por transportes torna-se uma importante ferramenta para o aprimoramento das estimativas (SMIRNOV, 2010; PITOMBO et al., 2015), uma vez que as decisões são sabidamente influenciadas pela existência de autocorrelação espacial, que deve ser atestada previamente por algum método exploratório.

Assim, muitos trabalhos passaram a levar em conta o aspecto espacial nos modelos de escolha discreta (KOPPELMAN; SETHI, 2005; MOHAMMADIAN; HAIDER; KANA-ROGLOU, 2005; SMIRNOV, 2010; BILLÉ; ARBIA, 2013; PÁEZ et al., 2013; PITOMBO et al., 2015), almejando maior precisão aos seus modelos ao considerar a influência que variáveis adjacentes exercem nos seus dados de estudo. De cunho geral, modelos de escolha discreta aplicam-se como ferramenta confirmatória para quaisquer tipos de dados, desde

que a variável dependente seja baseada num conjunto finito de opções e que o caráter espacial seja incorporado às variáveis independentes.

3.4 Análise espacial exploratória e confirmatória aplicadas ao doutorado

Indicadores globais de autocorrelação compõem a análise espacial exploratória desse estudo. Nesse trabalho se utilizou os indicadores globais Moran e SivarG. O indicador global Moran é baseado em correlogramas (também conhecido como diagramas de autocorrelação, são gráficos das autocorrelações da amostra *versus* unidades de distância). Indica a similaridade dos dados e varia de -1 até 1. Valores negativos indicam autocorrelação espacial negativa, ao passo que valores positivos indicam autocorrelação espacial positiva. O valor 0 indica a ausência de correlação espacial (MORAN, 1950; ANSELIN, 1995).

O indicador global SivarG (*Global Spatial Indicator Based on Variogram*) é baseado em pressupostos da geoestatística, mais especificamente na ferramenta semivariograma (teórico e experimental). O semivariograma permite representar a variação quantitativa das variáveis regionalizadas através do cálculo da semivariância segundo a distância entre pontos, medindo a dissimilaridade entre os dados. Quanto menor o valor da semivariância, mais parecidos são os dados entre os pontos analisados. Portanto, se há dependência espacial é natural que o valor da semivariância cresça com a distância e se estabilize em um patamar, a partir do qual não há mais dependência espacial (BURROUGH, 1986). O semivariograma é dado pela Equação 3.1.

$$\gamma(h) = \frac{1}{2N} \sum_{i=1}^N [Z(x+h) - Z(x)]^2, \quad (3.1)$$

em que $\gamma(h)$ é função semivariograma em h ; h é distância entre variáveis; $Z(x)$ a variável aleatória Z em x ; N número de elementos contabilizados.

Para cálculo de variograma experimental, é necessário determinar os valores de ângulo (direção), distância entre pontos (*lag-distance*) e tolerâncias de banda e angular. Os pares de pontos utilizados para o cálculo incluem o ponto de origem do vetor e os pontos dentro da área delimitada pelas tolerâncias. Todos os pontos são analisados para distâncias múltiplas de h , de forma a ser possível construir um gráfico relacionando variância γ e distância entre os pares (h).

A modelagem dos variogramas experimentais é realizada a partir da análise do melhor modelo teórico que se ajusta aos pontos do semivariograma experimental. Os modelos mais usuais são: efeito pepita puro (patamar constante indicando ausência de autocorrelação espacial), esférico, exponencial e gaussiano. Os pontos notáveis do variograma teórico (ajustado) são : alcance (a), corresponde à distância até a qual há correlação espacial; patamar (C), corresponde ao valor da semivariância no alcance; efeito pepita (C0), efeito observado devido às discontinuidades do semivariograma para pequenas distâncias; contribuição (C1), diferença entre o patamar e o efeito pepita, corresponde à contribuição do modelo teórico (Figura 1).

Figura 1: Esquema semivariograma teórico/experimental. Adaptado de Burrough (1986).

O indicador SivarG é baseado nos valores padronizados do variograma teórico (eixo das ordenadas), associados a um teste de hipótese para aleatoriedade espacial. Indica a dissimilaridade dos dados e varia de 0 até 1, onde o valor 1 indica a completa ausência de autocorrelação espacial e o valor 0 indica completa presença de autocorrelação espacial. Os trabalhos originais de Moran (1950), Naizer e Pitombo (2017) e Naizer (2018) devem ser consultados caso o leitor deseje maiores detalhes sobre o formalismo matemático empregado.

Wise, Haining e Signoretta (1998) e Lopes (2005) frisam que, na análise espacial exploratória, os métodos empregados são descritivos e não confirmatórios, destinando-se apenas a detectar padrões espaciais e permitir a elaboração de hipóteses preliminares. Faz-se necessário, portanto, confirmar essa etapa exploratória.

Assim sendo, a etapa de análise espacial confirmatória, relativa a este trabalho, consiste na calibração de dois modelos de escolha discreta (*logit* multinomial): o primeiro utilizando somente as variáveis originais do banco de dados (características dos indivíduos e das alternativas); e um segundo com o acréscimo de covariáveis espaciais, cuja utilização justifica-se pela análise espacial exploratória prévia. A incorporação das covariáveis espaciais ao modelo *logit* segue a tendência dos recentes trabalhos de levar em conta os aspectos espaciais, sendo que as covariáveis incluídas são extraídas por um método baseado nos preceitos geoestatísticos, que é outro notório ferramental de análise espacial.

4 MATERIAIS E MÉTODO

Este capítulo versa sobre os materiais e método da tese. Como materiais foram utilizadas duas bases de dados: A base de dados em painel, principal conjunto de dados dessa tese, que foi obtida por coleta via *smartphones* e utilizada em todas as etapas desse trabalho. A base de dados da cidade sintética, secundária e advinda de um estudo prévio, foi utilizada como protótipo para as etapas de modelagem paramétrica desse trabalho.

O método apresenta todos os passos para a realização dos objetivos, desde a obtenção dos dados de deslocamentos por meio de telefones móveis, passando pelo tratamento dos mesmos, sua caracterização com respeito aos padrões comportamentais dos indivíduos e culminando na modelagem paramétrica com a incorporação de covariáveis espaciais.

Ao término do capítulo, apresentam-se todas as ferramentas e aplicativos necessários ao desenvolvimento do projeto, avaliando suas características e apontando o momento de sua utilização.

4.1 Materiais

4.1.1 Base de dados principal: Dados em painel provenientes de *Smartphones*

A base de dados principal deste doutorado consiste dos dados em painel, provenientes do uso de *Smartphones*. Através desse conjunto de dados, todas as etapas metodológicas desse doutorado foram conduzidas, incluindo análise espacial e proposta de modelagem paramétrica com incorporação de covariáveis espaciais.

Este trabalho foi conduzido na cidade de São Carlos – SP (Figura 2). Segundo o último levantamento do Instituto Brasileiro de Geografia e Estatística (IBGE), o município tem uma população de 221.950 pessoas, possui área territorial de 1.136,907 km² e PIB per capita de R\$40.435,49. A cidade abriga duas grandes instituições de ensino superior: a Universidade de São Paulo (USP) que conta com dois campi e a Universidade Federal de São Carlos (UFSCar). Juntas, as instituições possuem cerca de 40 mil estudantes, o que representa aproximadamente 18% da população do município. A Figura 3, em seguida, apresenta um mapa do município exibindo a localidades dos campi citados.

Figura 2: Localização do município de São Carlos, SP – Brasil.

Figura 3: Mapa do município de São Carlos, SP – Brasil, apresentando a localização dos campi da Universidade de São Paulo e o campus da Universidade Federal de São Carlos.

O estabelecimento da nova proposta de coleta de dados e posterior agrupamentos baseados na análise comportamental individual relativa a viagens se deu por meio de amostragem por oportunidade. Contou-se com a participação de 104 alunos, pertencentes ao curso de graduação ou pós-graduação em engenharia civil da Escola de Engenharia de São Carlos, localizada no campus 1 da Universidade de São Paulo. Os alunos, voluntariamente, colaboraram com a coleta de informações durante um período de três dias úteis (terça, quarta e quinta-feira) do mês de agosto de 2017.

A escolha do mês de agosto se deu por ser um período letivo, com ausência de feriados, época conveniente para a realização de análises acerca de estudantes. A escolha dos dias úteis é baseada no trabalho de Ribeiro e Cunha (2017). Os autores afirmam que, com base numa análise de dados históricos para tráfego de veículos, os dias de terça, quarta e quinta-feira têm comportamentos similares, sendo que para meses sem feriados, refletem padrões bastante distintos dos dias de final de semana e das segundas e sextas-feiras.

A coleta promovida gerou um conjunto de dados brutos que indicavam todos os registros de movimentações realizados pelos indivíduos, caracterizados pela latitude/longitude e data/hora.

4.1.2 Base de dados auxiliar: A cidade fictícia

Para fins de validação metodológica, sobretudo em relação ao objetivo principal do trabalho, foi utilizada uma base de dados sintéticos com dados espacialmente dependentes.

A cidade fictícia, utilizada nesse trabalho, foi criada no trabalho de Assirati et al. (2016) e consiste de 580 indivíduos georreferenciados, localizados em um plano com 10 unidades de distância no eixo X e 5 unidades de distância no eixo Y. Além das coordenadas, cada um dos pontos da amostra é dotado de outras sete variáveis: Distância média de viagens para realização de atividades (em unidades de distância); Tempo de viagem para o modo de transporte ônibus (em min.); tempo de viagem para o modo de transporte automóvel (em min.), tarifa do ônibus (em reais), custos do uso do automóvel (em reais), indicador de renda (escala de 1 – menor renda, até 10 – maior renda) além de uma variável binária responsável por indicar a escolha modal do indivíduo: (0) uso de ônibus ou (1) uso do automóvel. A Tabela 1 apresenta medidas descritivas das variáveis e a Figura 4 apresenta a distribuição geográfica da variável binária.

Tabela 1: Medidas descritivas da base de dados auxiliar – Cidade Fictícia

	X	Y	Distância (Km)	Tempo ônibus (Min.)	Tempo Carro (Min.)	Preço ônibus (Reais)	Preço Carro (Reais)	Renda (Escala 1 a 10)
Média	5,2	2,68	2,47	5,4	3,7	1,99	2,1	5,43
Desvio Padrão	2,93	1,45	1,31	2,58	1,97	0,7	1,11	1,47
Valor Máximo	10	5	5,52	11,12	8,29	3	4,7	10
Valor Mínimo	0	0,15	0,05	1,08	0,07	1,5	0,04	1

Figura 4: Representação espacial do binário de escolha modal na cidade sintética.

4.2 Método

O fluxograma, apresentado na Figura 5, ilustra as etapas metodológicas seguidas neste trabalho.

Figura 5: Fluxograma das etapas metodológicas.

4.2.1 Coleta de dados via *smartphones*

A coleta de dados foi realizada por meio de aparelhos *smartphone*. Para tal, o usuário ativou a ferramenta de “histórico de localização” no aplicativo de mapas de seu dispositivo. Com tal ferramenta ativa, o telefone passa a acumular e armazenar informações de horário, latitude e longitude, referentes aos deslocamentos realizados no período de três dias. O processo é feito de maneira automática, não havendo a necessidade de invocar a aplicação a todo momento, nem da tela permanecer ligada. O indivíduo, realizando suas viagens em posse do aparelho, é suficiente para que o trajeto seja arquivado, compondo um rico e preciso banco de dados.

Difere-se, substancialmente, da abordagem tradicional de diário de atividades por questionário por três motivos. O primeiro é o de que os questionários geralmente contemplam apenas um dia de atividades, já a nova abordagem abrange três dias úteis de atividades. O segundo é o de que, por muitas vezes, informações úteis ao desenvolvimento do projeto são perdidas devido aos esquecimentos por parte do indivíduo no ato da entrevista. Esse esquecimento é eliminado na aquisição proposta, pois desde que os deslocamentos sejam feitos portando o *Smartphone*, a coleta contemplará todas as movimentações sem perda de dados. Finalmente, o terceiro refere-se à questão do tempo dos deslocamentos. Na modalidade de entrevista, é comum o entrevistado prover informações imprecisas pois pode ter uma percepção equivocada dos tempos de deslocamento. Nessa proposta, essa imprecisão é eliminada pois a posse do *Smartphone* durante as movimentações computa o tempo real despendido. Vale ressaltar como principal desvantagem da coleta proposta a não aferição, por exemplo, do tipo de moto de transporte motorizado ou/e motivo de viagem.

Ao término dos registros de três dias, o usuário é instruído a submeter seus dados em caráter de anonimato em uma plataforma *on-line*, implementada para acumular os dados submetidos e compor o painel. Durante essa submissão, um questionário socioeconômico (compreendendo informações sobre a idade, o sexo, o grau de escolaridade, a renda, o tipo de moradia, a execução de atividade remunerada, a quantidade de automóveis disponíveis na residência e a quantidade de motocicletas disponíveis na residência) é preenchido pelos indivíduos. A associação da coleta de dados via *smartphones* aos questionários *on-line* são apontadas por Bricka e Bhat (2006) e Schönfelder et al. (2002) como uma excelente alternativa para se obter maior precisão na definição do número total de viagens e do tempo de deslocamentos coletados.

4.2.1.1 Pré-processamento dos dados brutos

Ao receber os dados brutos, é necessário pré-processar esses dados a fim de torná-los compatíveis com os rumos que se deseja para a pesquisa. Seguindo técnicas apresentadas no trabalho de Anda, Erath e Fourie (2017), desenvolveu-se um algoritmo para a clusterização de tempos e distâncias para identificar os locais de atividades dos indivíduos a partir de seus dados brutos.

Quando um indivíduo se encontra em uma determinada localidade por muito tempo (por exemplo um aluno no campus da universidade), este gera uma quantidade grande de registros (pontos georreferenciados) espacialmente próximos. É preciso então aglutinar esses registros todos, pois referem-se à mesma localidade em um grande intervalo de tempo.

Para promover essa aglutinação, define-se uma cerca geográfica (no caso desse estudo, o valor foi de 150 metros). Os pontos aglutinados são aqueles que estão compreendidos dentro desta cerca, que tem seu centro nas coordenadas do primeiro ponto de uma suposta

localidade de atividades.

Após, compara-se os registros de horários dos pontos aglutinados. Se a diferença entre o primeiro e o último for maior que um determinado limite temporal (neste caso, o valor foi de 10 minutos), considera-se aquela aglutinação como um local de realização de atividades.

Por fim, substitui-se a aglutinação por um único registro, geolocalizado nas coordenadas do primeiro ponto, e com duração temporal começando no horário do primeiro ponto e se encerrando no horário do último. A Figura 6 apresenta um exemplo do processo de clusterização, partindo de um estado (A) com vários pontos compreendidos dentro da mesma cerca geográfica até o estado (B), aglutinados em um único ponto com longo tempo de permanência na localidade.

Figura 6: Exemplo de clusterização de tempos e distâncias para identificação dos locais de atividades dos indivíduos: (A) dados brutos; e (B) resultado do pré-processamento, com o local de atividade identificado.

Reforça-se que o algoritmo tem cunho genérico e deve ser adaptado para a localidade que pretende-se estudar. Cidades diferentes demandam cercas geográficas maiores ou menores e limites temporais compatíveis com o nível de tráfego da região por exemplo.

Após esse pré-processamento, conta-se com uma base de dados com as informações originais da coleta, pronta para ter informações extraídas visando a composição do painel.

4.2.1.2 Informação original

Dispondo do banco de dados, devidamente pré-processado, deve-se compor o painel, derivando informações a partir dos dados georreferenciados submetidos. Em sua forma original, os dados recebidos contêm as informações de data e hora além da latitude e longitude. A Figura 7 exemplifica dois registros de dados, do mesmo indivíduo (ID), em sua forma original.

ID	Data / Hora	Latitude	Longitude
1	08/08/2017 08:00	-22.005013	-47.894870
1	08/08/2017 08:15	-22.011498	-47.892853

Figura 7: Exemplo de dados originais submetidos pelos participantes.

A partir dos dados obtidos originalmente, são derivadas as seguintes informações: tempo e distância de deslocamentos, velocidade média do deslocamento, recorrência dos locais de atividades e a escolha do modo de transporte motorizado ou não-motorizado. A seguir, essas informações embasam a composição das cadeias de viagens cuja quantificação é parâmetro de entrada para o algoritmo de agrupamento. A obtenção das informações derivadas, além do método de composição das cadeias de viagens e de sua quantificação são descritos nas próximas subseções.

4.2.1.3 Tempo e distância de deslocamento

A partir dos dados originais é possível extrair duas informações: o tempo de deslocamento entre duas localidades diferentes e a distância desse. O tempo (em minutos) é obtido pela simples subtração dos horários de partida da localidade inicial e chegada da localidade final. Já a distância (em metros) do deslocamento é dada pela dimensão da ortodromia entre os pontos, segundo as coordenadas dos pontos de origem e destino. A Figura 8 exemplifica a derivação do tempo e da distância a partir dos registros exemplo.

ID	Data / Hora	Latitude	Longitude	Tempo (min)	Distância (m)
1	08/08/2017 08:00	-22.005013	-47.894870	15	1000
1	08/08/2017 08:15	-22.011498	-47.892853		

Figura 8: Exemplo de obtenção de tempo e distância de deslocamento.

4.2.1.4 Estimação de velocidade média de deslocamento

A seguir, estima-se a velocidade média do deslocamento (km/h), que é calculada a partir da razão entre as variáveis distância e tempo, com conversões de unidade apropriadas. A Figura 9 ilustra a obtenção dessa informação.

ID	Data / Hora	Latitude	Longitude	Tempo (min)	Distância (m)	Veloc. (Km/h)
1	08/08/2017 08:00	-22.005013	-47.894870	15	1000	4
1	08/08/2017 08:15	-22.011498	-47.892853

Figura 9: Exemplo de estimação da velocidade média do deslocamento.

4.2.1.5 Análise de recorrência das localidades

Ao se estudar padrões comportamentais de deslocamento, é relevante ter ciência se a movimentação foi feita para um local recorrente ou não-recorrente, isto é, distinguir as localidades que o indivíduo frequenta corriqueiramente (por exemplo a universidade para o caso de alunos universitários) das localidades que o indivíduo frequenta esporadicamente (cinema, *shoppings*, parques, etc.), já que não é possível a aferição do motivo de viagem. Para a realização dessa tarefa, foram examinados todos os pontos georreferenciados, por participantes, buscando por repetições. Caso a localidade se repetisse no intervalo de três dias, era rotulada como “R” por ser um local recorrente. Caso a localidade fosse única no mesmo intervalo de tempo analisado, era rotulada como “N” de não-recorrente. A Figura 10 exemplifica a detecção de um local recorrente.

ID	Data / Hora	Latitude	Longitude	Tempo (min)	Distância (m)	Veloc. (Km/h)	Recorrência
1	08/08/2017 08:00	-22.005013	-47.894870	15	1000	4	N
1	08/08/2017 08:15	-22.011498	-47.892853	7	1300	11,2	R
1	08/08/2017 08:22	-22.006087	-47.897831	9	1300	8,7	N
1	08/08/2017 08:31	-22.011498	-47.892853	3	500	10	R

Figura 10: Exemplo de obtenção de recorrência das localidades.

4.2.1.6 Caráter de motorização individual

Finalmente, o painel de dados se completa com o caráter de motorização acerca dos deslocamentos de cada indivíduo. O caráter de motorização distingue os deslocamentos entre motorizados e não-motorizados e é baseada na velocidade média do deslocamento.

Admitiu-se que uma pessoa caminha com velocidade média de 6 km/h. Assim, examinando todos os registros coletados, considerou-se que os deslocamentos que têm velocidade média igual ou inferior aos 6 km/h foram feitos por modos de transporte não-motorizados, sendo rotulados com o binário “0”. Já os deslocamentos com velocidade média superior a 6 km/h foram rotulados com o binário “1”, indicando que a movimentação se deu por modo de transporte motorizados. A Figura 11 exibe exemplos de deslocamentos motorizados e não-motorizados.

ID	Data / Hora	Latitude	Longitude	Tempo (min)	Distância (m)	Veloc. (Km/h)	Recorrência	Motorização
1	08/08/2017 08:00	-22.005013	-47.894870	15	1000	4	N	0
1	08/08/2017 08:15	-22.011498	-47.892853	7	1300	11,2	R	1
1	08/08/2017 08:22	-22.006087	-47.897831	9	1300	8,7	N	1
1	08/08/2017 08:31	-22.011498	-47.892853	3	500	10	R	1

Figura 11: Exemplo de obtenção da variável "*Caráter de motorização individual*".

4.2.1.7 Validação da coleta de dados através de diário de atividades tradicional

A validação da coleta de dados é realizada comparando os dados coletados pela abordagem via *Smartphones* com o método de entrevistas e composição de diários de atividades, consagrado método de identificação de dinâmicas cotidianas.

Como essa etapa almeja averiguar a qualidade das informações (quantidade de viagens, tempo e distâncias de deslocamentos, etc.) extraídas pelos algoritmos *Python* dos dados originais, decidiu-se pela composição de um grupo de 10 novos indivíduos, de perfil semelhante aos 104 indivíduos da coleta de dados que compôs o painel. Isso porque esses 10 novos indivíduos, além de realizarem o procedimento de coleta de dados via *Smartphone*, também responderam uma entrevista presencial, a fim de compor um diário de suas atividades.

Na entrevista presencial deve-se indagar os indivíduos quanto aos seus deslocamentos realizados no dia anterior: quantidade de viagens realizadas, distância e tempo dispendidos em cada deslocamento e o modo de transporte (motorizado ou não-motorizado) utilizado.

Após, para os mesmos indivíduos e o mesmo dia de referência, se extrai, via algoritmos computacionais propostos, informações análogas às indagadas na entrevista. Em posse dos resultados, apresenta-se o relatório emitido pelos algoritmos para o entrevistado e compara-se com as respostas que o mesmo proveu em tempo de entrevista. Assim é possível identificar eventuais esquecimentos ou imprecisões no provimento de informações via entrevistas.

Então, testes de Mann-Whitney e mediana são aplicados às variáveis “quantidade total de viagens”, “soma total de distâncias”, “soma total de tempos”, “quantidade total de viagens motorizadas” e “quantidade total de viagens não-motorizadas” para a verificação da similaridade dos dados declarados e computados quanto a distribuições populacionais e valores típicos.

4.2.2 Construção das cadeias de viagens

Após o estabelecimento do painel completo e realizadas a validação da coleta, iniciou-se o processo de análise e investigação acerca dos comportamentos individuais. Para tal, estabeleceu-se as cadeias de viagens, que são um encadeamento dos deslocamentos, realizados por cada indivíduo no período de três dias.

Para cada indivíduo, a cadeia é formada concatenando os pares de rótulos de motorização (0 – Não motorizado; 1 – Motorizado) e recorrência (R – local recorrente; N – local não-recorrente), considerando-se todas as movimentações registradas no painel. A Figura 12 exemplifica a codificação de uma cadeia completa para um estudante, durante o período de três dias.

Figura 12: Exemplo de formação da cadeia de viagens.

No dia 1, em seu primeiro deslocamento, o indivíduo desloca-se de automóvel (1 – meio motorizado) de sua residência para a universidade (R - local recorrente). Em seu segundo deslocamento, o indivíduo desloca-se caminhando (0 – meio não-motorizado) da universidade para um restaurante fora do campus (N - local não-recorrente). Em seu terceiro deslocamento, o indivíduo desloca-se caminhando (0 – meio não-motorizado) do restaurante de volta para a universidade (R - local recorrente). Em seu quarto deslocamento, o indivíduo desloca-se de automóvel (1 – meio motorizado) da universidade para um cinema (N - local não-recorrente). Em seu quinto deslocamento, o indivíduo desloca-se de automóvel (1 – meio motorizado) do cinema de volta para sua residência (R - local recorrente). O conjunto de cinco pares compõem a cadeia de viagens do primeiro dia $\rightarrow [1;R]-[0;N]-[0;R]-[1;N]-[1;R]$.

No dia 2, o primeiro deslocamento é feito de automóvel (1) da residência para a universidade (R). Lá o aluno passa o dia, retornando (segundo deslocamento) de automóvel (1) para sua residência. Os dois pares compõem a cadeia de viagens do segundo dia $\rightarrow [1;R]-[1;R]$.

No dia 3, há um primeiro deslocamento feito de automóvel (1) da residência para a universidade (R). Após há um segundo deslocamento feito a pé (0) para um mercado (N); a seguir há um terceiro deslocamento feito a pé (0) do mercado para a universidade (R) sucedido do quarto deslocamento realizado de automóvel (1) da universidade para residência (R). O conjunto de quatro pares compõem a cadeia de viagens do terceiro dia $\rightarrow [1;R]-[0;N]-[0;R]-[1;R]$.

A concatenação das cadeias dos três dias resulta na cadeia total de viagens do indivíduo avaliado $\rightarrow [1;R]-[0;N]-[0;R]-[1;N]-[1;R]-[1;R]-[1;R]-[1;R]-[0;N]-[0;R]-[1;R]$.

Acrescenta-se ainda, ao final da estrutura total, informações relativas ao número total de viagens realizadas e distância média de deslocamentos. Para mensuração do comportamento individual relativo a viagens, para posterior classificação, são utilizadas as seguintes informações:

- (a) A porcentagem de motorização (razão entre a contagem do número de deslocamentos motorizados e a contagem total dos pares de rótulos);
- (b) A porcentagem de recorrência (razão entre a contagem do número de deslocamentos para locais recorrentes e a contagem total dos pares de rótulos);
- (c) O número total de viagens realizadas (contagem total dos pares de rótulos);
- (d) A distância média de deslocamentos (média das distâncias de todos os deslocamentos realizados - extraída do painel de dados).

A Figura 13 ilustra o acréscimo das quatro informações quantitativas aplicadas ao exemplo da cadeia total da Figura 12.

Figura 13: Acréscimo de informações quantitativas à cadeia de viagens.

Assim, observando-se a Figura 13, tem-se as seguintes informações quantitativas relativas a cadeia de viagens total de três dias, para cada indivíduo.

- Porcentagem de motorização (0,63) advinda de sete deslocamentos motorizados (1º, 4º e 5º no primeiro dia; 1º e 2º no segundo dia e 1º e 4º no terceiro dia) em um total de onze;
- Porcentagem de recorrência (0,72) obtida de três deslocamentos para locais recorrentes (1º, 3º e 5º no primeiro dia; 1º e 2º no segundo dia e 1º, 3º e 4º no terceiro dia) em um total de onze;
- Número de deslocamentos (11) proveniente da simples contagem do número de pares de viagem;
- Distância média dos deslocamentos, calculada a partir da média dos valores registrados no painel para o indivíduo.

A Tabela 2 sumariza as principais características dos dados obtidos no período de três dias, convencionadas às quatro variáveis quantitativas. Nota-se que há quem se locomova integralmente apenas por modos não-motorizados, entretanto não há quem se locomova somente por modos motorizados. Para a variável “Grau de recorrência”, não há quem somente vá a locais corriqueiros, mas também não há quem faça deslocamentos apenas para locais não frequentes. “Número de viagens” e “distância média de deslocamentos” têm grande variação entre valor máximo e mínimo. Com relação as variáveis socioeconômicas, a “faixa etária” dos 18 aos 22 anos é predominante e o “sexo” tem distribuição quase uniforme. Com relação ao “grau de escolaridade”, a vasta maioria dos alunos tem o ensino médio completo, o que significa que estão matriculados no curso de graduação em engenharia civil. A pequena taxa de indivíduos com ensino superior completo são os alunos do curso de pós-graduação. A grande maioria dos indivíduos não possui “renda” nem “exerce atividade remunerada”. Relativo à “moradia”, há um número razoável de alunos que residem em residência estudantil, mas a maioria dos alunos residem em casas ou apartamentos. Quanto à “presença de automóveis na residência”, pouco mais da metade dos alunos não tem acesso a este veículo sendo que a parcela complementar tem acesso a pelo menos um automóvel. Finalmente, relativo à “presença de motocicletas na residência”, é pequena a parcela de alunos que tem acesso a esse veículo.

Tabela 2: Sumarização das características dos dados coletados por *Smartphones*

Município	São Carlos - SP			
Extrato populacional	Estudantes universitários - Campus 1 Universidade de São Paulo			
Registros	104			
Variáveis	Mínimo	Máximo	Média	
Grau de motorização	0	0,83	0,42	
Grau de recorrência	0,2	0,92	0,67	
Número de viagens	4	20	11,68	
Distância média dos deslocamentos (m)	307	1099	585,19	
Faixas etárias	18-22 anos	23-27 anos	28+ anos	
	78%	19%	3%	
Sexo	Masculino		Feminino	
	56%		44%	
Escolaridade	Médio completo		Superior completo	
	95%		5%	
Renda	Sem renda	0-2 Salários Mínimos	2,1-4 Salários Mínimos	
	91%	6%	3%	
Moradia	Casa/Apartamento		Residência Estudantil	
	82%		18%	
Atividade remunerada	Sim		Não	
	9%		91%	
Número de automóveis disponíveis no domicílio	0	1	2	3
	54%	27%	6%	13%
Número de motocicletas disponíveis no domicílio	0		1	
	90%		10%	

4.2.3 Algoritmo de agrupamento k -médias

O estabelecimento das cadeias de viagens é de suma relevância para a etapa de agrupamento e classificação individual segundo o comportamento, pois cada uma das quatro informações quantitativas, provenientes das cadeias, servirá como parâmetro de entrada para o algoritmo k -médias (MACQUEEN et al., 1967). Tal algoritmo objetiva classificar em k grupos um conjunto de n observáveis, definidas em termos de seus respectivos vetores de características (COSTA; CESAR, 2009). Segundo MacQueen et al. (1967), o objetivo principal do algoritmo é definir k centroides, um para cada um dos k grupos, segundo as etapas:

- (a) Os centroides inicialmente são distribuídos randomicamente pelo espaço delimitado pelo conjunto das n observáveis (Figura 14 a).
- (b) O próximo passo é pegar cada um dos n pontos, pertencentes ao conjunto de observáveis, e associá-los ao centroide mais próximo. Quando nenhum ponto está pendente, o primeiro passo é concluído e um agrupamento inicial é feito (Figura 14 b).
- (c) Neste ponto, precisa-se recalcular k novos centroides como baricentros dos *clusters* resultantes da etapa anterior (Figura 14 c).
- (d) Depois de obtidos esses k novos centroides, uma nova ligação deve ser feita entre os mesmos pontos de conjunto de dados e o novo centroide mais próximo. Os passos (b), (c) e (d) devem ser repetidos enquanto os k centroides mudarem sua localização quando comparados aos centroides da iteração anterior. O algoritmo se encerra quando os centroides não mais se moverem (Figura 14 d).

Figura 14: Etapas do algoritmo k -médias.

Neste trabalho, os 104 registros do painel geraram as $n = 104$ cadeias totais de viagens, que são definidas pelo vetor de quatro características: porcentagem de motorização, porcentagem de recorrência, número de viagens e distância média das viagens.

“Porcentagem de motorização” e “porcentagem de recorrência” têm valores que atingem no máximo uma unidade. Já as duas outras características, “número de viagens” e “distância média de viagens”, têm valores máximos que extrapolam uma unidade (Tabela 2). Por se tratar de uma análise multidimensional, considerou-se que as características deveriam ter o mesmo peso analítico, caso contrário a magnitude dos valores mais altos da variável “distância média de viagens” poderia impor uma tendência equivocada no processo de segmentação dos grupos. Portanto, decidiu-se por normalizar os dados de “número de viagens” e de “distância média de viagens” de forma que todos os valores do vetor de características estivessem dentro de uma extensão que varia de 0 a 1.

O algoritmo k -médias para esse estudo contou com $k = 3$ grupos distintos: Grupo 1, Grupo 2 e Grupo 3. As análises posteriores caracterizarão cada um dos grupos fazendo com que sejam nomeados de acordo com seus atributos.

4.2.4 Análise espacial exploratória

Após a classificação individual obtida na etapa anterior, têm-se a modelagem paramétrica para investigação do comportamento (grupos de indivíduos). A etapa de modelagem, que determinará a análise da vizinhança no comportamento, inicia-se a partir da análise espacial exploratória.

A análise espacial exploratória, deste trabalho, consistiu no cálculo de dois indicadores de associação espacial global (Moran e SivarG) para a variável de interesse (Grupos: 3 categorias). Para isso, demandava-se uma divisão dos dados em faixas de atuação, inicialmente para a análise espacial exploratória e, posteriormente, para a extração das covariáveis espaciais.

As dimensões do espaço amostral da base de dados principal são de 4,5 km no eixo X e 4,5 km no eixo Y. Assim, a maior distância possível nesse espaço é a diagonal principal que vale aproximadamente 6,4 km. Entretanto, a análise variográfica omnidirecional apontou um alcance (a) de 11,3 km atrelado a *cut-distance* de 2,6 km (Figura 15). A partir do alcance (a) é considerado que não há mais grande variância entre os dados (aleatoriedade espacial), sendo que a *cut-distance* impõe um limite superior para cálculos, sejam eles de probabilidade, de variância, de correlação, dentre outros. Constituiu-se quatro faixas de vizinhança através da divisão igualitária da dimensão da *cut-distance* em quatro parcelas, considerando duas faixas antes do alcance (a) e duas faixas após o alcance (a): (I) 0-650 metros; (II) 650-1300 metros; (III) 1300-1950 metros; (IV) 1950-2600 metros. A constituição das faixas também é outro conceito herdado da geoestatística e é denominado *lag-distance*. Refere-se aos segmentos, delimitados pela *cut-distance*, usados para auxiliar na aplicação dos cálculos (probabilidade, variância, correlação) por pontos. Todos as *lag-distances* comumente tem o mesmo tamanho e são equidistantes.

Figura 15: Variograma omnidirecional para a base de dados principal.

Após o cálculo de ambos os indicadores, para as faixas de distâncias determinadas, deverão ser avaliados não apenas os valores dos indicadores, mas também a significância estatística para autocorrelação espacial. Assim, as vizinhanças consideradas para determinação das covariáveis espaciais serão aquelas consideradas estatisticamente significativas ($P \leq 0,05$) para os testes de hipóteses associados aos indicadores globais calculados. A hipótese nula, associada aos testes de hipótese para ambos os indicadores de associação espacial, é de aleatoriedade espacial.

4.2.5 Proposta de extração de covariáveis espaciais

Considera-se um conjunto de observáveis distribuídos em um espaço. Segundo Tobler (1970), todos os elementos desse conjunto estão relacionados, entretanto elementos mais próximos são mais similares entre si, enquanto elementos mais distantes são menos similares.

Baseado nos preceitos geoestatísticos, que consistem da caracterização estatística de variáveis regionalizadas (YAMAMOTO; LANDIM, 2015), se estabeleceu um método a partir da mensuração das probabilidades que um indivíduo tem de pertencer a uma determinada categoria (variável analisada) segundo os valores de seus vizinhos (caráter de regionalidade). O intuito da abordagem é o de complementar a análise paramétrica tradicional através da inclusão de covariáveis regionais, pois segundo o trabalho de Naizer

(2018), o acréscimo dessas informações pode ser benéfico para modelagens de escolha discreta se for atestado o caráter de associação espacial da amostra.

As covariáveis espaciais deste trabalho são as probabilidades dos vizinhos (determinados a partir das quatro vizinhanças) pertencerem a cada uma das três categorias de padrões comportamentais estabelecidas pelo algoritmo k -médias.

Para o cálculo das probabilidades, o primeiro passo consiste do estabelecimento de vizinhanças de análise. Novamente, sugere-se aferir a distância máxima para cálculos (*cut-distance*) e dividir em porções iguais essa distância a fim de se determinar n vizinhanças de interesse (faixas de *lag-distance*), sendo que as mesmas não devem se sobrepor. Padronizou-se a utilização das mesmas faixas previamente estabelecidas para aplicação dos indicadores.

Estabelecidas as vizinhanças, efetua-se o cálculo, por indivíduos, das probabilidades de pertencimento às categorias, dadas as categorias dos seus vizinhos. Para cada indivíduo e para cada vizinhança, conta-se a quantidade de elementos compreendidos na região que pertençam a cada categoria existente no estudo. A probabilidade desse indivíduo pertencer a cada uma das categorias é a razão entre o número de elementos de uma dada categoria e o número total de elementos, encontrados na vizinhança considerada. A Figura 16 apresenta um exemplo dos procedimentos citados, num espaço simplificado e com apenas duas categorias possíveis.

Figura 16: Exemplo de cálculo de probabilidades, por categorias, para o ponto X, segundo quatro vizinhanças de interesse.

Caso deseje-se caracterizar um indivíduo localizado no ponto X segundo as probabilidades de pertencimento a uma determinada categoria baseadas em vizinhanças, procede-se:

- (a) Dadas as dimensões máximas da distribuição de dados, divide-se o espaço em um número n de vizinhanças de interesse. Para o exemplo, os dados estão dispostos em um quarto de circunferência com raio de tamanho $4L$. Assim, é possível delimitar $n = 4$ faixas de tamanho L cada. Cada faixa será, portanto, uma vizinhança de análise.
- (b) Conta-se o número de elementos de todas as categorias presentes na primeira vizinhança. Para o exemplo, na primeira faixa há 4 elementos vermelhos da categoria A e nenhum elemento preto da categoria B. Logo, um indivíduo localizado em X tem, segundo a primeira vizinhança ($n = 1$):
 - $P1(A) = 1$ (4 elementos de um total de 4) de ser da categoria A; e
 - $P1(B) = 0$ (0 elementos de um total de 4) de ser da categoria B.
- (c) Atua-se de maneira análoga para as demais vizinhanças ($n = 2, 3$ e 4). Assim, o indivíduo X é caracterizado:
 - $P2(A) = 0,7$ (7 elementos de um total de 10) de ser da categoria A;
 - $P2(B) = 0,3$ (3 elementos de um total de 10) de ser da categoria B;
 - $P3(A) = 0,5$ (8 elementos de um total de 16) de ser da categoria A;
 - $P3(B) = 0,5$ (8 elementos de um total de 16) de ser da categoria B;
 - $P4(A) = 0,2$ (4 elementos de um total de 20) de ser da categoria A; e
 - $P4(B) = 0,8$ (16 elementos de um total de 20) de ser da categoria B.
- (d) Repetem-se os procedimentos anteriores para todos os pontos da amostra.

Desta forma, comprovada a dependência espacial na etapa de análise espacial exploratória e calculados os indicadores de associação espacial global, as covariáveis espaciais são propostas, baseadas nas vizinhanças consideradas estatisticamente significativas na análise exploratória e nas probabilidades de os vizinhos pertencerem às escolhas em análise.

4.2.6 Calibração de funções utilidade apenas com variáveis originais

Esta etapa corresponde à calibração do modelo *logit* não espacial. A determinação das funções utilidade, nesta etapa, é baseada nas variáveis que caracterizam os indivíduos: três faixas etárias (*i1* dos 18-22 anos, *i2* dos 23-27 anos e *i3* para mais de 28 anos), sexo (masculino ou feminino), escolaridade (ensino médio completo e ensino superior completo), renda (sem renda, 0-2 salários mínimos, 2,1-4 salários mínimos), tipo de moradia (residência em moradia estudantil ou residência em casa/apartamento), desempenho de atividade remunerada (sim ou não), quantidade de automóveis disponíveis na residência (0,1,2 ou 3 automóveis), quantidade de motocicletas disponíveis na residência (0 ou 1 motocicleta). A variável resposta são as três categorias de padrões comportamentais (Grupos 1, 2 e 3). As Equações 4.1, 4.2 e 4.3, na sua forma literal, representam as funções utilidade do modelo não espacial e as funções utilidade (*V1*, *V2* e *V3*) correspondem aos três grupos comportamentais, provenientes do algoritmo k-médias.

$$V1 = 0, \quad (4.1)$$

$$\begin{aligned} V2 = & CTE2 + \beta_{2_idade1} * i1 + \beta_{2_idade2} * i2 + \beta_{2_sexo} * sexo \\ & + \beta_{2_escol} * escolaridade + \beta_{2_renda} * renda \\ & + \beta_{2_moradia} * moradia + \beta_{2_ativ_remun} * ativ_remun \\ & + \beta_{2_auto} * auto + \beta_{2_moto} * moto, \end{aligned} \quad (4.2)$$

$$\begin{aligned} V3 = & CTE3 + \beta_{3_idade1} * i1 + \beta_{3_idade2} * i2 + \beta_{3_sexo} * sexo \\ & + \beta_{3_escol} * escolaridade + \beta_{3_renda} * renda \\ & + \beta_{3_moradia} * moradia + \beta_{3_ativ_remun} * ativ_remun \\ & + \beta_{3_auto} * auto + \beta_{3_moto} * moto, \end{aligned} \quad (4.3)$$

- β_{k_idadeJ} - Coeficiente estimado para categoria J de idade do indivíduo do grupo k,
- β_{k_sexo} - Coeficiente estimado para sexo do indivíduo do grupo k,
- β_{k_escol} - Coeficiente estimado para escolaridade do indivíduo do grupo k,
- β_{k_renda} - Coeficiente estimado para renda do indivíduo do grupo k,
- $\beta_{k_moradia}$ - Coeficiente estimado para tipo de moradia do indivíduo do grupo k,
- $\beta_{k_ativ_remun}$ - Coeficiente estimado para o desempenho de atividade remunerada pelo indivíduo do grupo k,
- β_{k_auto} - Coeficiente estimado para quantidade de automóveis disponíveis ao indivíduo do grupo k,
- β_{k_moto} - Coeficiente estimado para quantidade de motocicletas disponíveis ao indivíduo do grupo k.

4.2.7 Calibração de funções utilidade com variáveis originais e covariáveis espaciais propostas

Uma vez que a análise espacial exploratória apontou que existe autocorrelação espacial entre as variáveis, espera-se que a inclusão de informações espaciais aumente o potencial de acerto e qualidade da modelagem paramétrica. Aplica-se, então, o cálculo das probabilidades de pertencimento às categorias, baseado nos valores de vizinhanças, consideradas significativas na análise espacial exploratória.

O estabelecimento das probabilidades, por categoria, em cada uma das vizinhanças provê o acréscimo informacional de caráter regionalizado que se buscava para incrementar as análises paramétricas de escolha discreta (*logit* multinomial). Sem contar as constantes e a primeira função utilidade que definiu-se como nula por conveniência de cálculos, antes se levava em conta nove atributos em cada função utilidade: faixa etária *i1*, faixa etária *i2*, sexo, escolaridade, renda, tipo de moradia, desempenho de atividade remunerada, quantidade de automóveis disponíveis na residência, quantidade de motocicletas disponíveis na residência.

Agora, além destes, conta-se com novos doze atributos dos indivíduos (para serem aplicados cada qual na sua respectiva função utilidade): $PI(1)$, $PI(2)$, $PI(3)$, $PII(1)$, $PII(2)$, $PII(3)$, $PIII(1)$, $PIII(2)$, $PIII(3)$, $PIV(1)$, $PIV(2)$, $PIV(3)$ que são respectivamente a probabilidade P de um indivíduo pertencer ao grupo de padrão comportamental 1 a partir dos valores dos elementos na primeira vizinhança I , probabilidade P de um indivíduo pertencer ao grupo de padrão comportamental 2 a partir dos valores dos elementos na primeira vizinhança I , probabilidade P de um indivíduo pertencer ao grupo de padrão comportamental 3 a partir dos valores dos elementos na primeira vizinhança I , e assim por diante, alternando as probabilidade P de pertencimento aos grupos comportamentais até a vizinhança IV .

Ressalva-se que a utilização de todas as vizinhanças é um cenário ideal onde todas foram apontadas como significativas pelos indicadores. A existência das covariáveis está, portanto, condicionado ao resultado aferido na análise espacial exploratória. Deve-se incluir na equação paramétrica apenas as covariáveis espaciais relativas às regiões onde ao menos um indicador global apontou significância ($P \leq 0,05$) para autocorrelação espacial. Caso contrário a região (vizinhança) deve ser desconsiderada.

O acréscimo informacional locacional define agora novas funções utilidade, representadas pelas Equações 4.4, 4.5 e 4.6:

$$V1 = \beta_{1_vI} * PI(1) + \beta_{1_vII} * PII(1) + \beta_{1_vIII} * PIII(1) + \beta_{1_vIV} * PIV(1), \quad (4.4)$$

$$\begin{aligned} V2 = & CTE2 + \beta_{2_idade1} * i1 + \beta_{2_idade2} * i2 + \beta_{2_sexo} * sexo \\ & + \beta_{2_escol} * escolaridade + \beta_{2_renda} * renda \\ & + \beta_{2_moradia} * moradia + \beta_{2_ativ_remun} * ativ_remun \\ & + \beta_{2_auto} * auto + \beta_{2_moto} * moto \\ & + \beta_{2_vI} * PI(2) + \beta_{2_vII} * PII(2) + \beta_{2_vIII} * PIII(2) + \beta_{2_vIV} * PIV(2), \end{aligned} \quad (4.5)$$

$$\begin{aligned} V3 = & CTE3 + \beta_{3_idade1} * i1 + \beta_{3_idade2} * i2 + \beta_{3_sexo} * sexo \\ & + \beta_{3_escol} * escolaridade + \beta_{3_renda} * renda \\ & + \beta_{3_moradia} * moradia + \beta_{3_ativ_remun} * ativ_remun \\ & + \beta_{3_auto} * auto + \beta_{3_moto} * moto \\ & + \beta_{3_vI} * PI(3) + \beta_{3_vII} * PII(3) + \beta_{3_vIII} * PIII(3) + \beta_{3_vIV} * PIV(3), \end{aligned} \quad (4.6)$$

- β_{k_idadeJ} - Coeficiente estimado para categoria J de idade do indivíduo do grupo k,
- β_{k_sexo} - Coeficiente estimado para sexo do indivíduo do grupo k,
- β_{k_escol} - Coeficiente estimado para escolaridade do indivíduo do grupo k,
- β_{k_renda} - Coeficiente estimado para renda do indivíduo do grupo k,
- $\beta_{k_moradia}$ - Coeficiente estimado para tipo de moradia do indivíduo do grupo k,
- $\beta_{k_ativ_remun}$ - Coeficiente estimado para o desempenho de atividade remunerada pelo indivíduo do grupo k,
- β_{k_auto} - Coeficiente estimado para quantidade de automóveis disponíveis ao indivíduo do grupo k,
- β_{k_moto} - Coeficiente estimado para quantidade de motocicletas disponíveis ao indivíduo do grupo k,
- β_{k_vi} - Coeficiente estimado para probabilidade de pertencimento do indivíduo do grupo k na vizinhança i.

4.2.8 Análise comparativa entre modelos de escolha discreta não espacial e espacial

Para a comparação de modelos *logit* (espaciais e não espaciais) foram utilizados quatro métricas: Os valores de rho-quadrado, rho-quadrado ajustado, valor do critério Akaike e taxa de acertos por validação cruzada e logaritmo da verossimilhança para amostra de validação.

A métrica rho-quadrado é definida pela equação:

$$\rho^2 = 1 - \frac{L^*}{L_0}, \quad (4.7)$$

onde L_0 é o valor de verossimilhança obtida ao assumir todos os parâmetros β do modelo como zero e L^* é o valor de máximo-verossimilhança obtida quando os parâmetros β correspondem aos valores estimados. Assim, um modelo ideal tende a unidade pois a razão L^* (caso onde os parâmetros β tem seus valores ótimos) por L_0 (caso onde os parâmetros β são todos nulos), tende a zero por L^* ser muito menor que L_0 .

A métrica rho-quadrado ajustado é definida pela equação:

$$\rho^2_* = 1 - \frac{L^* - K}{L_0}, \quad (4.8)$$

L_0 e L^* são análogos a Equação 4.7. K é o número de parâmetros estimados. Assim, a interpretação da métrica ajustada é análoga à convencional, levando-se em consideração o desconto do número K de parâmetros do valor L^* .

O critério Akaike é definido por:

$$A = 2K - 2 \ln L^*, \quad (4.9)$$

Os valores de K e L^* são análogos às Equações 4.7 e 4.8. Estabelecida por meio de uma subtração entre o número K de parâmetros e o logaritmo do valor L^* de máximo-verossimilhança, a formulação Akaike faz com que o critério penalize o *overfitting* (ato de acréscimo demasiado de variáveis as equações com o intuito de obter melhores ajustes, carecendo de critérios para tal adição) e é por essa razão que se busca menores valores para esse critério.

Finalmente, a validação cruzada consiste em segregar uma parcela da amostra para ser usada no processo de estimação de parâmetros e uma outra parcela usada no processo de validação dos parâmetros estimados. Após, aplica-se a modelagem junto dos parâmetros calibrados advindos do grupo de calibração, nos elementos do grupo de validação, obtendo-se assim valores estimados para esse grupo.

Como os valores do grupo de validação são conhecidos, pode-se aferir a taxa de acertos ao se comparar os valores reais com os valores estimados. Bons modelos tendem a ter taxas de acerto altas e a qualidade deste valor pode ser mensurada pelo cálculo de verossimilhança:

$$L = p^y \cdot (1 - p)^{(n-y)}, \quad (4.10)$$

Para o número total n de elementos considerados no teste, tem-se uma quantidade y de elementos avaliados corretamente. Define-se assim, a razão $p = \frac{y}{n}$ entre o número de elementos avaliados corretamente e o total de elementos considerados. Quando o valor de p tende a unidade (100% de taxas de acertos), o valor de L também tende a unidade, e consequentemente, $\log L$ tende ao valor nulo.

4.3 Aplicativos e ferramentas

A elaboração deste trabalho de doutorado fez uso de uma série de aplicativos e ferramentas. Uma breve descrição de cada ferramenta é listada a seguir:

4.3.1 Aplicativo para telefones móveis *Google Maps*

A aplicação *Google Maps* para telefones móveis foi de suma relevância para o desenvolvimento desse projeto. A função “*Linha do tempo*” desse aplicativo de mapas, com a opção de “*histórico de localizações*” ativada (Figura 17), foi responsável pela etapa de coleta de dados em sua forma bruta. É a aplicação de mapas nativa dos sistemas operativos *Android* e uma das mais usuais na plataforma *iOS*.

Figura 17: Aplicativo *Google Maps* com as configurações necessárias para coleta de dados.

A aplicação, quando configurada da maneira descrita, torna-se uma potente ferramenta de coleta de dados passiva. O usuário não precisa se lembrar de invocar a aplicação e nem mesmo há a necessidade da tela estar ligada. A coleta de dados é constante e de baixíssimo consumo de bateria.

4.3.2 Ferramentas de desenvolvimento *web* (*WebDev*)

Uma vez coletados os dados de movimentação pela aplicação, estes precisavam ser enviados para serem processados. Para isso, foi desenvolvido um servidor para armazenamento dos dados, interfaceado por meio de um *website*. Neste portal, além da submissão dos dados de movimentação, o indivíduo também preenche um abreviado questionário socioeconômico. O servidor que hospeda o *website* foi configurado com o sistema operativo Linux e o armazenamento de dados é feito por meio de bancos de dados programados em *mySQL*. O interfaceamento foi estabelecido por meio das linguagens web HTML, CSS, PHP e Javascript. A Figura 18 apresenta a interface programada para coleta de dados.

EESC • USP
Planejamento Urbano
São Carlos / SP

INSTRUÇÕES ENVIAR DADOS

Colabore com a pesquisa

Os dados coletados ajudam no planejamento de ações como:

- ampliação das linhas de transporte público,

Informações socioeconômicas

Idade

Sexo

Grau de Instrução

Atividade

Renda

Num de Automóveis no Domicílio

Num de Motocicletas no Domicílio

ENVIAR

Envio de dados

Envie o arquivo .json exportado do seu histórico de localização:

Selecionar Arquivo nenhum arquivo selecionado

ENVIAR

Figura 18: Interface *web* para submissão dos dados coletados.

O sistema possui certificação SSL (padrão consagrado de segurança *web*) o que protege o anonimato do usuário no envio de dados.

4.3.3 Linguagem de programação *Python*

Python é uma linguagem de programação multi-plataforma e destaca-se pela simplicidade de seus códigos e pela enorme quantidade de módulos adicionais desenvolvidos para as mais diversas necessidades. Nesse projeto, a linguagem de programação é utilizada nos algoritmos responsáveis pela realização das seguintes etapas:

- Pré-processamento dos dados brutos;
- Estabelecimento das variáveis derivadas dos dados originais para composição do painel:
 - Tempo e distância de deslocamentos;
 - Estimativa da velocidade média dos deslocamentos;
 - Análise de recorrência das localidades;
 - Caráter de motorização individual.
- Construção das cadeias de viagens;
- Agrupamento por meio da técnica k-médias;
- Análise espacial exploratória (em associação com algoritmos desenvolvidos por Naizer (2018) em linguagem de programação R);
- Extração de covariáveis espaciais;

- Análise espacial confirmatória (em associação com o módulo da aplicação *Biogeme* desenvolvido para *Python*).

4.3.4 *Biogeme*

A ferramenta *Biogeme* (2016) foi desenvolvido por Michel Bierlaire na *Ecole Polytechnique Fédérale de Lausanne* - Suíça. Trata-se de uma aplicação utilizada na estimação da máxima verossimilhança de modelos paramétricos em geral, com ênfase especial em modelos de escolha discreta. Possui um módulo denominado *pythonBiogeme* que permite a realização de modelos paramétricos de escolha discreta integrados às variáveis já presentes no ambiente Python. Foi utilizado na etapa de análise espacial confirmatória desse projeto.

4.3.5 Alocação de ferramentas por etapas

A Tabela 3 sumariza a utilização das ferramentas nas etapas desse projeto:

Tabela 3: Sumarização da utilização de ferramentas durante as etapas do projeto

Etapas - Ferramentas	Google Maps	WebDev	Python	Biogeme
Coleta de dados	X	X		
Pre-processamento de dados			X	
Composição do painel			X	
Construção das cadeias de viagens			X	
Agrupamento k-médias			X	
Análise espacial exploratória			X	
Extração de covariáveis espaciais			X	
Análise espacial confirmatória			X	X

5 RESULTADOS DO PROCESSO DE CLASSIFICAÇÃO E AGRUPAMENTO

Este capítulo descreve os resultados acerca do protótipo que validou o método de coleta de dados, atestando que a coleta via dispositivos móveis é mais precisa que os métodos tradicionais de entrevista. Em seguida, o capítulo apresenta os resultados do processo de classificação individual segundo comportamento relativo a viagens. Foram utilizados quatro diferentes tipos de informações quantificadas, referentes aos deslocamentos realizados em três dias consecutivos, para formar grupos de indivíduos e determinar a variável dependente, utilizada posteriormente na modelagem paramétrica.

5.1 Validação do método de coleta de dados

O estudo inicia-se com a coleta de dados através dos *smartphones*, advindos dos 104 estudantes que, voluntariamente, participaram da pesquisa. Esses dados chegam em sua forma bruta e precisam ser pré-processados. O pré-processamento consistiu do método de determinação das localidades de atividades por aglutinação, baseado em tempos e distâncias. A efetividade dessa abordagem foi comprovada por meio de uma análise comparativa com um método tradicional de obtenção de dados, baseado em entrevistas e composição de um diário de atividades.

Foi indagado, para um conjunto de dez indivíduos, algumas perguntas referentes aos seus deslocamentos realizados no dia anterior (quantidade de viagens realizadas, distância e tempo dispendidos em cada deslocamento e o modo de transporte utilizado). A seguir, para os mesmos indivíduos e o mesmo dia de referência, comparou-se os dados extraídos pela entrevista com os dados pré-processados computacionalmente. Os resultados são apresentados nas Tabelas 4 e 5:

Tabela 4: Comparativo entre dados declarados e computados - Amostra piloto (parte1).

ID	Quantidade total de viagens declaradas	Quantidade total de viagens computadas	Soma total de distâncias declaradas (Km)	Soma total de distâncias computadas (Km)	Soma total de tempos declarados (min.)	Soma total de tempos computados (min.)
1	7	7	8	9,3	65	60
2	5	5	6,5	6,2	24	28
3	6	8	5	5,7	26	33
4	7	9	10,5	7,1	70	86
5	10	12	9	13,8	95	81
6	5	5	7	4,6	30	34
7	7	7	10	7,3	30	35
8	6	8	7	10,2	75	81
9	5	5	12	7,6	65	86
10	4	4	5	4,7	10	12
TOTAL	62	70	80	76,5	490	536

Tabela 5: Comparativo entre dados declarados e computados - Amostra piloto (parte2).

ID	Quantidade total de viagens motorizadas declaradas	Quantidade total de viagens motorizadas computadas	Quantidade total de viagens não-motorizadas declaradas	Quantidade total de viagens não-motorizadas computadas
1	5	5	2	2
2	5	5	0	0
3	6	6	0	2
4	5	5	2	4
5	8	10	2	2
6	5	5	0	0
7	7	7	0	0
8	4	6	2	2
9	0	0	5	5
10	4	4	0	0
TOTAL	49	53	13	17

Nota-se que há disparidades entres os valores declarados em entrevista e os valores computados computacionalmente. Quanto ao total de viagens, nota-se que 4 dos 10 entrevistados declararam um valor diferente do computado pelo algoritmo computacional, totalizando uma diferença de 4 viagens. Ao se apresentar os valores computados aos entrevistados, averiguou-se que a essa diferença se deve ao esquecimento de determinados deslocamentos realizados, sendo que dois deles são referentes a deslocamentos não-motorizados e dois a deslocamentos motorizados.

Com relação ao somatório das distâncias percorridas e tempos dispendidos há diferenças, mesmo que mínimas, entre os valores declarados e computados de todos os indivíduos. Atribui-se essa diferença ao fato de que tais respostas são baseadas no juízo dos entrevistados, e a noção de tempo e distância, que é diferente para cada pessoa, nunca será tão precisa quanto a de um aparelho que registra intermitentemente locomoções com o auxílio de um GPS, dotado de enorme precisão locacional.

O percentual da diferença absoluta entre valores declarados e computados frente ao universo para cada variável é apresentado na Figura 19:

Figura 19: Percentual da diferença absoluta entre valores declarados e computados frente ao universo para cada variável.

As taxas de diferença entre valores declarados e computados variam de 4,3% até 23,52%, sendo que a maior taxa ocorre nas viagens não-motorizadas.

Após, foram calculados testes não paramétricos para amostras independentes (Mann-Whitney e Mediana) a fim de averiguar se os valores declarados e calculados eram similares (hipótese nula) segundo suas distribuições populacionais e medidas típicas. Os valores são apresentados na Tabela 6.

Tabela 6: Resultados dos testes não paramétricos para amostras independentes.

Variáveis	Mann-Whitney		Mediana	
	Significância	H0	Significância	H0
Quantidade total de viagens	0,656	Retém	0,481	Retém
Soma total de distâncias	1	Retém	0,684	Retém
Soma total de tempos	1	Retém	0,481	Retém
Quantidade total de viagens motorizadas	1	Retém	0,684	Retém
Quantidade total de viagens não-motorizadas	1	Retém	0,631	Retém

H0 teste U de Mann-Whitney: Os valores declarados e computados são similares para distribuições

H0 teste da mediana: Os valores declarados e computados são similares para medidas típicas

Em ambos os testes e para todas as variáveis em questão, corroborou-se a hipótese de similaridade entre dados coletados e computados.

Em suma, os resultados atestam que a abordagem de coleta de dados por *Smartphones* é, segundo os testes paramétricos, similar ao método tradicional de entrevistas e composição de um diário de atividades. Entretanto, a nova abordagem traz consigo vantagens para a pesquisa tais como: aumento da precisão relativo aos tempos de deslocamento e as distâncias percorridas, além da coleta ser realizada de maneira mais prática.

A melhora com relação aos aspectos citados se dá pelo fato de que a coleta via dispositivos móveis minimiza erros de esquecimento ou omissão das viagens e reduz, significativamente, a imprecisão das informações de distâncias e tempos de deslocamentos frutos de percepções equivocadas dos entrevistados. Vale lembrar que, apesar de colaborar com tais benefícios, um painel constituído por dados de *Smartphones* não consegue identificar com exatidão a motivação para o deslocamento nem o modo utilizado.

Com relação à motivação para o deslocamento, por meio de dados em painel, pode-se fazer excelentes inferências se os deslocamentos eram realizados para locais recorrentes ou não-recorrentes (inclusive com identificação de locais de moradia e trabalho/estudo através da análise de maiores tempos de permanência), mas o exato motivo de viagem ainda é melhor identificado em entrevistas presenciais.

De maneira análoga, através de dados em painel, conta-se com boa inferência com relação ao uso de modos motorizados e não-motorizados, mediante análise das velocidades de deslocamento. Entretanto, a exata escolha modal do indivíduo é melhor identificada na abordagem tradicional de entrevistas.

5.2 Resultados do processo de agrupamento

Comprovada a precisão dos dados pré-processados, deu-se continuidade na composição do painel, seguido do estabelecimento e quantificação das cadeias de viagens e chegando, finalmente, no algoritmo de agrupamento.

O particionamento do algoritmo k -médias resultou em $k = 3$ grupos distintos, os quais são exibidos (Figura 20) em diagramas de Voronoi, plotados em plano bidimensional para cada uma das características consideradas.

Figura 20: Particionamento estabelecido pelo algoritmo k -médias.

A partir da Figura 20, são explicitados três grupos de padrões de viagens para a amostra estudada. Diagramas de Voronoi expressam, duas-a-duas, as variáveis usadas como vetor de características usadas pelo algoritmo de agrupamento. A diagonal principal da imagem expressa a distribuição por meio de histogramas das variáveis envolvidas.

O algoritmo de agrupamento, para o presente trabalho, deve ser considerado efetivo pois segmentou, com sucesso, as amostras em três categorias de padrões de deslocamentos. Tal afirmação sustenta-se uma vez que cada uma das células de Voronoi (região delimitada por elementos pertencentes a um mesmo grupo) exibidas na Figura 20, está bem demarcada, havendo pouca ou nenhuma superposição de elementos (fator que induz classificações equivocadas) em todos os diagramas analisados.

Quanto aos grupos obtidos, convencionou-se sua nomenclatura como:

- Grupo 1 (verde) – Estudantes universitários com baixo grau de motorização, pouca recorrência de localidades, poucas viagens e distâncias curtas: *Estudantes não motorizados com motivos diversos de viagens, predominantemente curtas.*
- Grupo 2 (vermelho) – Estudantes universitários com grau de motorização moderada, recorrência moderada, número alto de viagens e distâncias médias: *Estudantes com opções motorizadas, com viagens recorrentes e também motivos variados e distâncias médias.*
- Grupo 3 (azul) – Estudantes universitários com alto grau de motorização, alta recorrência das viagens, poucas viagens e distâncias longas: *Estudantes motorizados com viagens, longas, às atividades obrigatórias (recorrentes) predominantemente.*

As Tabelas 7, 8 e 9 apresentam a mesma sumarização de dados vista na Tabela 2, porém agora, aplicada aos grupos estipulados.

Tabela 7: Sumarização das características do Grupo 1

Município	São Carlos - SP			
Grupo 1	Estudantes não motorizados com motivos diversos de viagens, predominantemente curtas			
Elementos	18			
Variáveis	Mínimo	Máximo	Média	
Grau de motorização	0	0,26	0,07	
Grau de recorrência	0,2	0,53	0,38	
Número de viagens	4	11	7,5	
Distância média dos deslocamentos (m)	307	536	408,66	
Faixas etárias	18-22 anos	23-27 anos	28+ anos	
	72%	28%	0%	
Sexo	Masculino		Feminino	
	61%		39%	
Escolaridade	Médio completo		Superior completo	
	94%		6%	
Renda	Sem renda	0-2 Salários Mínimos	2,1-4 Salários Mínimos	
	89%	11%	0%	
Moradia	Casa/Apartamento		Residência Estudantil	
	78%		22%	
Atividade remunerada	Sim		Não	
	11%		89%	
Número de automóveis disponíveis no domicílio	0	1	2	3
	83%	11%	0%	6%
Número de motocicletas disponíveis no domicílio	0		1	
	94%		6%	

Tabela 8: Sumarização das características do Grupo 2

Município	São Carlos - SP			
Grupo 2	Estudantes com opções motorizadas, com viagens recorrentes e também motivos variados e distâncias médias			
Elementos	70			
Variáveis	Mínimo	Máximo	Média	
Grau de motorização	0,2	0,64	0,44	
Grau de recorrência	0,53	0,92	0,7	
Número de viagens	7	20	13,97	
Distância média dos deslocamentos (m)	510	687	596,54	
Faixas etárias	18-22 anos	23-27 anos	28+ anos	
	84%	11%	5%	
Sexo	Masculino		Feminino	
	53%		47%	
Escolaridade	Médio completo		Superior completo	
	97%		3%	
Renda	Sem renda	0-2 Salários Mínimos	2,1-4 Salários Mínimos	
	99%	1%	0%	
Moradia	Casa/Apartamento		Residência Estudantil	
	81%		19%	
Atividade remunerada	Sim		Não	
	1%		99%	
Número de automóveis disponíveis no domicílio	0	1	2	3
	56%	24%	4%	16%
Número de motocicletas disponíveis no domicílio	0		1	
	91%		9%	

Tabela 9: Sumarização das características do Grupo 3

Município	São Carlos - SP			
Grupo 3	Estudantes motorizados com viagens, longas, às atividades obrigatórias (recorrentes) predominantemente			
Elementos	16			
Variáveis	Mínimo	Máximo	Média	
Grau de motorização	0,53	0,83	0,73	
Grau de recorrência	0,73	0,91	0,83	
Número de viagens	4	9	6,37	
Distância média dos deslocamentos (m)	599	1099	734,12	
Faixas etárias	18-22 anos	23-27 anos	28+ anos	
	56%	44%	0%	
Sexo	Masculino		Feminino	
	63%		37%	
Escolaridade	Médio completo		Superior completo	
	88%		12%	
Renda	Sem renda	0-2 Salários Mínimos	2,1-4 Salários Mínimos	
	62%	19%	19%	
Moradia	Casa/Apartamento		Residência Estudantil	
	88%		12%	
Atividade remunerada	Sim		Não	
	38%		62%	
Número de automóveis disponíveis no domicílio	0	1	2	3
	12%	56%	19%	13%
Número de motocicletas disponíveis no domicílio	0		1	
	81%		19%	

Observando-se as Tabelas 7, 8 e 9, verifica-se que valores mínimos globais são todos advindos do grupo de baixo grau de motorização (1). O Grupo com grau de motorização moderada (2) colabora com os valores máximos globais das variáveis “Grau de recorrência” e “Número de viagens”. O grupo com alto grau de motorização (3) colabora com os valores máximos globais das variáveis “Grau de motorização” e “Distância média dos deslocamentos”.

A convenção de nomenclatura estipulada é reforçada pela estatística referente ao número de automóveis disponíveis no domicílio: apenas 17% dos indivíduos do grupo de baixa motorização têm acesso a pelo menos um automóvel em sua residência. Esse número aumenta para 44% no grupo de moderado grau de motorização até culminar no valor de 88% para o grupo de alto grau de motorização.

Estabelecida a nomenclatura dos grupos, procede-se as seguintes análises referentes aos diagramas da Figura 20:

- (i) Figura 20 [A][4] - Grau de motorização vs. distância média de viagens: Há uma correlação positiva entre o grau de motorização e a distância média das viagens, uma vez que a medida que o grau de motorização aumenta, a distância média de viagens também aumenta. Tais comportamentos são triviais de se entender, uma vez que a posse de um veículo motorizado privilegia deslocamentos mais extensos. Ressalta-se também a distinção dos grupos neste gráfico, onde o grupo de baixo grau de motorização está concentrado no canto inferior esquerdo e gradativamente vai aumentando ambas as características, passando pelo grupo de motorização moderada, até culminar no grupo com alto grau de motorização, plotado no canto superior direito.
- (ii) Figura 20 [A][3] – Grau de motorização vs. número de viagens: Através deste diagrama é possível perceber que o maior número de viagens está associado ao grupo com grau de motorização moderada. Indivíduos deste conjunto combinam dois tipos de modos de transporte ao longo do dia, modos motorizados e não-motorizados. Segundo a Tabela 8, 19% dos elementos desse grupo residem em moradia estudantil, aumentando as probabilidades destes indivíduos terem acesso a pelo menos um automóvel, experimentando viagens feitas por meio de carona. Isto é, rotineiramente o modo principal de locomoção deste grupo é não motorizado, entretanto, sempre que possível valem-se de viagens motorizadas apesar de muitos não possuírem veículos.

Os indivíduos do grupo de baixa motorização, por não terem acesso a motocicletas ou automóveis (83% dos indivíduos deste grupo não têm acesso a automóveis em suas residências e 94% não têm acesso a motocicletas) geralmente realizam um número reduzido de viagens.

Os indivíduos do grupo de alta motorização, também realizam um número reduzido de viagens, visto que seus deslocamentos são, em média, os mais longos (análise do item i). Como há uma estreita relação entre maiores distâncias percorridas e um aumento do tempo de viagem, estes acabam por realizar menos viagens a fim de não dispendem muito tempo nas mesmas. O elevado grau de motorização desse grupo explica-se pelo fato deste concentrar a maior taxa de alunos de pós-graduação (13% possui ensino superior completo) e também a maior taxa de indivíduos que exercem atividade remunerada (38%).

- (iii) Figura 20 [A][2] – Grau de motorização vs. grau de recorrência: Assim como a análise do item i, estas são duas grandezas que têm correlação positiva entre si, sendo que os elementos pertencentes ao grupo de baixa motorização são os mais propensos a um comportamento menos recorrente, uma vez que estes estão mais aptos a realizarem viagens de oportunidade, uma vez que a não motorização privilegia grande variação de caminhos realizados e locais visitados. Ambas as variáveis aumentam gradativamente até chegarem no comportamento do grupo com alto grau de motorização que normalmente restringem-se a realizar deslocamentos para desempenhar suas atividades regulares, já que, possivelmente, residem mais distantes das localidades recorrentes.
- (iv) Figura 20 [B][4] – Grau de recorrência vs. distância média: Outro gráfico que apresenta correlação positiva entre as variáveis. Trata-se de um complemento informacional das análises (i) e (iii) e evidencia o fato de que viagens para locais não recorrentes normalmente têm uma menor distância média, enquanto viagens para locais recorrentes estão associados aos maiores deslocamentos. Tal resultado corrobora a relação, encontrada na literatura (AGUILERA; MASSOT; PROULHAC, 2009; ANDERSON, 2010; WILLIAMS, 2017), entre maiores distâncias de viagens para atividades de trabalho ou estudo e menores distâncias para atividades esporádicas.

Tal fato é corroborado pelas análises (i) e (iii). Segundo (i), deslocamentos para localidades não recorrentes são realizados por meios não-motorizados, meios estes que, segundo a análise (iii), contemplam menores distâncias, expressos no diagrama [B][4] apenas por indivíduos pertencentes ao grupo de baixa motorização.

Já o comportamento relativo aos locais recorrentes com distância média elevada, associados ao modo motorizado, são expressos no diagrama [B][4] pelos elementos dos conjuntos de moderado e alto grau de motorização.

- (v) Figura 20 [B][3] – Grau de recorrência vs. número de viagens: Este diagrama esclarece que ambos os extremos, viajantes que majoritariamente deslocam-se para locais não recorrentes e viajantes que majoritariamente deslocam-se para locais recorrentes, estão associados a um menor número de viagens realizadas por dia, sendo contemplados

pelos grupos de baixo grau de motorização e alto grau de motorização, respectivamente. O comportamento intermediário que condiz a um misto entre viagens com propósitos recorrentes e não-recorrentes, concentra o maior número de deslocamentos no período analisado, e manifesta-se no gráfico [B][3] pelos elementos do grupo de motorização moderada.

- (vi) Figura 20 [C][4] – Número de viagens vs. distância média: Diagrama complementar às informações das análises (ii) e (v), explicita que um baixo número de viagens está associado a curtas distâncias ou longas distâncias, enquanto que valores mais altos de número de viagens correlacionam-se a distâncias médias percorridas.

Tem cunho complementar pois, segundo a análise (ii), a ausência da motorização (grupo de baixa motorização) ou a presença com uso predominante (grupo de alta motorização) está ligado a um menor número de viagens, que de acordo com (v), tem íntima relação com os extremos de baixo e alto grau de recorrência. Já o grupo que combina os modos motorizado e não-motorizado realiza um número maior de viagens em suas jornadas diárias, sendo que tais deslocamentos não são nem tão curtos como o dos indivíduos do grupo de baixa motorização e nem tão longos como o dos indivíduos do grupo de alta motorização.

- (vii) Figura 20 [A][1] – Histograma da variável grau de motorização: A distribuição do histograma evidencia que há uma transição entre o grau de motorização e os elementos de cada grupo. As caixas mais a esquerda são completamente compostas por elementos do grupo de baixa motorização. A seguir, vê-se que as taxas centrais são compostas predominantemente por indivíduos do grupo de motorização moderada, havendo alguma superposição da baixa motorização nas caixas à esquerda e da alta motorização à direita. Finalmente, as caixas mais a direita são completamente compostas por elementos do grupo de alta motorização.

- (viii) Figura 20 [B][2] – Histograma da variável grau de recorrência: As taxas de baixa recorrência são dominadas pelos indivíduos com baixo grau de motorização. À medida que a taxa chega a um valor médio, começa a haver grande predominância dos indivíduos com moderada taxa de motorização. Tal predominância estende-se até os mais elevados graus de recorrência, quando a composição é mista entre elementos de moderado e alto grau de motorização.

- (ix) Figura 20 [C][3] – Histograma da variável número de viagens: Os menores valores de viagens realizadas são compostos por elementos dos grupos de baixo e alto grau de motorização. Para os valores médios de número de viagens há uma superposição entre os três grupos. Os valores altos para número de viagens são predominantemente realizados pelo grupo de motorização moderada.

- (x) Figura 20 [D][4] – Histograma da variável distância média de viagens: Outro histograma que apresenta transição entre o grau de motorização e os elementos de cada grupo. Caixas mais a esquerda, predominantemente compostas por indivíduos do grupo de baixa motorização, apontam os menores deslocamentos. Caixas mais a esquerda, completamente compostas por indivíduos do grupo de alta motorização, apontam os maiores deslocamentos. As caixas mais altas, que indicam maior predominância no conjunto de dados apontam que a maior parte dos deslocamentos são de cunho médio, vastamente realizados por elementos do grupo de motorização moderada, apresentando alguma intersecção com os outros grupos.

Além das análises comportamentais descritas, uma vez que se conta com a localidade da residência de cada um dos registros, é possível construir um mapa (Figura 21) onde observa-se a distribuição geográfica dos três grupos obtidos. Por essa perspectiva é possível corroborar muitas das análises originárias dos diagramas de Voronoi.

Figura 21: Mapa construído a partir das coordenadas da residência dos indivíduos estudados.

O conjunto de dados analisado compete estudantes universitários, portanto é natural que o seu local de estudo seja uma localidade recorrente e um ponto agregador de viagens. É exatamente o que é expresso pela circunferência alaranjada, que representa o campus da universidade envolta por diversos marcadores representando os muitos estudantes avaliados.

Os elementos do grupo de baixa motorização, representados por marcadores verdes, em sua maioria residem nas proximidades do campus pois não estão aptos a realização de longos deslocamentos.

Já os elementos representados pelos marcadores vermelhos, têm comportamento dual, visto que muitos elementos moram nas proximidades do campus e tantos outros residem em localidades mais afastadas, o que está de acordo com a categoria de motorização moderada deste agrupamento, onde os indivíduos alternam entre deslocamentos motorizados e não-motorizados, havendo igual alternância entre a localidade da residência afastada ou próxima da universidade.

Finalmente, os marcadores em azul, em grande maioria, distam bastante do campus, estando em conformidade com a categoria “*alto grau de motorização*” por eles representada, que uma vez dotados do acesso aos modos motorizados, tornam-se aptos às maiores movimentações.

Desta forma, o mapa representa, de forma visual, autocorrelação espacial da variável categórica “*grupos*”. Tal variável é a variável dependente a ser utilizada, posteriormente, nos modelos *logit* (binomial para a cidade fictícia e multinomial para amostra de estudantes) espaciais e não espaciais.

6 RESULTADOS DA MODELAGEM PARAMÉTRICA

Uma vez que se obteve os três grupos de comportamentos, na etapa anterior da pesquisa, este capítulo versa sobre os resultados acerca da modelagem paramétrica, responsável pela investigação da influência exercida, pela vizinhança, no comportamento individual relativo a viagens. Entretanto, antes de se chegar aos resultados para a base de dados principal da pesquisa, testes foram realizados em uma base de dados auxiliar (compostas de dados espacialmente dependentes), de maneira a averiguar a eficiência da análise espacial incorporada de covariáveis espaciais, prevendo assim as diretrizes para a aplicação na base de dados principal.

6.1 Pré-requisitos da modelagem na base de dados auxiliar

Antes da realização da análise espacial exploratória e confirmatória para a base de dados principal, decidiu-se por estabelecer testes a fim de validar as etapas metodológicas que compreendem a modelagem paramétrica. Assim, inicialmente trabalhou-se com uma base de dados fictícia, no sentido de avaliar um cenário conhecido e controlado, pois, pautando-se no trabalho de Naizer (2018), já se tinha ciência da dependência espacial desses dados.

A utilização da base de dados sintética emula as etapas metodológicas de coleta de dados, construção de cadeias de viagens e agrupamento por algoritmo k -médias. Isso porque a base de dados já conta com uma variável que é discreta e segrega os dados em dois grupos: o binário de escolha modal (0) ônibus e (1) automóvel.

Com relação à análise espacial exploratória, necessitou-se estabelecer as vizinhanças para a aplicação dos indicadores e extração das covariáveis espaciais. Como as dimensões do espaço amostral são de 10 unidades no eixo x e 5 unidades no eixo y , sabe-se que a maior distância possível nesse espaço é a diagonal principal que vale aproximadamente 12 unidades. Esse valor foi adotado como limitante *cut-distance* e, a partir dele, constituiu-se quatro faixas de vizinhança através da divisão igualitária da diagonal em quatro parcelas: (I) 0 – 3 unidades; (II) 3,1 – 6 unidades; (III) 6,1 – 9 unidades; (IV) 9,1 – 12 unidades.

Com relação à análise confirmatória, a calibração do modelo *logit* binomial não espacial levou em conta, na composição das funções utilidade, as variáveis que representam as alternativas modais (tempo do ônibus, tempo do automóvel, preço do ônibus e preço do automóvel), da viagem (distância de deslocamento), bem como variáveis individuais (renda). A variável resposta é o binário de escolha modal (0) ônibus e (1) automóvel. As Equações 6.1 e 6.2, na sua forma literal, representam as funções utilidade do modelo não espacial.

$$V_0 = \beta_{0_dist} * dist + \beta_{0_renda} * renda + \beta_{tempo} * tempo_onibus + \beta_{preco} * preco_onibus, \quad (6.1)$$

$$V_1 = CTE + \beta_{1_dist} * dist + \beta_{1_renda} * renda + \beta_{tempo} * tempo_auto + \beta_{preco} * preco_auto, \quad (6.2)$$

- β_{k_dist} - Coeficiente estimado para distância percorrida do indivíduo usuário do modo k ,
 β_{k_renda} - Coeficiente estimado para renda do indivíduo usuário do modo k ,
 β_{tempo} - Coeficiente estimado para tempo despendido pelo indivíduo,
 β_{preco} - Coeficiente estimado para o valor despendido pelo indivíduo.

Já a calibração do modelo *logit* espacial, além dos atributos originais das equações 6.1 e 6.2, idealmente incorpora novos oito atributos dos indivíduos: $PI(0)$, $PI(1)$, $PII(0)$, $PII(1)$, $PIII(0)$, $PIII(1)$, $PIV(0)$, $PIV(1)$, que são respectivamente a probabilidade P de um indivíduo escolher realizar deslocamentos por ônibus (0) a partir dos valores dos elementos na primeira vizinhança I , probabilidade P de um indivíduo escolher realizar deslocamentos por automóvel (1) a partir dos valores dos elementos na primeira vizinhança I , e assim por diante, alternando as probabilidade P de escolha modal até a vizinhança IV . Probabilidades estas que foram advindas do método de extração das covariáveis espaciais.

O acréscimo informacional locacional define agora novas funções utilidade, representadas pelas Equações 6.3 e 6.4:

$$V_0 = \beta_{0_dist} * dist + \beta_{0_renda} * renda + \beta_{tempo} * tempo_onibus + \beta_{preco} * preco_onibus + \beta_{0_vI} * PI(0) + \beta_{0_vII} * PII(0) + \beta_{0_vIII} * PIII(0) + \beta_{0_vIV} * PIV(0) \quad (6.3)$$

$$V_1 = CTE + \beta_{1_dist} * dist + \beta_{1_renda} * renda + \beta_{tempo} * tempo_auto + \beta_{preco} * preco_auto + \beta_{1_vI} * PI(1) + \beta_{1_vII} * PII(1) + \beta_{1_vIII} * PIII(1) + \beta_{1_vIV} * PIV(1) \quad (6.4)$$

- β_{k_dist} - Coeficiente estimado para distância percorrida do indivíduo usuário do modo k ,
 β_{k_renda} - Coeficiente estimado para renda do indivíduo usuário do modo k ,
 β_{tempo} - Coeficiente estimado para tempo despendido pelo indivíduo,
 β_{preco} - Coeficiente estimado para o valor despendido pelo indivíduo,
 β_{k_vi} - Coeficiente estimado para probabilidade de escolha do indivíduo pelo modo k na vizinhança i .

É relevante frisar que é dito que, idealmente, se incorpora todos esses componentes espaciais pois parte-se da premissa de que todos os parâmetros estimados são estatisticamente significativos. Se esse não for o caso, os coeficientes não-significativos devem ser desconsiderados. Além disso, são incorporados apenas as covariáveis espaciais pertencentes às vizinhanças consideradas espacialmente dependentes pelos testes de hipótese realizados na análise espacial exploratória.

6.2 Resultados da modelagem paramétrica para a cidade fictícia

Inicialmente, foi realizada a análise espacial exploratória, com base no cálculo dos indicadores Moran e SivarG para a variável escolha modal, apresentados na Tabela 10. Vale ressaltar aqui que, as vizinhanças escolhidas, para a modelagem posterior, foram aquelas que os indicadores foram considerados estatisticamente significativos pelos testes de hipótese, associados a pelo menos um dos indicadores. A hipótese nula, para os testes de hipótese de ambos os indicadores, é de aleatoriedade espacial. Neste caso, as vizinhanças de interesse são aquelas onde a hipótese nula foi rejeitada em pelo menos um dos indicadores (Vizinhanças I, II e III).

Tabela 10: Resultado dos indicadores Moran e SivarG aplicados ao banco de dados sintético.

Vizinhança	Moran				SivarG			
	Índice	Z	p-valor	H0	Índice	Z	p-valor	H0
I	0,269	65,119	0	Rejeita	0,954	-7,867	0	Rejeita
II	-0,064	-43,602	0	Rejeita	1	0	0,6	Aceita
III	-0,01	-30,338	0	Rejeita	1	0	0,6	Aceita
IV	-0,002	0	1	Aceita	1	0	0,6	Aceita

Verifica-se, a partir da Tabela 10, que os indicadores são grandezas quase que complementares. A proximidade do valor 1, para o índice de Moran, indica alta autocorrelação espacial, enquanto que o valor 1 indica muito baixa autocorrelação espacial para o indicador SivarG.

Nota-se que os valores dos indicadores, estatisticamente significativos, não são altos (para Moran) ou baixos (para SivarG). Isso indica que, mesmo que pequena, existe alguma autocorrelação espacial nessas localidades e é estatisticamente significativa. Assim, a inclusão de covariáveis, extraídas dessas vizinhanças, proverá um acréscimo informacional aos modelos de escolha discreta.

Posteriormente, foram calculadas as probabilidades dos vizinhos, que pertencem às vizinhanças *I*, *II* e *III* escolherem o modo ônibus (categoria 0) ou o modo automóvel (categoria 1). Reforça-se que essas covariáveis espaciais são individuais. A cada indivíduo, são associados seis valores de probabilidades dos seus vizinhos: $P(1)I$; $P(1)II$; $P(1)III$; $P(0)I$; $P(0)II$; $P(0)III$.

A próxima etapa, relativa à calibração das funções utilidades das variáveis originais, baseou-se nas Equações 6.1 e 6.2. Ao realizar a calibração pelo método da máxima verossimilhança, foram obtidos os parâmetros β estimados. Entretanto, nem todos foram significativos. Sendo assim, foram retiradas, das funções utilidade, aquelas variáveis, associadas aos parâmetros não significativos (β_{renda} em ambas as Equações 6.1 e 6.2). Assim sendo, faz-se necessário recalibrar o modelo, desconsiderando as variáveis que não são pertinentes ao estudo, conforme Equações 6.5 e 6.6.

$$V_0 = \beta_{0_dist} * dist + \beta_{tempo} * tempo_onibus + \beta_{preco} * preco_onibus, \quad (6.5)$$

$$V_1 = CTE + \beta_{1_dist} * dist + \beta_{tempo} * tempo_auto + \beta_{preco} * preco_auto, \quad (6.6)$$

- β_{k_dist} - Coeficiente estimado para distância percorrida do indivíduo usuário do modo k,
 β_{tempo} - Coeficiente estimado para tempo despendido pelo indivíduo,
 β_{preco} - Coeficiente estimado para o valor despendido pelo indivíduo.

Os resultados são apresentados na Tabela 11. Ao se segregar a amostra em uma proporção 7/3 e promover testes de validação cruzada com os coeficientes calibrados obteve-se uma taxa de acertos de 74% para a amostra de validação, com valor de verossimilhança $L = 2,26 \times 10^{-44}$ e $\log(L) = -100,49$.

Tabela 11: Calibração pelo método da máxima verossimilhança considerando as variáveis originais significativas da base de dados sintética

	Coeficiente	Valor	Erro	p-valor
	β_{0_dist}	-9.46	1.07	0.03
	β_{1_dist}	9.46	1.37	0.05
	β_{preco}	-20.0	0.844	0.00
	β_{tempo}	20.8	1.33	0.00
	CTE	-7.87	1.05	0.00
Rho-quadrado:		0.756		
Rho-quadrado ajustado:		0.738		
Critério de informação Akaike:		210.455		

Em seguida, é realizada a calibração do modelo espacial com a incorporação das covariáveis espaciais relativas às vizinhanças I, II e III, por indivíduo. Como os testes de hipótese não apontaram significância para a região IV, as variáveis relativas a essa região e seus respectivos parâmetros foram retirados das Equações 6.3 e 6.4. Novamente, como na primeira modelagem, após a calibração, houve um coeficiente não-significativo (coeficiente

relativo à terceira vizinhança na Equação 6.4) demandando a eliminação de sua variável da equação e uma nova calibração.

Desconsiderando os coeficientes estimados não significativos e suas respectivas variáveis associadas, promove-se a nova calibração de acordo com as Equações 6.7 e 6.8.

$$V_0 = \beta_{0_dist} * dist + \beta_{tempo} * tempo_onibus + \beta_{preco} * preco_onibus + \beta_{0_vI} * PI(0) + \beta_{0_vII} * PII(0) + \beta_{0_vIII} * PIII(0), \quad (6.7)$$

$$V_1 = CTE + \beta_{1_dist} * dist + \beta_{tempo} * tempo_auto + \beta_{preco} * preco_auto + \beta_{1_vI} * PI(1) + \beta_{1_vII} * PII(1), \quad (6.8)$$

- β_{k_dist} - Coeficiente estimado para distância percorrida do indivíduo usuário do modo k ,
 β_{tempo} - Coeficiente estimado para tempo despendido pelo indivíduo,
 β_{preco} - Coeficiente estimado para o valor despendido pelo indivíduo,
 β_{k_vi} - Coeficiente estimado para probabilidade de escolha do indivíduo pelo modo k na vizinhança i .

Os resultados são expressos na Tabela 12. Novamente, segregando a amostra em uma proporção 7/3 para testes de validação cruzada com os coeficientes calibrados obteve-se uma taxa de acertos de 89% para a segunda modelagem, com valor de verossimilhança $L = 1,82 \times 10^{-39}$ e $\log(L) = -79,98$.

Tabela 12: Calibração pelo método da máxima verossimilhança considerando as covariáveis espaciais acrescidas significativas, para base de dados sintética.

	Coeficiente	Valor	Erro	p-valor
	β_{0_dist}	-19.9	1.80	0.04
	β_{0_vI}	13.0	1.41	0.01
	β_{0_vII}	9.1	1.60	0.02
	β_{0_vIII}	8.66	2.71	0.02
	β_{1_dist}	19.9	1.35	0.05
	β_{1_vI}	14.1	5.91	0.03
	β_{1_vII}	7.66	4.67	0.03
	β_{preco}	-47.1	3.2	0.00
	β_{tempo}	56.1	4.1	0.00
	CTE	1.41	1.66	0.00
Rho-quadrado:		0.929		
Rho-quadrado ajustado:		0.904		
Critério de informação Akaike:		77.323		

Ao se comparar a primeira modelagem, que considerava as variáveis significativas originais da base de dados (Tabela 11), com a segunda modelagem, que considerava as

variáveis significativas originais e as covariáveis regionais (Tabela 12), corrobora-se a hipótese de que, se averiguada a existência de associação espacial das variáveis por meio de análise espacial exploratória, pode-se incluir o aspecto espacial ao modelo pois este trará maior precisão aos resultados. A Figura 22 apresenta os valores das métricas comparativas para os modelos não-espacial e espacial:

Figura 22: Valores das métricas comparativas para os modelos não-espacial e espacial, usando-se a base de dados auxiliar.

O acréscimo informacional é positivo proporcionando, para este estudo de caso, um aumento de 15% das taxas de acertos da validação cruzada (com correspondente aumento de verossimilhança). Os valores que aferem a qualidade do modelo também foram positivos: rho-quadrado e rho-quadrado ajustado aumentaram e se aproximaram consideravelmente da unidade. Já o critério Akaike reduziu-se de 210,455 para 77,323. Salienta-se a importância de tal redução uma vez que o critério Akaike penaliza *overfitting* e é por essa razão que menores valores são almejados para esse critério. O segundo modelo mostra-se mais desejável pois apesar de contar com cinco coeficientes a mais que o primeiro, apresenta critério Akaike aproximadamente 2,72 vezes menor.

Os bons resultados obtidos para essa base de dados sintética encorajaram a aplicação do procedimento de análise espacial na base de dados original, visando a obtenção de modelos mais precisos.

6.3 Resultados da modelagem paramétrica para a base de dados principal

Novamente, foi realizada a análise espacial exploratória, aplicada agora à base de dados principal do estudo (Tabela 13). Mais uma vez, as vizinhanças escolhidas foram aquelas que os indicadores foram considerados estatisticamente significativos pelos testes de hipótese, associados a pelo menos um dos indicadores. As vizinhanças de interesse são aquelas onde a hipótese nula (aleatoriedade espacial) foi rejeitada em pelo menos um dos indicadores (Vizinhanças I, II e III).

Tabela 13: Resultado dos indicadores Moran e SivarG aplicados ao banco de dados principal.

Vizinhança	Moran				SivarG			
	Índice	Z	p-valor	H0	Índice	Z	p-valor	H0
I	0,131	3,424	0,01	Rejeita	0,928	-4,616	0	Rejeita
II	0,091	4,008	0	Rejeita	0,985	-0,927	0,01	Rejeita
III	0,026	2,058	0,04	Rejeita	0,998	-0,121	0,45	Aceita
IV	-0,02	-0,333	0,73	Aceita	1	0	0,50	Aceita

Verifica-se, a partir da Tabela 13, que mais uma vez obteve-se o cenário onde os valores para os indicadores estatisticamente significativos, não são altos (para Moran) ou baixos (para SivarG) dando indícios da existência estatisticamente significativa de alguma autocorrelação espacial nessas localidades. A inclusão de covariáveis, extraídas dessas vizinhanças, é recomendada pois entende-se que proverá um acréscimo informacional aos modelos *logit* para a base de dados em painel.

Após, foram calculadas as probabilidades dos vizinhos, localizados nas vizinhanças I, II e II (aquelas que foram estatisticamente significativas na análise espacial exploratória) pertencerem a cada um dos três grupos de padrões comportamentais (Grupos 1, 2 e 3). Sendo as covariáveis espaciais individuais, a cada indivíduo, são incorporados nove valores de probabilidades dos seus vizinhos: $P(1)I$; $P(1)II$; $P(1)III$; $P(2)I$; $P(2)II$; $P(2)III$; $P(3)I$; $P(3)II$; $P(3)III$.

Na sequência, a calibração das funções utilidades das variáveis originais baseou-se nas Equações 4.1, 4.2 e 4.3 (Capítulo 4). Ao realizar a calibração pelo método da máxima verossimilhança, foram obtidos os parâmetros β estimados. Entretanto, nem todos foram significativos, sendo assim, foram retiradas das funções utilidade, aquelas variáveis, associadas aos parâmetros não significativos. Para a Equação 4.2 foram desconsiderados os parâmetros: β_{2_idade1} , β_{2_idade2} , β_{2_sexo} , β_{2_escol} , β_{2_renda} , $\beta_{2_ativ_remun}$. Para a Equação 4.3 foram desconsiderados os parâmetros: β_{3_idade1} , β_{3_idade2} , β_{3_sexo} , β_{3_escol} , $\beta_{3_moradia}$. Assim sendo, faz-se necessário recalibrar o modelo, desconsiderando as variáveis que não são pertinentes ao estudo, conforme Equações 6.9, 6.10 e 6.11.

$$V1 = 0, \quad (6.9)$$

$$\begin{aligned} V2 = & CTE2 \\ & + \beta_{2_moradia} * moradia \\ & + \beta_{2_auto} * auto + \beta_{2_moto} * moto \end{aligned} \quad (6.10)$$

$$\begin{aligned} V3 = & CTE3 \\ & + \beta_{3_renda} * renda + \beta_{3_ativ_remun} * ativ_remun \\ & + \beta_{3_auto} * auto + \beta_{3_moto} * moto \end{aligned} \quad (6.11)$$

- β_{k_renda} - Coeficiente estimado para renda do indivíduo do grupo k,
 $\beta_{k_moradia}$ - Coeficiente estimado para tipo de moradia do indivíduo do grupo k,
 $\beta_{k_ativ_remun}$ - Coeficiente estimado para o desempenho de atividade remunerada pelo indivíduo do grupo k,
 β_{k_auto} - Coeficiente estimado para quantidade de automóveis disponíveis ao indivíduo do grupo k,
 β_{k_moto} - Coeficiente estimado para quantidade de motocicletas disponíveis ao indivíduo do grupo k.

Os resultados são apresentados na Tabela 14. Ao se segregar a amostra em uma proporção 7/3 e promover testes de validação cruzada com os coeficientes calibrados obteve-se uma taxa de acertos de 71%, com valor de verossimilhança $L = 3,76 \times 10^{-8}$ e $\log(L) = -17,09$.

Tabela 14: Calibração pelo método da máxima verossimilhança considerando as variáveis originais significativas da base de dados principal

	Coeficiente	Valor	Erro	p-valor
	β_{2_auto}	-23.6	0.976	0.00
	β_{3_auto}	0.449	0.249	0.03
	β_{2_moto}	-13.2	0.943	0.00
	β_{3_moto}	0.367	0.13	0.01
	$\beta_{2_moradia}$	36.4	1.54	0.00
	$\beta_{3_ativ_remun}$	0.05	0.024	0.00
	β_{3_renda}	0.739	0.25	0.03
	CTE2	-5.10	2.37	0.03
	CTE3	-2.02	0.355	0.00
Rho-quadrado:		0.417		
Rho-quadrado ajustado:		0.338		
Critério de informação Akaike:		151.199		

Em seguida, é realizada a calibração do modelo espacial, com a incorporação das covariáveis espaciais relativas às vizinhanças I, II e III, por indivíduo, uma vez que os testes de hipótese não apontaram significância para a região IV. Sendo assim, as variáveis relativas a essa região e seus respectivos parâmetros (β_{vIV}) foram retirados das Equações 4.4, 4.5 e 4.6 (Capítulo 4).

Novamente, como na primeira modelagem, após a calibração, houve coeficientes não-significativo demandando a eliminação dos mesmos e de suas variáveis associadas da equação, além do procedimento de uma nova calibração. Para a Equação 4.4 foram desconsiderados os parâmetros: β_{1_vII} , β_{1_vIII} . Para a Equação 4.5 foi desconsiderado os parâmetro: β_{2_vIII} . Para a Equação 4.6 foram desconsiderados os parâmetros: β_{3_vII} , β_{3_vIII} .

Desconsiderando os coeficientes não significantes e suas respectivas variáveis associadas, promoveu-se a nova calibração de acordo com as Equações 6.12, 6.13 e 6.14:

$$V1 = \beta_{1_vI} * PI(1), \quad (6.12)$$

$$\begin{aligned} V2 = & CTE2 \\ & + \beta_{2_moradia} * m_moradia \\ & + \beta_{2_auto} * auto + \beta_{2_moto} * moto \\ & + \beta_{2_vI} * PI(2) + \beta_{2_vII} * PII(2) \end{aligned} \quad (6.13)$$

$$\begin{aligned} V3 = & CTE3 \\ & + \beta_{3_renda} * renda + \beta_{3_ativ_remun} * ativ_remun \\ & + \beta_{3_auto} * auto + \beta_{3_moto} * moto \\ & + \beta_{3_vI} * PI(3) \end{aligned} \quad (6.14)$$

- β_{k_renda} - Coeficiente estimado para renda do indivíduo do grupo k,
 $\beta_{k_moradia}$ - Coeficiente estimado para tipo de moradia do indivíduo do grupo k,
 $\beta_{k_ativ_remun}$ - Coeficiente estimado para o desempenho de atividade remunerada pelo indivíduo do grupo k,
 β_{k_auto} - Coeficiente estimado para quantidade de automóveis disponíveis ao indivíduo do grupo k,
 β_{k_moto} - Coeficiente estimado para quantidade de motocicletas disponíveis ao indivíduo do grupo k,
 β_{k_vi} - Coeficiente estimado para probabilidade de pertencimento do indivíduo do grupo k na vizinhança i.

Os resultados são expressos na Tabela 15. Novamente, segregando a amostra em uma proporção 7/3 para testes de validação cruzada, com os coeficientes calibrados, obteve-se uma taxa de acertos de 76% para a segunda modelagem, com valor de verossimilhança $L = 1,09 \times 10^{-7}$ e $\log(L) = -16,02$.

Tabela 15: Calibração pelo método da máxima verossimilhança considerando as covariáveis espaciais acrescidas significativas, para base de dados principal.

Coeficiente	Valor	Erro	p-valor
β_{2_auto}	-25.2	1.32	0.00
β_{3_auto}	0.475	0.257	0.05
β_{2_moto}	-11.3	1.18	0.00
β_{3_moto}	0.373	0.11	0.03
$\beta_{2_moradia}$	38.6	2.32	0.00
$\beta_{3_ativ_remun}$	0.202	0.12	0.03
β_{3_renda}	0.704	0.502	0.04
β_{1_vI}	0.218	0.11	0.03
β_{2_vI}	4.28	2.02	0.04
β_{3_vI}	0.323	0.17	0.03
β_{2_vII}	5.81	2.93	0.05
CTE2	-7.33	3.47	0.03
CTE3	-2.06	0.31	0.01
Rho-quadrado:	0.587		
Rho-quadrado ajustado:	0.391		
Critério de informação Akaike:	139.186		

Ao se comparar a primeira modelagem, que considerava as variáveis significativas originais da base de dados (Tabela 14) com a segunda modelagem, que considerava as variáveis significativas originais e as covariáveis regionais (Tabela 15), corrobora-se, agora para a base de dados principal, a hipótese de que, se averiguada a existência de associação espacial das variáveis por meio de análise espacial exploratória, pode-se incluir o aspecto espacial ao modelo pois este trará maior precisão aos resultados, atestando assim que existe a influência da vizinhança nesse conjunto de dados. A Figura 23 apresenta os valores das métricas comparativas para os modelos não-espacial e espacial.

Figura 23: Valores das métricas comparativas para os modelos não-espacial e espacial - Base de dados principal.

O acréscimo informacional é positivo, proporcionando, para este conjunto de dados, um aumento de 5% das taxas de acertos da validação cruzada. Os valores que aferem a qualidade do modelo também foram positivos: rho-quadrado e rho-quadrado ajustado aumentaram e o critério Akaike reduziu de 151.199 para 139.186. Uma vez que o critério Akaike penaliza *overfitting* e, por isso, almeja-se menores valores para essa métrica, conclui-se que o segundo modelo é mais atrativo pois, apesar de contar com quatro coeficientes a mais que o primeiro, apresenta critério Akaike aproximadamente 1,08 vezes menor.

Com relação aos coeficientes estatisticamente significativos das variáveis originais da base de dados, nota-se que para o grupo 2 (grau de motorização moderada), há coeficientes negativos para as variáveis “*disponibilidade de automóveis*” e “*disponibilidade de motocicletas*” na residência, entretanto há um coeficiente positivo e alto para a variável “*moradia em residência estudantil*”. Esses valores de coeficientes estão de acordo com os resultados do Capítulo 5, que apontavam que esse grupo não dispõe de veículos motorizados (mais da metade dos elementos desse grupo não tem acesso a pelo menos um automóvel em sua residência, tampouco a motocicletas), portanto é natural que a posse de veículos entre como um coeficiente negativo na sua função utilidade. Entretanto, uma parcela considerável de elementos desse grupo (aproximadamente 1/5 dos elementos) tem sua moradia em residências estudantis, o que faz com que o coeficiente para essa variável tenha alto valor na função utilidade. Esse fato embasa as opções de locomoção motorizadas desse grupo, que as fazem experimentando viagens por caronas de seus colegas de residência.

Essas viagens colaboram efetivamente para o caráter misto de locomoção deste grupo.

Já o grupo 3 (grau elevado de motorização) apresenta coeficientes positivos para as variáveis “*disponibilidade de automóveis na residência*”, “*disponibilidade de motocicletas na residência*”, “*desempenho de atividade remunerada*” e “*renda*”. Mais uma vez, os coeficientes estão de acordo com os resultados do Capítulo 5 pois, nesse grupo, mais de 80% dos elementos têm acesso a pelo menos um automóvel em suas residências, além de alguns outros terem acesso a motocicletas, justificando os coeficientes positivos da função utilidade para os veículos. Quanto às variáveis de desempenho de atividade remunerada e renda, os valores positivos dos coeficientes da função utilidade também se justificam pois esse é o grupo com maior número de elementos que trabalham (aproximadamente 38%) e, conseqüentemente, possuem renda não nula.

Sobre os valores obtidos para os novos coeficientes, estatisticamente significativos, nota-se, que para todos os grupos, a primeira vizinhança exerce influência na determinação de um indivíduo segundo seu padrão comportamental, sendo que essa influência se estende até a vizinhança II para o grupo 2. Os coeficientes são todos positivos.

Entende-se que os elementos do grupo 1 (baixa motorização) em sua maioria residem nas proximidades do campus pois não estão aptos a realização de longos deslocamentos. Sendo assim um indivíduo desse grupo sofre influência de sua primeira vizinhança, pois diversos indivíduos de igual categoria 1 estão próximos.

Os elementos do grupo 2 (motorização moderada), caracterizaram-se pelo comportamento dual. Muitos elementos moram nas proximidades do campus e tantos outros residem em localidades mais afastadas pois alternam entre deslocamentos motorizados e não motorizados. Sendo assim, os elementos desse grupo, que realizam a maioria das movimentações por meios não motorizados, sofrem influência de sua primeira vizinhança, de maneira análoga ao que ocorre com os elementos do grupo 1. Entretanto, essa influência estende-se até a segunda vizinhança, contemplando assim aqueles elementos que realizam a maioria das movimentações por meios motorizados.

Finalmente, os elementos do grupo 3 (alto grau de motorização), em grande maioria, distam bastante do campus, pois uma vez dotados do acesso aos modos motorizados, tornam-se aptos às maiores movimentações. Entretanto, estão restritos apenas à influência da primeira vizinhança. Isso se dá pois, próximo de um elemento desse grupo, residem, comumente, muitos outros elementos desse mesmo grupo, uma vez que a longa distância do *campus* demanda intensa movimentação motorizada, inviabilizando tais localidades para a grande maioria de elementos do grupo 1 e boa parte dos elementos do grupo 2.

7 CONSIDERAÇÕES FINAIS

7.1 Síntese do trabalho

A análise do comportamento individual relativo a viagens é de suma relevância para atividades como entendimento da demanda por transportes e embasamento de planejamentos para os sistemas de transportes. Sabe-se que este comportamento individual é influenciado por uma série de fatores do meio urbano, dentre eles a vizinhança. Assim, almeja-se identificar, de maneira satisfatória, a influência que a vizinhança exerce no comportamento dos indivíduos relativo a viagens.

Esse trabalho de doutorado objetivou o desenvolvimento de um procedimento metodológico para a identificação da influência da vizinhança por meio de aquisição de dados em painel, viabilizados por meio da utilização de *Smartphones*.

Em posse do painel, advindo da coleta de dados dos dispositivos móveis, é possível realizar a composição de cadeias de viagens individuais e mensurá-las por meio de quatro variáveis: grau de motorização da cadeia de deslocamentos, grau de recorrência das localidades, número total de deslocamentos realizados e distância média dos deslocamentos realizados.

Definidas e mensuradas as cadeias, uma classificação individual é promovida, através do algoritmo k-médias, de forma a classificar os indivíduos em grupos, derivando assim, a variável dependente para a etapa de modelagem paramétrica.

Para a base de dados de estudantes universitários na cidade de São Carlos, podem-se classificar os indivíduos em três grupos distintos de acordo com o padrão comportamental de deslocamentos dos mesmos a partir do seu encadeamento de viagens.

O grupo de baixo grau de motorização realiza um numero reduzido de viagens e estas são de curta distância, entretanto é o grupo que está mais apto a realizar deslocamentos para locais não recorrentes. Residem nas proximidades do campus da universidade, o que é de grande conveniência, uma vez que o acesso aos modos motorizados é reduzido nesse grupo, devido ao caráter de renda e a baixa taxa de indivíduos que exercem atividade remunerada.

O grupo com grau de motorização moderada é o conjunto que realiza maior número de viagens, visto que combinam os modos motorizados e não motorizados, muito provavelmente por experimentarem viagens realizadas por meios de caronas, possíveis graças aos automóveis presentes nas residências estudantis, moradia de parcela considerável desse grupo. Os deslocamentos são médio/longos e a recorrência varia entre locais recorrentes e não-recorrentes, mas há uma maior tendência aos locais rotineiros. Seu local de residência

varia de acordo com a tendência de uso de modos motorizados.

O grupo com alto grau de motorização também realiza um número reduzido de viagens, entretanto estas são de longas distâncias. Dotados de veículos motorizados, residem em locais afastados da universidade. A posse de veículos se justifica por este ser o grupo com as maiores taxas de nível de escolaridade e renda.

A modelagem paramétrica (*logit* multinomial) é conduzida através de dois modelos. O primeiro modelo, não espacial, contempla apenas as variáveis originais da base de dados. O segundo modelo, espacial, deve incorporar covariáveis regionais.

A incorporação das covariáveis regionais se dá por meio de análise espacial. Num primeiro momento, a análise espacial exploratória, realizada pelos indicadores globais de autocorrelação espacial Moran e SivarG, aponta os locais da base de dados onde existe autocorrelação espacial, dando os primeiros indícios das localizações onde a vizinhança pode exercer influência no comportamento dos indivíduos. Apontadas as regiões dotadas de autocorrelação espacial, é possível, por um método baseado em preceitos geoestatísticos, extrair covariáveis espaciais para incorporarem ao segundo modelo *logit*, que agora além das variáveis originais, dispõem também de covariáveis espaciais. Compostos os dois modelos, parte-se para a etapa da análise espacial confirmatória, onde métricas estatísticas são aplicadas a fim de comparar os dois modelos.

Com relação à análise espacial exploratória, a observação dos testes de hipótese, associados aos indicadores Moran e SivarG, determinou duas vizinhanças a serem acrescentadas à análise. Em seguida, na etapa de análise espacial confirmatória observou-se um aprimoramento na modelagem, quando consideradas as covariáveis espaciais, atestado por quatro métricas comparativas.

Os resultados das modelagens (análise dos valores dos coeficientes estimados) foram compatíveis com os agrupamentos pré-determinados, sendo que nas funções utilidade, os valores dos coeficientes de variáveis que indicavam posse de veículos, foram positivos para o grupo 3 (alto grau de motorização) e negativos para o grupo 2 (motorização moderada), estando de acordo com a escolha modal predominante dos grupos.

Os coeficientes de renda e desempenho de atividade remunerada foram positivos para o grupo 3, o que é coerente pois esse é o grupo com maior número de pessoas formadas e também o grupo com maior número de pessoas que desempenham atividade remunerada.

O coeficiente de local de moradia foi positivo para o grupo 2, o que também é coerente pois sendo os elementos desse grupo pessoas que têm um comportamento que alterna deslocamentos motorizados e não-motorizados, residir em moradias estudantis é um fator positivo para o acesso a veículos, e conseqüentemente, a realização de deslocamentos motorizados.

Além disso, os coeficientes relativos às vizinhanças estavam de acordo com a hipótese

estudada nessa tese, que buscava atestar a influência exercida pela vizinhança nos padrões comportamentais. Todos os grupos sofrem influência de sua primeira vizinhança, e tal influência se estende até a segunda vizinhança no caso do grupo 2.

7.2 Conclusões e contribuições

Este trabalho de doutorado atestou que modelos acrescidos de covariáveis espaciais são capazes de obter resultados mais precisos. Quanto ao objetivo principal da tese, conclui-se que, para dados espacialmente dependentes, a vizinhança exerce influência no comportamento individual relativo a viagens e, portanto, é uma variável que deve ser considerado em análises comportamentais.

Destaca-se que a aquisição de dados em painel via *Smartphones* mostra-se uma alternativa pouco onerosa e eficaz para a construção de um painel de dados que permita o estudo de padrões comportamentais de deslocamentos. A economia de recursos monetários se dá pelo elevado número de aparelhos que a população detém. Como o *smartphone* está sempre no bolso ou na bolsa do seu proprietário ao longo do dia, mostram-se uma ferramenta em potencial para aquisição de precisos dados sem a necessidade de dispendir recursos com entrevistadores. A aquisição de dados via *smartphone* permite coletar informações referentes a múltiplos dias, enriquecendo assim a análise comportamental.

O painel de dados foi validado por meio de comparação com o procedimento tradicional de diários de atividades. Dados em painel são igualmente descritivos quanto à identificação das dinâmicas de deslocamentos cotidianos, entretanto conta-se com maior praticidade de coleta e aumento da precisão de alguns tipos de informação como tempo e distância de deslocamentos.

O procedimento metodológico proposto para identificação de categorias comportamentais, através da composição de cadeias de viagens e mensuração de suas características, é capaz de segregar as amostras em grupos, gerando assim a variável dependente para modelagens paramétricas posteriores. Além disso, o procedimento metodológico proposto para incorporação de covariáveis espaciais é capaz de prover acréscimo informacional visando o aprimoramento e aumento de precisão da modelagem espacial frente a modelagem não espacial.

Ressalva-se que as discussões apresentadas se aplicam ao grupo de estudantes universitários avaliados, não sendo possível generalizar tais comportamentos para toda a população do município. Entretanto, os métodos aplicados para aquisição e processamento de dados, algoritmo de agrupamento e mecanismos de análises são de cunho geral, sendo facilmente aplicáveis a outro extrato populacional ou região que se deseje compreender os padrões de deslocamentos.

Deve-se destacar também o papel da geoestatística no procedimento metodológico. A

ferramenta da geoestatística foi fundamental para a composição do indicador global SivarG (baseado no variograma teórico, associado à um teste de hipótese), na determinação das vizinhanças (baseadas nos valores dos *lags*/distâncias entre observações dos variogramas), e, indiretamente, da determinação das covariáveis espaciais (na determinação de distâncias dos vizinhos – baseados nos variogramas).

Salienta-se que, os resultados provenientes da base de dados auxiliar promoveram contribuições metodológicas ao trabalho, ditando a ordem de aplicação das etapas para a análise espacial: (i) Análise Espacial Exploratória, inicialmente, a qual utiliza dois indicadores de associação espacial global, sendo um deles baseado no variograma teórico (SivarG). (ii) Extração de covariáveis regionais, onde a delimitação das vizinhanças também é determinada a partir de conceitos variográficos. (iii) Calibração das funções utilidade com e sem covariáveis regionais.

Novamente, os resultados são pertinentes ao banco de dados utilizado nessa pesquisa e não podem ser diretamente aplicados para toda a população da cidade. Mas, o método proposto para análise espacial, baseado em preceitos da geoestatística, também é genérico e pode ser replicado para qualquer estudo de caso, desde que os dados utilizados sejam espacialmente dependentes.

7.3 Recomendações

Recomenda-se, como continuidade deste trabalho, a expansão do espaço amostral para que as conclusões possam se estender para todo o município e não somente para um estrato da amostra.

Sugere-se também aplicar o procedimento metodológico ao mesmo estrato de estudantes, para outros *campi* da cidade ou até mesmo em outros municípios a fim de averiguar semelhanças e diferenças nos padrões comportamentais.

Além disso, o método proposto avalia o caráter de motorização das viagens relativo a somente duas categorias de deslocamentos: motorizados ou não motorizados. A implementação de uma etapa que promova o cruzamento de informações com as linhas de sistemas de transporte público pode ser capaz de trazer mais precisão ao caráter de motorização, podendo avaliar, além dos deslocamentos não motorizados, se os deslocamentos motorizados foram realizados por modos motorizados particulares ou modos motorizados coletivos.

Um ressalva metodológica deste trabalho é a de que usuários de bicicleta podem atingir velocidades equivalentes aos veículos motorizados em fluxo regular, tendo assim seus deslocamentos classificados erroneamente como modo motorizado. Sugere-se a criação de um método capaz de distinguir o uso da bicicleta, classificando-a corretamente como uma locomoção não motorizada.

REFERÊNCIAS

- ACKER, V. V.; WITLOX, F. Car ownership as a mediating variable in car travel behaviour research using a structural equation modelling approach to identify its dual relationship. **Journal of Transport Geography**, Elsevier, v. 18, n. 1, p. 65–74, 2010.
- AGUILERA, A.; MASSOT, M.-H.; PROULHAC, L. Exploring the relationship between work and travel behavior on weekdays: Analysis of Paris region travel survey over 20 years. **Transportation Research Record: Journal of the Transportation Research Board**, Transportation Research Board of the National Academies, n. 2135, p. 69–77, 2009.
- AHMADI, S. H.; SEDGHAMIZ, A. Geostatistical analysis of spatial and temporal variations of groundwater level. **Environmental monitoring and assessment**, Springer, v. 129, n. 1-3, p. 277–294, 2007.
- AJAY, K.; LEVINSON, D. Chained trips in Montgomery County, Maryland. **ITE Journal**, University of Minnesota: Nexus Research Group, p. 27–32, 1995.
- AMANCIO, M. A. **Relacionamento entre a forma urbana e as viagens a pé**. 2005. Dissertação (Mestrado) — Universidade Federal de São Carlos, 2005.
- ANAS, A. **Residential Location Markets and Urban Transportation. Economic Theory, Econometrics and Policy Analysis With Discrete Choice Models**. New York: Academic Press, 1982.
- ANDA, C.; ERATH, A.; FOURIE, P. J. Transport modelling in the age of big data. **International Journal of Urban Sciences**, Taylor & Francis, v. 21, n. 1, p. 19–42, 2017.
- ANDERSON, M. K. Characteristics of trips and travellers in private and public transportation in the Danish travel survey data. In: **Selected Proceedings from the Annual Transport Conference at Aalborg University**. [S.l.: s.n.], 2010.
- ANDRESEN, M. A. A spatial analysis of crime in Vancouver, British Columbia: A synthesis of social disorganization and routine activity theory. **The Canadian Geographer/Le Géographe canadien**, Wiley Online Library, v. 50, n. 4, p. 487–502, 2006.
- ANSELIN, L. Local indicators of spatial association—LISA. **Geographical analysis**, Wiley Online Library, v. 27, n. 2, p. 93–115, 1995.
- _____. Thirty years of spatial econometrics. **Papers in regional science**, Wiley Online Library, v. 89, n. 1, p. 3–25, 2010.
- ANTONINI, C.; ARGILÉS-BOSCH, J. M. Productivity and environmental costs from intensification of farming. a panel data analysis across EU regions. **Journal of Cleaner Production**, Elsevier, v. 140, p. 796–803, 2017.

ARENTZE, T.; TIMMERMANS, H. Information gain, novelty seeking and travel: a model of dynamic activity-travel behavior under conditions of uncertainty. **Transportation Research Part A: Policy and Practice**, Elsevier, v. 39, n. 2-3, p. 125–145, 2005.

ASSIRATI, L. et al. Interpolação bi-linear para simulação de dados espacialmente correlacionados: Uma aplicação relativa à demanda por transportes. In: ANPET. **Anais do XXX Congresso da Associação Nacional de Ensino e Pesquisa em Transportes**. Rio de Janeiro, 2016.

AXHAUSEN, K. W. et al. Observing the rhythms of daily life: A six-week travel diary. **Transportation**, Springer, v. 29, n. 2, p. 95–124, 2002.

BAILEY, T.; GATTREL, A. **Spatial data analysis by example**. London: Longman, 1995.

BALTAGI, B. H. **The Oxford handbook of panel data**. Oxford: Oxford Handbooks, 2014.

BAMBERG, S.; RÖLLE, D.; WEBER, C. Does habitual car use not lead to more resistance to change of travel mode? **Transportation**, Springer, v. 30, n. 1, p. 97–108, 2003.

BATELLA, W. B.; DINIZ, A. M. A. Análise espacial dos condicionantes da criminalidade violenta no estado de Minas Gerais. **Sociedade & Natureza**, Universidade Federal de Uberlândia, Instituto de Geografia, Programa de Pós-Graduação em Geografia, p. 151–163, 2010.

BEKHOR, S.; PRASHKER, J. N. Gev-based destination choice models that account for unobserved similarities among alternatives. **Transportation Research Part B: Methodological**, Elsevier, v. 42, n. 3, p. 243–262, 2008.

BEN-AKIVA, M. E.; LERMAN, S. R. **Discrete choice analysis: theory and application to travel demand**. [S.l.]: MIT press, 1985. v. 9.

BERALDI, F. B. **Os efeitos da Operação Direção Segura sobre a mortalidade no trânsito terrestre dos municípios de São Paulo**. 2017. Tese (Doutorado) — Fundação Getúlio Vargas, 2017.

BERRY, B. J. L.; MARBLE, D. F. **Spatial analysis: a reader in statistical geography**. [S.l.]: Prentice-Hall, 1968.

BHAT, C. An analysis of evening commute stop-making behavior using repeated choice observations from a multi-day survey. **Transportation Research Part B: Methodological**, Elsevier, v. 33, n. 7, p. 495–510, 1999.

BHAT, C. R. Flexible model structures for discrete choice analysis. **Handbook of Transport Modelling**, Ch, v. 5, p. 75–104, 2008.

BHAT, C. R.; GUO, J. A mixed spatially correlated logit model: formulation and application to residential choice modeling. **Transportation Research Part B: Methodological**, Elsevier, v. 38, n. 2, p. 147–168, 2004.

- BHAT, C. R.; KOPPELMAN, F. S. A conceptual framework of individual activity program generation. **Transportation Research Part A: Policy and Practice**, Elsevier, v. 27, n. 6, p. 433–446, 1993.
- BIELENKI, C.; RAIA, A.; SILVA, A. O uso de geoprocessamento no planejamento de pesquisa origem-destino no município de São Carlos. **Revista Minerva**, v. 5, n. 2, p. 159–167, 2008.
- BILLÉ, A. G.; ARBIA, G. Spatial discrete choice and spatial limited dependent variable models: a review with an emphasis on the use in regional health economics. **arXiv preprint arXiv:1302.2267**, 2013.
- BIOGEME, a. Pythonbiogeme: A short introduction, technical report transp-or 160706, Bierlaire, M. Transport and Mobility Laboratory, ENAC, EPFL, 2016.
- BLANC, E.; SCHLENKER, W. The use of panel models in assessments of climate impacts on agriculture. **Review of Environmental Economics and Policy**, Oxford University Press, v. 11, n. 2, p. 258–279, 2017.
- BOARNET, M. G.; SARMIENTO, S. Can land-use policy really affect travel behaviour? a study of the link between non-work travel and land-use characteristics. **Urban Studies**, Sage: London, England, v. 35, n. 7, p. 1155–1169, 1998.
- BORCARD, D.; GILLET, F.; LEGENDRE, P. Spatial analysis of ecological data. In: **Numerical ecology with R**. New York: Springer, 2011. p. 227–292.
- BOWMAN, J. L.; BEN-AKIVA, M. E. Activity-based disaggregate travel demand model system with activity schedules. **Transportation Research Part A: Policy and practice**, Elsevier, v. 35, n. 1, p. 1–28, 2001.
- BRICKA, S.; BHAT, C. Comparative analysis of global positioning system-based and travel survey-based data. **Transportation Research Record: Journal of the Transportation Research Board**, Transportation Research Board of the National Academies, n. 1972, p. 9–20, 2006.
- BROCK, W. A.; DURLAUF, S. N. Discrete choice with social interactions. **The Review of Economic Studies**, Wiley-Blackwell, v. 68, n. 2, p. 235–260, 2001.
- BRUTON, M.; LEE, D. **Introduction to Transportation-Planning**. Washington, DC: Amer Inst Planners 1776 Massachusetts Ave NW 20036, 1978.
- BURROUGH, P. Principles of geographical information systems for land resources assessment. **Monographs on Soil and Resources Survey**, n. 12, 1986.
- CÂMARA, G. et al. Análise espacial e geoprocessamento. **Análise espacial de dados geográficos**, Embrapa Cerrados Brasília, v. 2, 2002.
- CAMBARDELLA, C.; KARLEN, D. Spatial analysis of soil fertility parameters. **Precision Agriculture**, Springer, v. 1, n. 1, p. 5–14, 1999.
- CAO, X.; MOKHTARIAN, P. L.; HANDY, S. L. Do changes in neighborhood characteristics lead to changes in travel behavior? a structural equations modeling approach. **Transportation**, Springer, v. 34, n. 5, p. 535–556, 2007.

CARVALHO, C. H. R. et al. Tarifação e financiamento do transporte público urbano. Instituto de Pesquisa Econômica Aplicada (IPEA), 2013.

CARVALHO, C. H. R. d.; PEREIRA, R. H. M. Efeitos da variação da tarifa e da renda da população sobre a demanda de transporte público coletivo urbano no Brasil. Instituto de Pesquisa Econômica Aplicada (IPEA), 2011.

CERVERO, R.; GOLUB, A.; NEE, B. City carshare: longer-term travel demand and car ownership impacts. **Transportation Research Record: Journal of the Transportation Research Board**, TRB, n. 1992, p. 70–80, 2007.

CHAKRAVARTI, D.; SINHA, A. R.; KIM, J. Choice research: A wealth of perspectives. **Marketing Letters**, Springer, v. 16, n. 3-4, p. 173–182, 2005.

CLARK, B.; CHATTERJEE, K.; MELIA, S. Changes to commute mode: The role of life events, spatial context and environmental attitude. **Transportation Research Part A: Policy and Practice**, Elsevier, v. 89, p. 89–105, 2016.

CLARK, P. J.; EVANS, F. C. Distance to nearest neighbor as a measure of spatial relationships in populations. **Ecology**, Wiley Online Library, v. 35, n. 4, p. 445–453, 1954.

CORNWELL, C.; TRUMBULL, W. N. Estimating the economic model of crime with panel data. **The Review of economics and Statistics**, JSTOR, p. 360–366, 1994.

COSTA, L. da F.; CESAR, R. M. **Shape classification and analysis: theory and practice**. [S.l.]: CRC Press, Inc., 2009.

CURTIS, C.; PERKINS, T. Travel behaviour: A review of recent literature. **Perth, WA: Urbanet, Curtin University of Technology**, v. 2, n. 3, p. 77, 2006.

DAISY, N. S. et al. Understanding and modeling the activity-travel behavior of university commuters at a large canadian university. **Journal of Urban Planning and Development**, American Society of Civil Engineers, v. 144, n. 2, p. 04018006, 2018.

DEATON, A. Panel data from time series of cross-sections. **Journal of econometrics**, Elsevier, v. 30, n. 1-2, p. 109–126, 1985.

DISSANAYAKE, D.; MORIKAWA, T. Household travel behavior in developing countries: Nested logit model of vehicle ownership, mode choice, and trip chaining. **Transportation Research Record: Journal of the Transportation Research Board**, TRB, n. 1805, p. 45–52, 2002.

ENGEBRETSEN, Ø.; CHRISTIANSEN, P. Urban structure and travel behaviour. **Bystruktur og transport—En studie av personreiser i byer og tettsteder**, 2011.

FOTHERINGHAM, A. S.; BRUNSDON, C.; CHARLTON, M. **Quantitative geography: perspectives on spatial data analysis**. [S.l.]: Sage, 2000.

FOTHERINGHAM, A. S.; CHARLTON, M. E.; BRUNSDON, C. Geographically weighted regression: a natural evolution of the expansion method for spatial data analysis. **Environment and planning A**, Sage: London, England, v. 30, n. 11, p. 1905–1927, 1998.

FRANK, L. et al. Urban form, travel time, and cost relationships with tour complexity and mode choice. **Transportation**, Springer, v. 35, n. 1, p. 37–54, 2008.

FRUSTI, T.; BHAT, C.; AXHAUSEN, K. Exploratory analysis of fixed commitments in individual activity—travel patterns. **Transportation Research Record: Journal of the Transportation Research Board**, TRB, n. 1807, p. 101–108, 2002.

GEARY, R. C. The contiguity ratio and statistical mapping. **The Incorporated Statistician**, JSTOR, v. 5, n. 3, p. 115–146, 1954.

GETIS, A.; ORD, J. K. The analysis of spatial association by use of distance statistics. **Geographical Analysis**, Wiley Online Library, v. 24, n. 3, p. 189–206, 1992.

GIULIANO, G.; DARGAY, J. Car ownership, travel and land use: a comparison of the US and Great Britain. **Transportation Research Part A: Policy and Practice**, Elsevier, v. 40, n. 2, p. 106–124, 2006.

GOLOB, T. F. A simultaneous model of household activity participation and trip chain generation. **Transportation Research Part B: Methodological**, Elsevier, v. 34, n. 5, p. 355–376, 2000.

GOMES, V. A. et al. Estudo comparativo do uso de krigagem para estimação da escolha do modo particular motorizado. **Revista Brasileira de Cartografia**, v. 68, n. 4, 2016.

_____. Kriging geostatistical methods for travel mode choice: a spatial data analysis to travel demand forecasting. **Open Journal of Statistics**, Scientific Research Publishing, v. 6, n. 03, p. 514, 2016.

GOOVAERTS, P. Factorial kriging analysis: a useful tool for exploring the structure of multivariate spatial soil information. **Journal of Soil Science**, Wiley Online Library, v. 43, n. 4, p. 597–619, 1992.

_____. Geostatistical analysis of disease data: accounting for spatial support and population density in the isopleth mapping of cancer mortality risk using area-to-point Poisson kriging. **International Journal of Health Geographics**, BioMed Central, v. 5, n. 1, p. 52, 2006.

GOUDIE, D. Zonal method for urban travel surveys: sustainability and sample distance from the CBD. **Journal of Transport Geography**, Elsevier, v. 10, n. 4, p. 287–301, 2002.

GRIFFITH, D. A. **Advanced spatial statistics: special topics in the exploration of quantitative spatial data series**. Netherlands: Springer Science & Business Media, 2012. v. 12.

GUO, L. et al. Global and local indicators of spatial association between points and polygons: A study of land use change. **International Journal of Applied Earth Observation and Geoinformation**, Elsevier, v. 21, p. 384–396, 2013.

HAMMADOU, H. et al. How to incorporate the spatial dimension in destination choice models: the case of Antwerp. **Transportation Planning and Technology**, Taylor & Francis, v. 31, n. 2, p. 153–181, 2008.

HANSON, S.; HUFF, J. O. Assessing day-to-day variability in complex travel patterns. **Transportation Research Record**, v. 891, p. 18–24, 1981.

HERNANDEZ, A. R. et al. Analysis of infant mortality trends and risk factors in Porto Alegre, Rio Grande do Sul State, Brazil, 1996–2008. **Cadernos de Saude Publica**, SciELO Brasil, v. 27, n. 11, p. 2188–2196, 2011.

HESS, S. **Advanced discrete choice models with applications to transport demand**. 2005. Tese (Doutorado) — University of London, 2005.

HOSMER, D. W.; LEMESHOW, S.; STURDIVANT, R. X. **Applied Logistic Regression**. [S.l.]: John Wiley & Sons, 2013. v. 398.

HOWARD, C.; BURNS, E. Cycling to work in phoenix: route choice, travel behavior, and commuter characteristics. **Transportation Research Record: Journal of the Transportation Research Board**, Transportation Research Board of the National Academies, n. 1773, p. 39–46, 2001.

HSIAO, C. **Analysis of panel data**. [S.l.]: Cambridge University Press, 2014.

HUFF, J. O.; HANSON, S. Repetition and variability in urban travel. **Geographical Analysis**, Wiley Online Library, v. 18, n. 2, p. 97–114, 1986.

ISAAKS, E. H.; SRIVASTAVA, R. M. An introduction to applied geostatistics. Oxford University Press, New York, 1989.

JERRETT, M. et al. Spatial analysis of air pollution and mortality in Los Angeles. **Epidemiology**, JSTOR, p. 727–736, 2005.

KIM, S.; ULFARSSON, G. F.; HENNESSY, J. T. Analysis of light rail rider travel behavior: impacts of individual, built environment, and crime characteristics on transit access. **Transportation Research Part A: Policy and Practice**, Elsevier, v. 41, n. 6, p. 511–522, 2007.

KLOMP, J.; HOOGEZAND, B. Natural disasters and agricultural protection: A panel data analysis. **World Development**, Elsevier, v. 104, p. 404–417, 2018.

KOCKELMAN, K. Travel behavior as function of accessibility, land use mixing, and land use balance: evidence from San Francisco Bay Area. **Transportation Research Record: Journal of the Transportation Research Board**, TRB, n. 1607, p. 116–125, 1997.

KOPPELMAN, F. S.; SETHI, V. Incorporating variance and covariance heterogeneity in the generalized nested logit model: an application to modeling long distance travel choice behavior. **Transportation Research Part B: Methodological**, Elsevier, v. 39, n. 9, p. 825–853, 2005.

KOUSER, S.; QAIM, M. Impact of BT cotton on pesticide poisoning in smallholder agriculture: A panel data analysis. **Ecological Economics**, Elsevier, v. 70, n. 11, p. 2105–2113, 2011.

KREMPI, A. P. **Explorando recursos de estatística espacial para análise da acessibilidade da cidade de Bauru**. 2004. Tese (Doutorado) — Universidade de São Paulo, 2004.

KRIGE, D. G. A Statistical approach to some basic mine valuation problems on the Witwatersrand. **Journal of the Southern African Institute of Mining and Metallurgy**, Southern African Institute of Mining and Metallurgy, v. 52, n. 6, p. 119–139, 1951.

LAWRANCE, E. C. Poverty and the rate of time preference: evidence from panel data. **Journal of Political economy**, The University of Chicago Press, v. 99, n. 1, p. 54–77, 1991.

LEITÃO, N. C. Financial development and economic growth: a panel data approach. **Theoretical and Applied Economics**, General Association of the Economists in Romania, p. 15–24, 2010.

LISCO, T. **The Value of Commuters Travel Time: A Study in Urban Transportation**. Chicago: University of Chicago, 1967.

LOPES, S. B. **Efeitos da dependência espacial em modelos de previsão de demanda por transporte**. 2005. Tese (Doutorado) — Universidade de São Paulo, 2005.

MACQUEEN, J. et al. Some methods for classification and analysis of multivariate observations. In: OAKLAND, CA, USA. **Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability**. [S.l.], 1967. v. 1, n. 14, p. 281–297.

MARQUES, L. D. et al. Modelos dinâmicos com dados em painel: revisão de literatura. **Centro de Estudos Macroeconômicos e Previsão, Faculdade de Economia do Porto**, 2000.

MENDONÇA, M. J. C. d.; MOTTA, R. S. Saúde e saneamento no Brasil. Instituto de Pesquisa Econômica Aplicada (IPEA), 2005.

MILTON, J. C.; SHANKAR, V. N.; MANNERING, F. L. Highway accident severities and the mixed logit model: an exploratory empirical analysis. **Accident Analysis & Prevention**, Elsevier, v. 40, n. 1, p. 260–266, 2008.

MOHAMMADIAN, A.; HAIDER, M.; KANAROGLOU, P. S. Incorporating spatial dependencies in random parameter discrete choice models. In: **84th Annual Meeting of the Transportation Research Board, Washington, DC**. [S.l.: s.n.], 2005.

MORAN, P. A. Notes on continuous stochastic phenomena. **Biometrika**, JSTOR, v. 37, n. 1/2, p. 17–23, 1950.

MORGAN, W. B. et al. **Agriculture in the third world: a spatial analysis**. [S.l.]: Westview Press, 1978.

MURRAY, A. T. et al. Exploratory spatial data analysis techniques for examining urban crime: Implications for evaluating treatment. **British Journal of Criminology**, Oxford University Press, v. 41, n. 2, p. 309–329, 2001.

NAIZER, C. **Procedimento Metodológico para Proposta de Indicadores de Associação espacial Global e Local através de conceitos variográficos**. 2018. Dissertação (Mestrado) — Universidade de São Paulo, 2018.

NAIZER, C.; PITOMBO, C. S. Procedimento metodológico para proposta de indicadores de associação espacial global através de conceitos variográficos. In: ANPET. **Anais do XXV Congresso da Associação Nacional de Ensino e Pesquisa em Transportes**. Recife, 2017.

NEWBOLD, K. B. et al. Travel behavior within canada's older population: a cohort analysis. **Journal of Transport Geography**, Elsevier, v. 13, n. 4, p. 340–351, 2005.

NG, W.-S.; ACKER, A. Understanding urban travel behaviour by gender for efficient and equitable transport policies. 2018.

NOGUEIRA, E. L. P. **O uso da calculadora gráfica GeoGebra no smartphone como ferramenta para o ensino das funções exponencial e logarítmica**. 2018. Dissertação (Mestrado) — Brasil, 2018.

ODEN, N. L. Assessing the significance of a spatial correlogram. **Geographical Analysis**, v. 16, n. 1, p. 1–16, 1984.

ORTÚZAR, J.; WILLUMSEN, L. G. **Modelling transport**. London: John Wiley & Sons, 2011.

PÁEZ, A. Spatial perspectives on urban systems: developments and directions. **Journal of Geographical Systems**, Springer, v. 9, n. 1, p. 1–6, 2007.

PÁEZ, A. et al. Development of an indicator to assess the spatial fit of discrete choice models. **Transportation Research Part B: Methodological**, Elsevier, v. 56, p. 217–233, 2013.

PÁEZ, A.; SCOTT, D. M.; VOLZ, E. A discrete-choice approach to modeling social influence on individual decision making. **Environment and Planning B: Planning and Design**, SAGE Publications Sage UK: London, England, v. 35, n. 6, p. 1055–1069, 2008.

PAN, H.; SHEN, Q.; ZHANG, M. Influence of urban form on travel behaviour in four neighbourhoods of Shanghai. **Urban studies**, SAGE Publications Sage UK: London, England, v. 46, n. 2, p. 275–294, 2009.

PARENTE, S. T.; MCCULLOUGH, J. S. Health information technology and patient safety: evidence from panel data. **Health Affairs**, Project HOPE-The People-to-People Health Foundation, Inc., v. 28, n. 2, p. 357–360, 2009.

PAS, E. I. Intrapersonal variability and model goodness-of-fit. **Transportation Research Part A: General**, Elsevier, v. 21, n. 6, p. 431–438, 1987.

PAS, E. I.; SUNDAR, S. Intrapersonal variability in daily urban travel behavior: some additional evidence. **Transportation**, Springer, v. 22, n. 2, p. 135–150, 1995.

PAULSEN, D. J.; ROBINSON, M. B. **Spatial aspects of crime: Theory and practice**. [S.l.]: Pearson, 2004.

PENDYALA, R. Measuring day-to-day variability in travel behavior using gps data. **FHWA**, Washington, DC, 1999.

- PEROBELLI, F. S. et al. Produtividade do setor agrícola brasileiro (1991-2003): uma análise espacial. **Nova economia**, SciELO Brasil, v. 17, n. 1, p. 65–91, 2007.
- PFEIFFER, D. et al. **Spatial analysis in epidemiology**. [S.l.]: Oxford University Press Oxford, 2008. v. 142.
- PITOMBO, C. **Análise do comportamento subjacente ao encadeamento de viagens através do uso de minerador de dados**. 149 p. 2003. Dissertação (Mestrado) — Escola de Engenharia de São Carlos, Universidade de São Paulo, São Carlos, 2003.
- PITOMBO, C. S.; GOMES, M. M. Study of work-travel related behavior using principal component analysis. **Open Journal of Statistics**, Scientific Research Publishing, v. 4, n. 11, p. 889, 2014.
- PITOMBO, C. S.; KAWAMOTO, E.; SOUSA, A. J. An exploratory analysis of relationships between socioeconomic, land use, activity participation variables and travel patterns. **Transport Policy**, Elsevier, v. 18, n. 2, p. 347–357, 2011.
- PITOMBO, C. S. et al. A two-step method for mode choice estimation with socioeconomic and spatial information. **Spatial Statistics**, Elsevier, v. 11, p. 45–64, 2015.
- PIZZOL, B. **Padrões de atividades de residentes de Paraisópolis: Análise de dados de múltiplos dias coletados com smartphones**. 2018. Dissertação (Mestrado) — Universidade de São Paulo, 2018.
- POWELL, L. M. Fast food costs and adolescent body mass index: evidence from panel data. **Journal of Health Economics**, Elsevier, v. 28, n. 5, p. 963–970, 2009.
- PRITCHARD, J. P. et al. Spatial analysis of transportation-related social exclusion in the Lisbon metropolitan area. **Procedia-Social and Behavioral Sciences**, Elsevier, v. 111, p. 440–449, 2014.
- RAMADURAI, G.; UKKUSURI, S. Dynamic user equilibrium model for combined activity-travel choices using activity-travel supernetwork representation. **Networks and Spatial Economics**, Springer, v. 10, n. 2, p. 273–292, 2010.
- REIS, C. T. et al. A interiorização da epidemia de HIV/AIDS e o fluxo intermunicipal de internação hospitalar na Zona da Mata, Minas Gerais, Brasil: uma análise espacial. **Cadernos de Saúde Pública**, SciELO Public Health, v. 24, p. 1219–1228, 2008.
- REITER, G. R. Infraestrutura de transportes no Brasil: uma análise com dados em painel no período de 1995 a 2008. 2015.
- REZAEIAN, M. et al. Geographical epidemiology, spatial analysis and geographical information systems: a multidisciplinary glossary. **Journal of Epidemiology & Community Health**, BMJ Publishing Group Ltd, v. 61, n. 2, p. 98–102, 2007.
- RIBEIRO, E. R.; CUNHA, A. L. Análise exploratória de método para definição de dia típico utilizando transformada wavelet e análise de agrupamento. In: ANPET. **Anais do XXV Congresso da Associação Nacional de Ensino e Pesquisa em Transportes**. Recife, 2017.

RIPLEY, B. D. The second-order analysis of stationary point processes. **Journal of Applied Probability**, Cambridge University Press, v. 13, n. 2, p. 255–266, 1976.

ROCKOFF, J. E. The impact of individual teachers on student achievement: Evidence from panel data. **American Economic Review**, v. 94, n. 2, p. 247–252, 2004.

SANTOLIN, R.; JR, F. G. J.; REIS, J. C. d. Lei de responsabilidade fiscal e implicações na despesa de pessoal e de investimento nos municípios mineiros: um estudo com dados em painel dinâmico. **Estudos Econômicos (São Paulo)**, SciELO Brasil, v. 39, n. 4, p. 895–923, 2009.

SANTOS, A. **Summary of travel trends : 2009 National Household Travel Survey**. Washington, DC: U.S. Department of Transportation, Federal Highway Administration, 2011.

SCHÖNFELDER, S. et al. Exploring the potentials of automatically collected gps data for travel behaviour analysis: A swedish data source. **Arbeitsberichte Verkehrs-und Raumplanung**, ETH, Eidgenössische Technische Hochschule Zürich, Institut für Verkehrsplanung, Transporttechnik, Strassen-und Eisenbahnbau IVT, v. 124, 2002.

SERRANO, R. M.; VALCARCE, E. V. **Técnicas econométricas para el tratamiento de datos espaciales: la econometría espacial**. [S.l.]: Edicions Universitat Barcelona, 2000. v. 44.

SHARABY, N.; SHIFTAN, Y. The impact of fare integration on travel behavior and transit ridership. **Transport Policy**, Elsevier, v. 21, p. 63–70, 2012.

SILVA, A. Rodrigues da. Pesquisa origem-destino da cidade de são carlos. **Relatório do projeto de pesquisa**, FAPESP (Fundação de Amparo à Pesquisa do Estado de São Paulo), v. 4, p. 15843–4, 2008.

SILVA, G. J. C. da; MENEZES, L. B.; NEDER, H. D. Qualidade da malha rodoviária, custos econômicos associados e determinantes dos acidentes de trânsito no Brasil: avaliação e proposição de política. **Revista de Políticas Públicas**, v. 19, n. 1, p. 327–347, 2016.

SILVA, J. A. Spatial self-selection in land-use–travel behavior interactions: accounting simultaneously for attitudes and socioeconomic characteristics. **Journal of Transport and Land Use**, JSTOR, v. 7, n. 2, p. 63–84, 2014.

SIMMA, A.; AXHAUSEN, K. Commitments and modal usage: analysis of German and Dutch panels. **Transportation Research Record: Journal of the Transportation Research Board**, TRB, n. 1854, p. 22–31, 2003.

SMIRNOV, O. A. Modeling spatial discrete choice. **Regional Science and Urban Economics**, Elsevier, v. 40, n. 5, p. 292–298, 2010.

SOUSA, T. R. V.; LEITE, P. A. M. Análise por dados em painel do status de saúde no nordeste brasileiro. **Revista de Saúde Pública**, v. 42, p. 796–804, 2008.

SPRING, I. Sistema de processamento de informações georreferenciadas. **Manual de uso**, 2006.

- STOPHER, P. et al. Reducing burden and sample sizes in multiday household travel surveys. **Transportation Research Record: Journal of the Transportation Research Board**, TRB, n. 2064, p. 12–18, 2008.
- _____. **Methods for Household Travel Surveys**. Washington DC: National Academy Press, 1996. (National cooperative highway research program; Synthesis of highway practice). ISBN 9780309060097.
- SUSILO, Y.; KITAMURA, R. Analysis of day-to-day variability in an individual's action space: exploration of 6-week mobidrive travel diary data. **Transportation Research Record: Journal of the Transportation Research Board**, TRB, n. 1902, p. 124–133, 2005.
- SUSILO, Y. O.; AXHAUSEN, K. W. Repetitions in individual daily activity–travel–location patterns: a study using the Herfindahl–Hirschman index. **Transportation**, Springer, v. 41, n. 5, p. 995–1011, 2014.
- TAKANO, M. S. M. **Análise da influência da forma urbana no comportamento de viagens encadeadas com base em padrões de atividades**. 2011. Dissertação (Mestrado) — Universidade de Brasília, 2011.
- TEIXEIRA, G. **Uso de dados censitários para identificação de zonas homogêneas para planejamento de transportes utilizando estatística espacial**. 2003. Dissertação (Mestrado) — Universidade de Brasília, 2003.
- THØGERSEN, J. Understanding repetitive travel mode choices in a stable context: A panel study approach. **Transportation Research Part A: Policy and Practice**, Elsevier, v. 40, n. 8, p. 621–638, 2006.
- TOBLER, W. R. A computer movie simulating urban growth in the Detroit region. **Economic geography**, Taylor & Francis, v. 46, n. sup1, p. 234–240, 1970.
- TOTH, B. et al. Iregional shopping center linked trip distribution. **ITE Journal**, v. 60, n. 5, 1990.
- TRAPP, G. S. et al. On your bike! a cross-sectional study of the individual, social and environmental correlates of cycling to school. **International Journal of Behavioral Nutrition and Physical Activity**, BioMed Central, v. 8, n. 1, p. 123, 2011.
- _____. Increasing children's physical activity: Individual, social, and environmental factors associated with walking to and from school. **Health Education & Behavior**, Sage Publications Sage CA: Los Angeles, CA, v. 39, n. 2, p. 172–182, 2012.
- VASCONCELLOS, E. A. D. Urban change, mobility and transport in São Paulo: three decades, three cities. **Transport Policy**, Elsevier, v. 12, n. 2, p. 91–104, 2005.
- WACKERNAGEL, H. Geostatistics. **Wiley Statistics Reference**, Wiley Online Library, 2014.
- WILLIAMS, K. Spatial planning, urban form and sustainable transport: an introduction. In: **Spatial Planning, Urban Form and Sustainable Transport**. [S.l.]: Routledge, 2017. p. 15–28.

WISE, S.; HAINING, R.; SIGNORETTA, C. The role of visualization in the Exploratory Spatial Data Analysis of Area-based Data. In: **Proceedings of the 3rd International Conference on GeoComputation**. University of Bristol, UK: [s.n.], 1998.

YAMAMOTO, J. K.; LANDIM, P. M. B. **Geoestatística: conceitos e aplicações**. [S.l.]: Oficina de Textos, 2015.

YANG, Z. et al. Car ownership and urban development in chinese cities: A panel data analysis. **Journal of Transport Geography**, Elsevier, v. 58, p. 127–134, 2017.