UNIVERSIDADE ESTADUAL DE CAMPINAS

Faculdade de Engenharia Civil, Arquitetura e Urbanismo

ALEXANDRE FRAZÃO D' ANDRÉA

ANÁLISE ESPACIAL APLICADA A MODELOS TRADICIONAIS DE PRODUÇÃO DE VIAGENS NO ESTUDO DE CASO DA REGIÃO METROPOLITANA DE CAMPINAS

CAMPINAS

2018

ALEXANDRE FRAZÃO D' ANDRÉA

ANÁLISE ESPACIAL APLICADA A MODELOS TRADICIONAIS DE PRODUÇÃO DE VIAGENS NO ESTUDO DE CASO DA REGIÃO METROPOLITANA DE CAMPINAS

Dissertação de Mestrado apresentada a Faculdade de Engenharia Civil, Arquitetura e Urbanismo da Unicamp para a obtenção do título de Mestre em Engenharia Civil, na área de Transportes.

Orientadora: Profª. Drª. Maria Teresa Françoso

ESSE EXEMPLAR CORRESPONDE À VERSÃO FINAL DA DISSERTAÇÃO DEFENDIDA PELO ALUNO ALEXANDRE FRAZÃO D' ANDRÉA E ORIENTADO PELA PROFESSORA MARIA TERESA FRANCOSO

ASSINATURA DA ORIENTADORA:

_____

CAMPINAS

2018

Agência(s) de fomento e nº(s) de processo(s): Não se aplica.
ORCID: https://orcid.org/0000-0002-0227-5695

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca da Área de Engenharia e Arquitetura
Luciana Pietrosanto Milla - CRB 8/8129

An24a

Andréa, Alexandre Frazão D', 1971-
Análise espacial aplicada a modelos tradicionais de produção de viagens no estudo de caso da Região Metropolitana de Campinas / Alexandre Frazão D' Andréa. – Campinas, SP : [s.n.], 2018.

Orientador: Maria Teresa Françoso.
Dissertação (mestrado) – Universidade Estadual de Campinas, Faculdade de Engenharia Civil, Arquitetura e Urbanismo.

1. Análise espacial. 2. Autocorrelação espacial. 3. Transportes - Planejamento. 4. Transportes - Modelos matemáticos. I. Françoso, Maria Teresa, 1963-. II. Universidade Estadual de Campinas. Faculdade de Engenharia Civil, Arquitetura e Urbanismo. III. Título.

Informações para Biblioteca Digital

Título em outro idioma: Spatial analysis applied to traditional trip production models in the case study of the Campinas Metropolitan Region
Palavras-chave em inglês:
Spatial analysis
Spatial autocorrelation
Transport - Planning
Transportation - Mathematical models
Área de concentração: Transportes
Titulação: Mestre em Engenharia Civil
Banca examinadora:
Maria Teresa Françoso [Orientador]
Humberto de Paiva Junior
Flávia da Fonseca Feitosa
Data de defesa: 27-07-2018
Programa de Pós-Graduação: Engenharia Civil

UNIVERSIDADE ESTADUAL DE CAMPINAS
FACULDADE DE ENGENHARIA CIVIL, ARQUITETURA E URBANISMO

ANÁLISE ESPACIAL APLICADA A MODELOS TRADICIONAIS DE PRODUÇÃO DE VIAGENS NO ESTUDO DE CASO DA REGIÃO METROPOLITANA DE CAMPINAS

ALEXANDRE FRAZÃO D' ANDRÉA

Dissertação de Mestrado aprovada pela Banca Examinadora, constituída por:

Profª. Drª. Maria Teresa Françoso
Presidente e Orientadora/ UNICAMP

Prof. Dr. Humberto de Paiva Junior
UFABC

Profª. Drª Flávia da Fonseca Feitosa
UFABC

A Ata da defesa com as respectivas assinaturas dos membros encontra-se no processo de vida acadêmica do aluno.

Campinas, 27 de julho de 2018

DEDICATÓRIA

Dedico este trabalho à Priscila e a Olivia.

AGRADECIMENTOS

À minha esposa Priscila, por compreender o significado do mestrado para minha vida e por ter dado total apoio desde o início até o final da pesquisa, à Olivia, por me ensinar a filtrar o que realmente é importante.

Ao meu pai Edson "*in memorian*", por ser meu guia diante os desafios, à minha mãe Nita, por possibilitar que eu sempre estudasse, mesmo diante de enormes dificuldades, à minha irmã Silvia e ao Ricardo por sempre estarem por perto, à Leticia por me fazer pensar que o estudo nunca termina e que sempre vale a pena.

A todos os professores do programa de pós-graduação em engenharia civil da Faculdade de Engenharia Civil Arquitetura e Urbanismo da Unicamp, em especial à minha orientadora professora Maria Teresa Francoso.

Ao meu amigo Fabio Delospital, pelo companheirismo e principalmente por ter compartilhado comigo o grande desafio de nosso primeiro congresso internacional, Road Federation IRF World Meeting em 2013 na cidade de Riyadh na Arábia Saudita.

À GPO Sistran por compreender a importância do mestrado e possibilitar diversas vezes a minha ausência do trabalho.

Aos meus amigos Felipe Dias, Luciano Peron e David Soares, por suas ideias durante o período de desenvolvimento dessa pesquisa de mestrado.

“A melhor coisa sobre o futuro é que vem um só dia de cada vez”. (Abraham Lincoln)

RESUMO

As técnicas de regressões lineares são as mais utilizadas para o desenvolvimento de modelos de produção de viagens em zonas de tráfego, as quais partem das premissas de independência, linearidade, homocedasticidade dos dados da variável resposta. Estas premissas são restritivas e raramente atendidas completamente em modelos de produção de viagens, especialmente porque as variáveis explicativas (características socioeconômicas, demográficas e de uso do solo) não ocorrem de forma aleatória e dependem de fatores geralmente relacionados com sua localização. Segundo a lei de Tobler, os valores de variáveis espaciais são atribuídos à proximidade entre medidas, dessa mesma variável, em diferentes localidades, portanto, há uma tendência de que o valor de uma ou mais variáveis associadas a uma determinada localização assemelham-se mais aos valores observados em sua vizinhança do que ao restante das localizações do conjunto amostral. Assim é possível inferir que os procedimentos matemáticos clássicos de regressões lineares utilizados para a elaboração de modelos de produção de viagens violam principalmente as premissas de independência e homocedasticidade dos dados, da variável resposta, ao não considerar que as viagens produzidas são influenciadas pelas zonas vizinhas; e ao não apresentar uniformidade das relações que explicam as viagens em todas as partes do espaço considerado. Diante disso, o objetivo desta pesquisa é verificar a hipótese de que modelos de produção de viagens, que consideraram a dependência espacial, promovem resultados mais eficientes do que os que utilizam eventos dissociados da influência de zonas vizinhas. Para isto, desenvolveu-se uma comparação entre um modelo tradicional e um modelo baseado na dependência espacial, ambos elaborados a partir de dados de viagem e dados socioeconômicos da Região Metropolitana de Campinas. Com os resultados obtidos pode-se concluir que os modelos baseados em dependência espacial apresentam melhores resultados do que utilizam as técnicas tradicionais.

Palavras-chave: análise espacial, dependência espacial, autocorrelação espacial, modelos de geração de viagens, planejamento de transportes.

ABSTRACT

Linear regression techniques are the most used for the development of production trip models in traffic zones, which start from the premises of independence, linearity, and homoscedasticity of the response variable data. These premises are restrictive and rarely fully met in production trip models, especially since the explanatory variables (socioeconomic, demographic and land use characteristics) do not occur at random and depend on factors generally related to their location. According to Tobler's law, the values of spatial variables are attributed to the proximity between measures of that same variable in different sites, so there is a tendency that the values of one or more variables associated with a location are more similar to values observed in its neighborhood than to the rest of the sample set locations. Thus, it is possible to conclude that the classical mathematical procedures of linear regressions used to prepare production trip models violate mainly the assumptions of independence and homoscedasticity of the data, first of all, because it does not consider the trips produced are influenced by the neighbouring zones; secondly, by failing to show uniformity in the relationships that explain travel in all parts of the space under consideration. Therefore, the objective of this master dissertation is to verify the hypothesis that trip production models, which considered spatial dependence, promote more efficient results than those that use events dissociated from the influence of neighbouring areas. For this, a comparison between a traditional model and a spatial dependence model was developed, both models were elaborated with travel and socioeconomic data of the Campinas Metropolitan Region. With the results obtained it can be concluded that the models based on spatial dependence present better results than they use the traditional techniques.

Keywords: spatial analysis, spatial dependence, spatial autocorrelation, trip generation models, transport planning.

LISTA DE FIGURAS

Figura 1 – Modelo 4 Etapas 20
Figura 2 – Distribuições reais 24
Figura 3 – Distribuições do modelo 26
Figura 4 – Áreas 39
Figura 5 – Efeito da escala 45
Figura 6 – Efeito do zoneamento 46
Figura 7 - Representação esquemática dos valores do Índice de Moran 54
Figura 8 – Representação esquemática de valores do Índice de Geary 56
Figura 9 - *LISA Map* – Densidade Populacional de New York 69
Figura 10 - *Cluster Map* – Densidade Populacional de New York 71
Figura 11 – Fluxograma das etapas do desenvolvimento 74
Figura 12 - *Cluster Map* – Densidade Populacional de New York 83
Figura 13 – Distribuição espacial dos desvios 92
Figura 14 – *Box plot* das produções de viagem por pessoa por zona na RMC 96
Figura 15 – Configuração da Matriz de Vizinhança – software Geoda 97
Figura 16 – Matriz de vizinhança (arquivo de entrada) 102
Figura 17 – Diagramas de Espalhamento e Mapas temáticos – Quadrantes Q1 e Q3 104
Figura 18 – Diagramas de Espalhamento e Mapas temáticos – Quadrantes Q2 e Q4 105
Figura 19 – Indicadores de Moran Local para todas as zonas da RMC 109
Figura 20 – LISA *Map*– Valores da significância por zona 111
Figura 21 – Mapa RMC 113
Figura 22 – Mapa RMC 113

LISTA DE TABELAS

Tabela 1 – Modelos de produção da RMSP 22
Tabela 2 - Matriz de proximidade 39
Tabela 3 – Matrizes de pesos espaciais do ponto de vista teórico 41
Tabela 4 - Matrizes de pesos espaciais do ponto de vista topológico 42
Tabela 5 – Extrato do banco de dados tabulado (insumo para modelos e análise espacial) 84
Tabela 6 – Viagens classificadas por motivo 86
Tabela 7 – Correlações entre variáveis explicativas e viagens produzidas 87
Tabela 8 – Correlações Pearson e níveis de significância - variáveis explicativas 87
Tabela 9 – Coeficiente de Determinação do modelo 88
Tabela 10 – Teste F 89
Tabela 11 – Coeficientes das variáveis explicativas e teste t 90
Tabela 12 - Correlações entre variáveis explicativas e viagens produzidas 114
Tabela 13 - Correlações Pearson e níveis de significância - variáveis explicativas 115
Tabela 14 – Coeficiente de Determinação ($\boldsymbol{R2\ e\ R2ajustado}$) 116
Tabela 15 – Teste F 116
Tabela 16 - Coeficientes das variáveis explicativas e teste t 117
Tabela 17 – Comparações entre modelos 121

SUMÁRIO

1. OBJETIVOS 15

2. INTRODUÇÃO 15

3. MODELOS DE PREVISÃO DE DEMANDA 17

4. MODELOS DE PRODUÇÃO DE VIAGENS 20

4.1. QUALIDADE DOS MODELOS 23

4.2. PREMISSAS PARA A ELABORAÇÃO DO MODELO 24

5. ANÁLISE ESPACIAL 30

5.1. CONTEXTO HISTÓRICO 30

5.2. CONCEITOS 34

5.3. TÉCNICAS DE ANÁLISE ESPACIAL 35

5.3.1. ANÁLISE EXPLORATÓRIA DE DADOS ESPACIAIS (AEDE) 36

6. CARACTERÍSTICAS DOS DADOS ESPACIAIS 37

6.1. PROXIMIDADE ESPACIAL 37

6.2. ESTACIONARIEDADE 43

6.3. ESCALA ESPACIAL 43

6.4. PROBLEMA DE FRONTEIRA 44

6.5. PROBLEMA DA UNIDADE AÉREA MODIFICÁVEL 44

6.6. EFEITOS ESPACIAIS 46

6.6.1. HETEROGENEIDADE ESPACIAL 46

6.6.2. DEPENDÊNCIA ESPACIAL 47

7. *CLUSTERS* ESPACIAIS E AUTOCORRELAÇÃO ESPACIAL 50

7.1. INDICADOR DE MORAN GLOBAL 51

7.2. ÍNDICE DE GEARY GLOBAL 54

7.2.1. SIGNIFICÂNCIA ESTATÍSTICA 56

7.2.2. TESTE PARAMÉTRICO 57

7.2.3. TESTE NÃO PARAMÉTRICO 57

7.3. INDICADOR DE MORAN LOCAL 59

7.1. INDICADOR DE GETIS LOCAL 60

7.2. TÉCNICAS GRÁFICAS 63

7.2.1. DIAGRAMA DE ESPALHAMENTO 63

7.2.2. LISA MAP 67

7.2.3. CLUSTER MAP 69

8. METODOLOGIA 71

8.1. OBTENÇÃO E PREPARAÇÃO DOS INSUMOS 74

8.2. *CLUSTERS* ESPACIAIS 75

8.3. MODELOS DE PRODUÇÃO DE VIAGENS 78

8.3.1. VARIÁVEL RESPOSTA 79

8.3.2. VARIÁVEIS EXPLICATIVAS 79

8.3.3. PREMISSAS DAS REGRESSÕES 79

8.3.4. PARÂMETROS E TESTES DO MODELO 80

8.4. COMPARAÇÃO ENTRE MODELOS 80

9. DESENVOLVIMENTO 81

9.1. OBTENÇÃO E PREPARAÇÃO DOS INSUMOS 81

9.2. MODELO DE PRODUÇÃO CLÁSSICO 86

9.2.1. SELEÇÃO DAS VARIÁVEIS DO MODELO 86

9.2.2. ELABORAÇÃO DO MODELO CLÁSSICO 88

9.3. *CLUSTERS* ESPACIAIS 94

9.3.1. MATRIZ DE PROXIMIDADE ESPACIAL 95

9.3.2. INDICADOR DE MORAN GLOBAL 101

9.3.3. INDICADOR DE MORAN LOCAL 108

9.3.4. CLUSTERS ESPACIAIS 112

9.4. MODELO DE PRODUÇÃO COM *CLUSTERS* ESPACIAIS 114

9.4.1. SELEÇÃO DAS VARIÁVEIS DO MODELO 114

9.4.2. ELABORAÇÃO DO MODELO CLÁSSICO 115

10. COMPARAÇÃO DOS RESULTADOS DOS MODELOS 120

11. CONCLUSÕES E RECOMENDAÇÕES 123

REFERÊNCIAS BIBLIOGRÁFICAS 125

ANEXO 1: QUESTIONÁRIO DE DOMICÍLIOS 132

ANEXO 2: QUESTIONÁRIO SOCIOECONÔMICO 133

ANEXO 3: QUESTIONÁRIO DE VIAGENS 134

ANEXO 4: VIAGENS BASE DOMICILIAR MOTIVO TRABALHO HORA PICO MANHÃ 135

ANEXO 5: MÉTODO DOS MÍNIMOS QUADRADOS 136

ANEXO 6: COEFICIENTE DE DETERMINAÇÃO 137

1. OBJETIVOS

O objetivo desta pesquisa é verificar a hipótese de que modelos de produção de viagens, que consideraram a dependência espacial, promovem resultados mais eficientes do que os que utilizam eventos dissociados da influência de zonas vizinhas.

2. INTRODUÇÃO

O planejamento de transportes tem o objetivo de analisar alternativas de infraestrutura física, operação de sistemas e gerenciamento de demanda. Essas análises são realizadas tradicionalmente por meio de técnicas que procuram, em um primeiro momento, reproduzir as complexas relações entre viagens e características socioeconômicas, demográficas e de uso do solo em uma data inicial e posteriormente, reproduzir essas relações em diferentes anos para fazer previsões.

A principal técnica utilizada em estudos de previsão de demanda é denominada metodologia 4 Etapas, que contempla a elaboração de modelos de produção e atração de viagens em cada zona de tráfego, distribuição espacial entre zonas de origem e destino, divisão da demanda entre os diversos modos de transportes e, finalmente, a alocação de tráfego no sistema viário existente (BRUTON, 1975).

No que diz respeito a elaboração de modelos de produção de viagens, foco da presente dissertação de mestrado, as técnicas de regressões são as mais empregadas. Tais técnicas são baseadas em premissas matemáticas que tem a função de garantir que os modelos retratem o comportamento médio das relações entre as viagens e as variáveis que as explicam. Para tanto, preconizam que essas relações são independentes do espaço, ou seja, são idênticas em todo o território em análise, contrariando o fato de que as viagens não são eventos aleatórios,

dependem de sua localização no espaço e são influenciadas por áreas vizinhas (ORTUZAR e WILLUMSEN, 2011).

Essa característica de independência com relação ao espaço é dificilmente encontrada em análise de dados urbanos de grandes cidades ou regiões metropolitanas brasileiras, e modelos tradicionais baseados em comportamento médio das variáveis podem apresentar diferenças significativas entre os dados modelados e observados. Por outro lado, em qualquer estudo, as variáveis socioeconômicas e de uso do solo também não são completamente heterogêneas em todo seu território. Frequentemente as zonas se organizam em conjuntos homogêneos cujas viagens podem ser explicadas por variáveis próprias e distintas daquelas que representam as viagens do restante das zonas.

O uso de técnicas estatísticas de análise espacial pode auxiliar o desenvolvimento de modelos baseados em regressões lineares, minimizando os problemas intrínsecos do atendimento às premissas mencionadas anteriormente.

A análise espacial possibilita a identificação de conjuntos de zonas com comportamento homogêneo entre si e que se diferenciam do restante da área de estudo. O comportamento dos agrupamentos homogêneos pode ser reproduzido por meio de modelos específicos para esses conjuntos de dados homogêneos. Esse tipo de procedimento pode ser empregado para se obter as diferentes relações entre as viagens e os fatores que determinam a dependência da localização geográfica das zonas.

Nos capítulos a seguir são apresentados os conceitos e técnicas matemáticas utilizadas na elaboração de modelos de produções de viagens e suas limitações quanto ao atendimento às premissas de regressões. Em seguida são apresentadas técnicas alternativas pertencentes ao campo da análise espacial que

podem ser utilizadas para contribuir na melhoria dos resultados dos modelos de produção.

3. MODELOS DE PREVISÃO DE DEMANDA

Segundo Weiner, 2008a e Niemeier e Song, 2009, antes da década de 1950, a previsão de demanda era realizada baseado em dados do passado que eram extrapolados para o futuro utilizando-se fatores uniformes de crescimento. A premissa principal dessa técnica era que as relações entre os aspectos socioeconômicos e de uso do solo com as viagens eram consideradas inalteradas ao longo do tempo.

A partir dos anos 1950, grandes sistemas viários foram projetados para atender ao crescente consumo de automóveis e às significativas alterações do uso e ocupação do solo de importantes cidades americanas. Nesse período foram desenvolvidas as primeiras técnicas que pudessem prever a demanda e apoiar decisões sobre a implantação desses sistemas viários no sentido de diminuir os riscos de aplicação de recursos vultuosos de construção, operação e manutenção.

Segundo Weiner (2008b), a primeira técnica utilizada para previsão de demanda de transportes foi a técnica de geração de viagens por pessoa aplicada na cidade de San Juan, Porto Rico em 1948.

No início da década de 1950, desenvolveu-se um estudo de previsão de demanda para proposição de uma rede de vias expressas na cidade de Detroit. Esse estudo foi baseado nos diversos tipos de uso do solo daquela cidade e ficou conhecido como Detroit Metropolitan Area Traffic Study (DMATS).

Em 1955, desenvolveu-se um plano de transportes para a cidade de Chicago, onde foram aplicadas técnicas matemáticas que consideravam as produções e atrações de viagens nas diversas zonas da cidade, sua distribuição no

espaço urbano, além das preferências dos usuários pelos diversos modos de transportes e trajetos (MCNALLY, 2007).

As experiências de Chicago e Detroit foram replicadas em diversos planos de transportes no Reino Unido, tais como “London Traffic Survey (1960) e o “Highway Plan” (1962) (FURNISH E WIGNALL, 2009).

O aprimoramento dessas experiências iniciais serviu de base conceitual para o desenvolvimento de ferramentas matemáticas de previsão de demanda que configuram como uma das práticas mais importantes do processo de planejamento de transportes.

Tais ferramentas são conhecidas como modelos de previsão de demanda, e configuram representações simplificadas da realidade com foco em certos elementos considerados importantes para um determinado ponto de vista (ORTUZAR E WILLUMSEN, 2011).

Tais modelos têm o objetivo de estudar os impactos na demanda quando esta é submetida a cenários que combinam alternativas de infraestrutura, evolução temporal e espacial dos dados socioeconômicos, além de alterações dos padrões de ocupação do uso do solo (FURNISH E WIGNALL, 2009).

De acordo com Niemeier e Song (2009), os resultados esperados da aplicação de modelos de previsão de demanda são indicadores de performance e de benefícios a partir da implantação de alternativas de infraestrutura. Tais indicadores possibilitam subsidiar avaliações de viabilidade econômico-financeira de projetos, auxiliar nas tomadas de decisão sobre alternativas de tecnologia e infraestrutura de transportes e legitimar políticas públicas de transporte minimizando os custos e riscos antes de implementá-las.

Os modelos possibilitam ainda estimar o consumo de combustíveis, impactos econômicos relacionados aos níveis de emissões de poluentes na atmosfera e a quantidade de acidentes (FURNISH e WIGNALL, 2009).

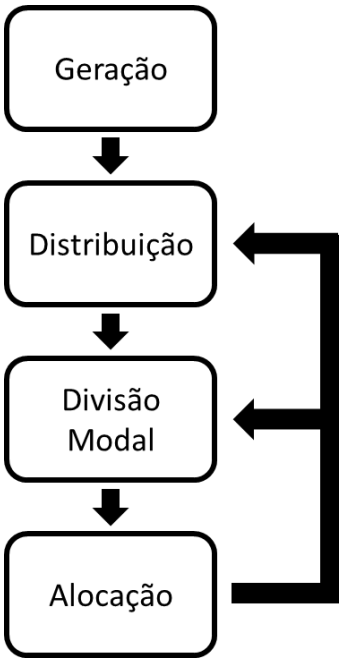
Um dos maiores benefícios da utilização de modelos é a possibilidade de testar complexas relações entre diversas atividades e o sistema de transportes antes de sua implementação e monitoramento.

A técnica que desponta como a mais utilizada na elaboração de estimativas de demanda de transportes é conhecida por Modelo de 4 etapas, que é constituído por submodelos sequenciais em que o resultado de um é o insumo do próximo, como apresentado na Figura 1 (ORTUZAR e WILLUMSEN, 2011).

O primeiro submodelo (primeira etapa), denomina-se modelo de geração e sua construção envolve o desenvolvimento de equações matemáticas que procuram representar os totais de produções e atrações de viagens em cada zona de tráfego da área de estudo na situação atual. O resultado é utilizado como insumo para a etapa seguinte, conhecida como modelo de distribuição, que procura inferir o arranjo espacial das produções e atrações na área de estudo, resultando em uma matriz de viagens entre zonas de tráfego.

Essa matriz é aproveitada na etapa seguinte, denominada modelo de divisão modal, que tem o objetivo de modelar as escolhas entre transporte coletivo e individual nas viagens consideradas. Finalmente na etapa de alocação de tráfego são utilizados algoritmos para prever as preferências de caminhos entre pares de origens e destinos (BRUTON, 1975).

Figura 1 – Modelo 4 Etapas



Fonte: Elaboração própria

4. MODELOS DE PRODUÇÃO DE VIAGENS

Os modelos de produção procuram reproduzir os valores mais prováveis dos totais de viagens originadas ou de retorno em zonas de tráfego a partir da relação com as características socioeconômicas, demográficas, geográficas e de uso do solo obtidas em pesquisas domiciliares de origem e destino (NIEMEIER E SONG, 2009 e ORTUZAR e WILLUMSEN, 2011; BOYCE e WILLIAMS, 2015).

A técnica mais utilizada na elaboração de desses modelos é a de regressões lineares, pelo método dos mínimos quadrados, que possibilita obter os parâmetros de uma equação de reta que minimizam os desvios entre dados observados e modelados (ver ANEXO 5).

A equação de reta que representa o modelo de produção de viagens resultante das regressões lineares é apresentada na equação 1, a seguir.

$$\hat{y}_i = \hat{b}_0 + \hat{b}x_i \pm \hat{\epsilon} \quad (1)$$

Em que:

$\hat{y}_i = variável\ resposta$ represetnativa das produções de viagem em cada zona i

$x_i = variável\ explicativa\ das\ viagens\ produzidas\ na\ zona\ i$

$\hat{\beta}_0 = intercepto\ no\ eixo\ y$

$\hat{\beta} = coeficiente\ da\ variável\ x_i$

$\hat{\epsilon} = desvios\ entre\ dados\ mdelados\ \hat{y}_i\ e\ observados\ y$

O **termo $\hat{y}_i$** é denominado variável resposta e refere-se aos totais de viagens produzidas previstas pelo modelo em cada zona de tráfego i e geralmente se referem ao período em que se deseja investigar, seja na hora de pico ou dia inteiro.

Geralmente os modelos são estratificados de acordo com as características particulares dos diferentes tipos de viagens que ocorrem no meio urbano. O primeiro nível de estratificação refere-se a viagens de base domiciliar, nas quais o domicílio encontra-se na origem ou o destino. Tipicamente essas viagens são por motivo trabalho, educação, saúde, compras, lazer, outros. O segundo nível de estratificação é definido de acordo com as viagens de base não domiciliar, tais como as que ocorrem entre locais de trabalho e de compras.

A **Erro! Fonte de referência não encontrada.** apresenta um exemplo de representação da estratificação de modelos de produção de viagem elaborados para a Região Metropolitana de São Paulo - RMSP de acordo com a base (domiciliar e não-domiciliar) e motivos.

Tabela 1 – Modelos de produção da RMSP

Base	Motivo	Estrato	Modelo de produção
Domiciliar	Trabalho	BDT	0,09754 população + 0,31512 frota + 6,24147 renda/100,000
Domiciliar	Educação	BDE	0,28174 frota - 0,04849 população + 2,39908 renda/100,000 + 0,21966 estudantes
Domiciliar	Outros	BDO	0,08293 frota + 0,01975 população
Não - Domiciliar	Todos	BND	0,02589 matrículas - 0,01459 população + 0,07162 empregros secundarios + 2,91852 renda/100,000

Fonte: SISTRAN ENGENHARIA (2016)

O **termo $\widehat{b_0}$,** também conhecido como constante linear do modelo, representa os desvios provocados pelas variáveis explicativas que afetam a magnitude da resposta e que não foram incluídas no modelo.

O **termo x_i** representa as variáveis explicativas encontradas em cada zona de tráfego de uma cidade ou região. A seleção das variáveis deve atender a algum critério teórico que as justifique sua presença no modelo. Segundo (Boyce e Williams, 2015) e (Ortuzar e Willumsen, 2011) as variáveis mais comumente recomendadas na literatura são: renda, totais de população, número de domicílios, número de pessoas empregadas por domicílios, renda média por domicílio, idade, sexo, número de automóveis por domicilio, frota, tamanho da família, valor do solo, densidade residencial, acessibilidade. A seleção dessas variáveis deve garantir que não apresentem elevada correlação r entre si (multicolinearidade). As variáveis são pouco relacionadas quando $r \leq 0{,}1$; medianamente relacionadas quando $0{,}1 \leq r \leq 0{,}4$ e fortemente relacionadas quando o $r \geq 0{,}5$. (AGRESTI e FINLAY, 2012).

Os **termos $\widehat{b_i}$** $(i = 1,2, \dots, n)$ são os coeficientes de cada variável, também conhecidos como coeficientes angulares da reta de regressão e representam o

incremento (ou decréscimo) da variável resposta y para cada unidade adicional da variável explicativa x (FÁVERO *et al.*, 2009). Esses coeficientes possibilitam verificar a influência de cada variável independente na previsão da variável resposta (BRUTON, 1975).

Os desvios $\hat{\epsilon}$ são definidos como as diferenças entre dados observados (valores reais obtidos de pesquisas) e os valores modelados (valores obtidos da aplicação do modelo de regressão) (KENNEDY, 2003).

4.1. Qualidade dos modelos

Segundo Andy (2009) e Marôco (2011), os parâmetros estimados pelo método dos mínimos quadrados devem ser submetidos a testes que garantam a generalização do modelo para o universo dos dados. Isto é feito a partir dos testes e apresentados a seguir e detalhados no ANEXO 6.

- Teste R^2: esse indicador compara a linha dos mínimos quadrados à média dos dados e dessa forma aponta o quanto da variação de y pode ser creditada ao modelo de regressão amostral. Observa-se que esse indicador não é recomendável para modelos multivariados, porque a incorporação de mais uma variável tende a aumentar o valor de R^2 mesmo que possua influência reduzida sobre a variável resposta.

- Teste $R^2\text{ajustado}$: tem o mesmo objetivo do teste R^2, porém é utilizado em sua substituição nos casos de modelos multivariados.

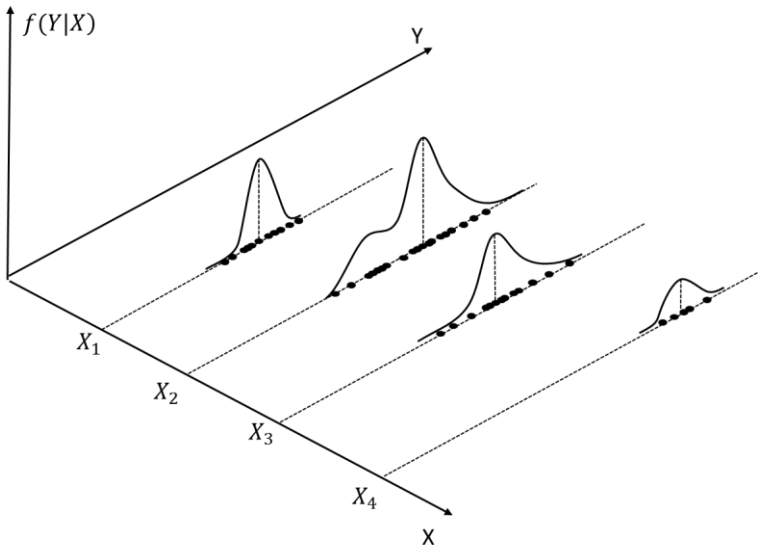
- Teste F: tem o objetivo de verificar a significância conjunta das variáveis explicativas, ou seja, se as variáveis explicativas contribuem para a elucidar a variável resposta.

- Teste t : possibilita verificar se as variáveis explicativas e o intercepto influenciam no comportamento da variável resposta. A escolha das variáveis do modelo deve ser apoiada por esse indicador.

4.2. Premissas para a elaboração do modelo

Os conjuntos de valores observados de y em cada valor de x, possibilitam construir funções de distribuição de probabilidade condicional $f_i(\, y_i \,|x_i)$. Essas funções respondem à pergunta: qual a probabilidade de se obter y a partir de um x conhecido? No entanto, se essas distribuições são diferentes entre si, como representado na Figura 2, a chance de se obter y em função de x é diferente a cada observação i , tornando impossível estabelecer um modelo linear para fazer estimativas de cada valor de y.

Figura 2 – Distribuições reais



Onde:

$y = variável\ resposta.$

$x = variável\ resposta$

$f_i(\, y_i \,|x_i) =\ distribuições\ de\ probabilidade\ dos\ valores\ de\ y_i\ para\ cada\ x_i$

$i = n - ésima\ observação$

Para possibilitar o uso de técnicas estatísticas na elaboração de um modelo de produção de viagens baseado em amostras, que represente as relações entre variáveis y e x no universo dos dados, é necessário atender às seguintes premissas simplificadoras sobre a regularidade das distribuições $f_i(\, y_i \,|x_i)$ (ORTUZAR e WILLUMSEN, 2011).

- **Linearidade:** a relação entre as variáveis resposta e a explicativa deve representar uma tendência da distribuição dos dados em forma linear (AGRESTI E FINLAY, 2012)**.**

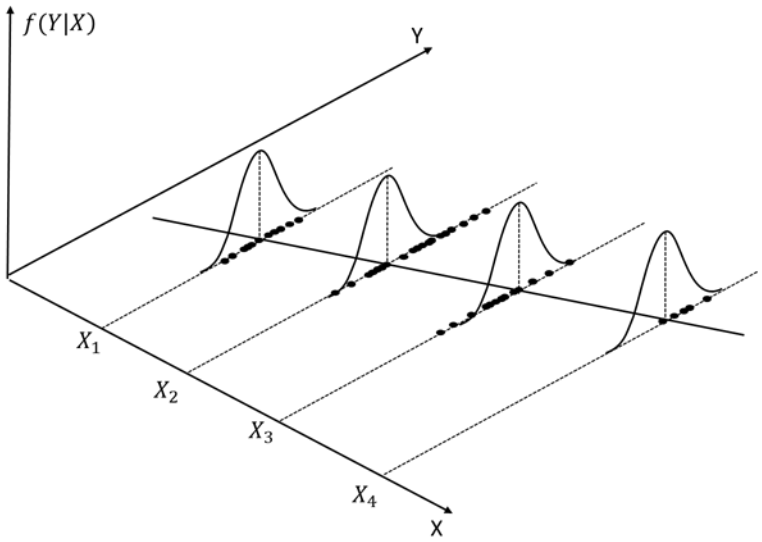
- **Independência dos desvios:** os valores previstos pela regressão linear não podem ser influenciados mutuamente, ou seja, um valor previsto pelo modelo não pode estar relacionado por qualquer outro valor previsto, (HAIR E JOSEPH, 2006).

- **Normalidade da distribuição dos desvios:** presume-se que o desvio ϵ é uma variável aleatória distribuída segundo uma curva normal.

- **Homocedasticidade dos desvios**: a premissa de homocedasticidade corresponde a hipótese de dispersão homogênea dos desvios ϵ.

A ideia central do atendimento a tais premissas é garantir que seja possível representar as relações que ocorrem no universo dos dados por uma linha reta, sendo que cada resultado esperado y_i corresponda a média de sua respectiva distribuição $f_i(\, y_i \,|x_i)$. Essas distribuições devem ser conhecidas em cada x_i e apresentar a mesma variabilidade em todas as amostras, como apresentado na Figura 3 (FÁVERO *et al.*, 2009).

Figura 3 – Distribuições do modelo



Onde:

$y = variável\ resposta.$

$x = variável\ resposta$

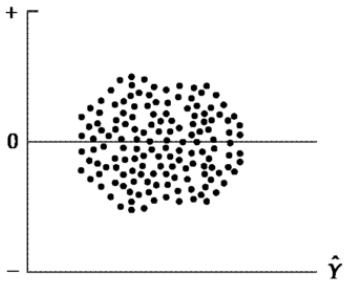
$f_i(\ y_i\ |x_i) =\ distribuições\ de\ probabilidade\ dos\ valores\ de\ y_i\ para\ cada\ x_i$

$i = n - ésima\ observação$

A principal medida para verificar se uma distribuição atende às premissas de regressões lineares é análise gráfica bidimensional onde os valores modelados de $\hat{y}_i$ são representados no eixo das abscissas e os valores dos desvios são representeados no eixo das ordenadas.

A situação ideal desse gráfico é apresentada no Gráfico 1, em que se observa um padrão aleatório da distribuição dos valores dos desvios da regressão, além de amplitude aproximadamente constante em todo o espectro de dados, padrão não curvilíneo e a presença de uma única nuvem de dados (HAIR, 2006).

Gráfico 1 - Aleatoriedade



Os demais gráficos apresentados a seguir (Gráfico 2, Gráfico 3, Gráfico 4 e Gráfico 5) são exemplos de distribuições de desvios, entre dados modelados e observados, que violam respectivamente as premissas de linearidade, independência, normalidade e homocedasticidade.

Gráfico 2 – Não linearidade

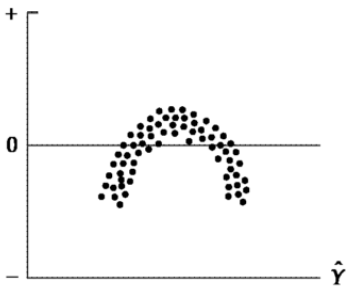


Gráfico 3 – Dependência

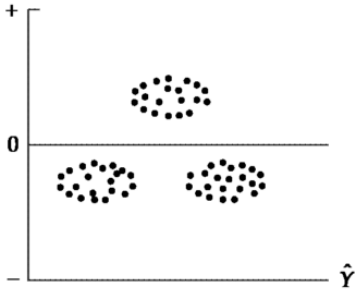


Gráfico 4 - Não normalidade

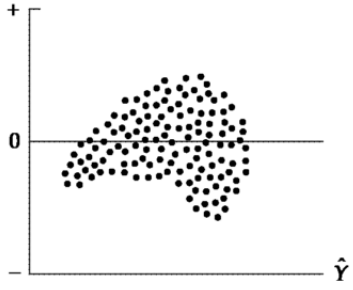
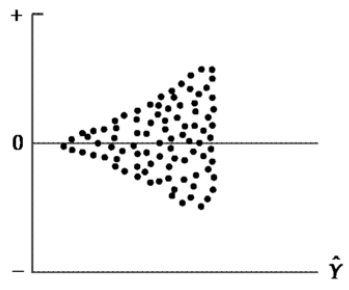


Gráfico 5 - Heterocedasticidade



FONTE (Hair, Joseph, 2006)

Algumas medidas simples podem ser incorporadas para solucionar problemas de não atendimentos das premissas, tais como: identificar e incorporar

variáveis explicativas que não fazem parte do modelo, identificar valores atípicos que podem causar alterações nos parâmetros da equação de reta, normalizar as variáveis para eliminar o problema do tamanho das zonas, além de transformações logarítmicas ou exponenciais das variáveis resposta e explicativas (ANDY, 2009).

Mesmo aplicando as medidas apontadas anteriormente, as premissas de regressões são restritivas e raramente atendidas completamente em modelos de produção de viagens, especialmente porque as variáveis explicativas (características socioeconômicas, demográficas e de uso do solo) não ocorrem de forma aleatória e dependem de fatores geralmente relacionados com sua localização.

Apesar dessas constatações, considera-se nesses modelos que os parâmetros $\widehat{\boldsymbol{b}}_{\boldsymbol{i}}$ de quaisquer variáveis explicativas são constantes, ou seja, significa afirmar que a lógica de produção de viagens é a mesma em todas as zonas da área de estudo, independentemente de sua localização espacial.

Um exemplo é o modelo de produção de viagens elaborado pela Companhia Paulista de Trens Metropolitanos – CPTM na Região Metropolitana de São Paulo. Esse modelo contava com apenas uma única equação linear para todas as 1895 zonas de tráfego, o que significa que as variáveis explicativas têm o mesmo nível de importância em qualquer parte da RMSP, ou seja, são independentes da dinâmica urbana particular de cada zona ou das discrepâncias socioeconômicas, demográficas e de uso do solo existentes.

A identificação de sub-regiões, cujas zonas se comportam de forma semelhante entre si e diferentemente do restante das demais, é uma prática recomendada por Fávero *et al.* (2009). Essa abordagem possibilita a elaboração de modelos de produção de viagens específicos para cada sub-região homogênea, com a vantagem de que seus dados não são influenciados por outras áreas que não fazem parte daquela sub-região.

Os gráficos a seguir são exemplos da ideia de elaboração de modelos específicos dos grupos de zonas homogêneas (*cluster* 1 e 2) formados segundo um critério de homogeneidade. O que se espera com esse tipo de abordagem é verificar se os modelos dos *cluster* 1 e 2 do Gráfico 7 são mais aderentes aos dados observados do que o modelo geral do Gráfico 6 e que os valores atípicos, que provocam resultados indesejáveis no modelo geral sejam transformados em valores que contribuem com os modelos de *cluster*s.

Gráfico 6 – Dados desagrupados

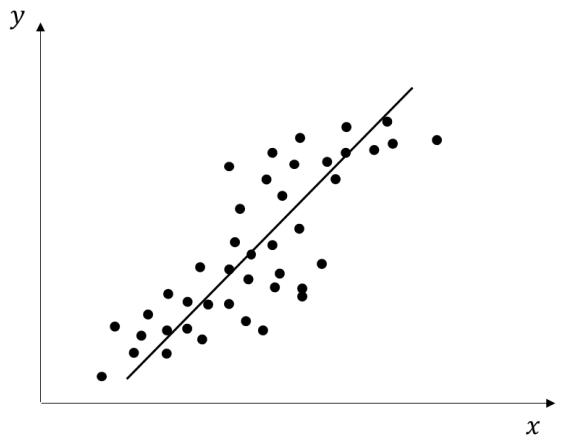
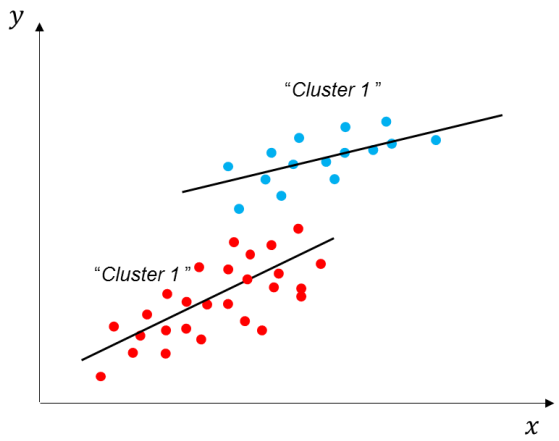


Gráfico 7 – Dados agrupados



A incorporação da localização como elemento formal dos modelos é um desafio que tem sido explorado nos últimos anos a partir do desenvolvimento de técnicas matemáticas de análise espacial em contraposição aos modelos tradicionais.

Nos capítulos a seguir serão apresentadas técnicas estatísticas desenvolvidas para determinar *clusters* de zonas homogêneas a partir de conceitos de vizinhança.

5. ANÁLISE ESPACIAL

5.1. Contexto histórico

As origens da Análise Espacial estão relacionadas com crescente interesse no estudo das iterações que regem a localização de atividades nos anos 1950 e também com o desenvolvimento da geografia quantitativa, que se caracterizou pelo rigor na aplicação da metodologia científica na verificação de hipóteses e, principalmente, pelo desenvolvimento de técnicas estatísticas que procuram reproduzir as relações entre as características de organização e distribuição espacial de qualquer variável de interesse (BURTON, 1963; LAMEGO, 2014).

A consideração do espaço como um fator influente nas análises de dados começou a ser estudado por Artu Robison, em 1956, quando estabeleceu pesos às observações de acordo sua influência, melhorando os resultados de suas análises de padrões espaciais. Na mesma época, pesquisadores da Universidade de Washington verificam a influência da localização relativa como o principal fator para compreender a natureza das atividades humanas (GETIS, 2008).

Os estudos de Moran (1948) e Geary (1954) foram pioneiros na elaboração de indicadores quantitativos do nível de associação entre variáveis considerando sua localização.

Do ponto de vista teórico, o estudo apresentado por Tobler (1970) resultou no principal conceito da análise espacial que estabelece a relação de dependência a partir da distância entre eventos no espaço

A partir dos anos 1970, com o advento dos Sistemas de Informações Geográficas - SIG, a Análise Espacial começou a ser aplicada em distintas áreas do conhecimento e foi nessa ocasião que surgiram subáreas da Análise Espacial com

diferentes abordagens. As principais subáreas são: estatística espacial, econometria espacial e geoestatística.

Atualmente, o diagnóstico de dados brutos e análise da aplicação de técnicas de estatística espacial são beneficiados pelo desenvolvimento de programas de computador específicos que incorporam ferramentas de estatística de dados espaciais à representação geográfica dos Sistemas de Informações Geográficas - SIG.

A aplicação das técnicas de análise espacial está difundida para além da geografia, seu campo original, e estende-se a um conjunto amplo de áreas do conhecimento, tais como: Urbanismo, Economia Espacial e Regional, Planejamento Urbano e Regional, Economia Urbana, Estudos de Análise Integrada de Uso do Solo e Transportes, Econometria, Ciências Sociais, Geografia, Ecologia, Epidemiologia, Sociologia, Geologia e Estudos Ambientais, Engenharia Civil, Estudos de Mercado Imobiliário (FRAME *et al*., 2000 e ANSELIN, 2010).

Segundo Mendonça (2008), a análise espacial consiste em entender como os dados decorrentes de fenômenos ocorridos no espaço se organizam e qual a relação entre eles. De acordo com Getis (2008), (“Spatial Autocorrelation”, [s.d.]) e Goodchild *et al*.(1992), atualmente, os principais usos da análise espacial são descritos a seguir:

- Identificação de conjuntos de zonas com padrões espaciais (*clusters* espaciais) a partir de um critério de semelhança.
- Determinação da extensão de efeitos espaciais em determinada variável.
- Encontrar possíveis relações dependentes que a realização de uma determinada variável pode ter em outras localidades.

- Identificação do papel do decaimento da distância ou interações espaciais em um modelo espacial.
- Verificação da influência que a geometria das unidades espaciais de análise pode ter sobre a variável.
- Compreensão dos processos responsáveis pela existência de padrões espaciais.
- Desenvolvimento de modelos de verificação de causa e efeito com a habilidade de prever e controlar eventos que ocorrem no espaço geográfico.

No Brasil, várias pesquisas incorporaram elementos de análise espacial no planejamento de transportes, podendo-se citar: Henrique e Loureiro (2004); Lopes (2005); Santos (2006); Mendonça (2008); Lopes *et al*. (2014) e Rocha (2014). A seguir são apresentados alguns dos exemplos.

Henrique e Loureiro (2004) utilizaram análise espacial para obter diagnósticos de acessibilidade e mobilidade na cidade de Fortaleza. Nesse trabalho os autores identificaram as ferramentas de análise estatística de visualização dos dados por meio de divisão os dados em intervalos iguais, por percentis e desvios padrão. Posteriormente utilizaram a média móvel para explorar a variabilidade dos dados no espaço. Finalizando, utilizaram indicadores globais e locais de dependência espacial, possibilitando identificar *clusters* espaciais por meio de testes de significância.

Lopes (2005) estudou os efeitos da dependência espacial nos modelos de previsão de demanda de transporte. Obteve os indicadores globais e locais de dependência espacial, identificou os *clusters* e *valores atípicos* espaciais e elaborou modelos para a Região Metropolitana de Porto Alegre, possibilitando testar a hipótese de que a introdução de indicadores de dependência espacial na

modelagem de demanda por transportes poderia produzir resultados mais confiáveis dos que os obtidos com modelos tradicionais.

Santos (2006) utilizou a análise espacial para avaliar acidentes de trânsito na cidade de São Carlos. Uma das premissas de seu trabalho é que o uso do solo exerce grande influência nos eventos ocorridos no sistema viário, dentre eles, os acidentes de tráfego. Assim como Lopes (2005), foram realizadas análises exploratórias de dados espaciais possibilitando identificar localizações atípicas ou *valores atípicos* espaciais, procurando descobrir padrões de associação espacial (*clusters*). O resultado desse trabalho possibilitou identificar padrões de ocorrência de acidentes em determinadas localizações da cidade.

Mendonça (2008), apresentou uma metodologia para previsão de demanda de passageiros do transporte rodoviário interestadual no Brasil a partir do desenvolvimento de modelos espaciais.

Braga *et al.* (2010), analisou a distribuição espacial do índice do produto interno bruto (IPIB) na mesorregião Metropolitana de Belo Horizonte (MMBH), no Estado de Minas Gerais, no ano de 2004, visando compreender a dependência ou semelhança entre os municípios dessa mesorregião.

Paiva e Khan (2011) utilizou análise espacial em seu trabalho para identificar se o setor industrial presente em um município influencia a formação do estoque de emprego formal no mesmo setor, mas em um município vizinho.

Souza (2013) obteve valores de produção e atração de viagens considerando-se a dependência espacial das variáveis normalmente atualizadas em modelos de previsão de demanda. O estudo de caso desse trabalho foi realizado sobre a base de dados da Região Metropolitana do Rio de Janeiro do ano de 2004.

Rocha (2014), apresentou uma estimativa da geração de viagens por transporte coletivo na Região Metropolitana de Salvador (RMS), por meio de ferramentas de análise espacial, especificamente, Geoestatística (Krigagem Ordinária e Krigagem com Deriva externa).

5.2. Conceitos

A Análise Espacial é definida como o conjunto de técnicas matemáticas que tem o objetivo de estudar a influência do espaço no comportamento de fenômenos naturais ou sociais (CARVALHO *et al*., 2004; LLOYD, 2007 e FISHCER e GETIS, 2010).

As técnicas de análise espacial diferem das técnicas tradicionais de análise de dados, sobretudo porque sua representação é realizada a partir de dois elementos fundamentais: a **localização** do objeto em análise e seus diversos **atributos**, enquanto que as técnicas tradicionais têm foco somente nos atributos (FISHCER e GETIS, 2010 e KUHN, 2016).

Segundo Anselin (1992); Carvalho *et al.*(2004); Fishcer e Getis (2010), a localização de objetos no espaço é realizada por meio de representações geométricas que podem ser de três tipos distintos: superfícies contínuas, pontos e também unidades territoriais de análise indivisíveis

- **Superfícies contínuas**: são estimadas a partir de um conjunto de amostras de campo, que podem estar regularmente ou irregularmente distribuídas. A variação contínua sobre o espaço está associada frequentemente com a distância. As superfícies contínuas são comumente relacionadas com fenômenos geológicos, climatológicos topográficos, ecológicos, fitogeográficos e pedológicos. Os principais estudos de geoestatística

envolvem a estimativa de valores em uma superfície a partir de interpolações.

- **Eventos ou padrões pontuais:** são identificados como pontos localizados no espaço com dimensão zero e relacionados a eventos discretos. Os fenômenos pontuais são expressos através de ocorrências, denominados processos pontuais. Nesse caso, o objeto de interesse é a própria localização espacial dos eventos em análise.

- **Áreas**: são representadas por polígonos bidimensionais fechados e possibilitam que valores de uma determinada variável localizada no espaço sejam agregados e sumarizados. As unidades territoriais de análise do tipo área geralmente estão associadas a problemas de contagens e taxas agregadas e podem ser, por exemplo, setores censitários, zonas de endereçamento postal, municípios, zonas de tráfego. O uso de áreas envolve a identificação de dependência espacial, padrões espaciais, elaboração de modelos econométricos (MARTIN, 1999).

5.3. Técnicas de análise espacial

Existem técnicas especificas a serem utilizadas em cada tipo de dado (superfície, pontos e áreas). As técnicas abordadas nessa dissertação estão relacionadas especificamente aos dados de áreas.

A Análise Espacial de Áreas contempla as seguintes técnicas de análise de dados espaciais: **seleção**, que contempla entradas e saídas de dados, consultas a banco de dados vinculados concomitantemente a mapas, possibilitando visualização de histogramas, estatísticas descritivas e diagramas de dispersão; **manipulação**, possibilitando criar novos dados espaciais, transformação de variáveis, destacando-se funções de agregação; desagregação, geração de áreas

de influência, sobreposição de camadas; **análise exploratória de dados espaciais**: indicadores de associação espacial, com inferência e visualização, além de identificação e apresentação de padrões espaciais.; **modelos espaciais**: que tem o objetivo de explicação de padrões espaciais, otimização, simulação, predição (LOPES, 2005; FISHCER e GETIS, 2010).

5.3.1. Análise exploratória de dados espaciais (AEDE)

A Análise exploratória de dados espaciais (AEDE) consiste em um conjunto de ferramentas gráficas e descritivas que possibilitam identificar propriedades espaciais dos dados, especialmente a dependência espacial, descrever e visualizar distribuições, identificar localizações espaciais atípicas (*outliers*), identificar padrões semelhantes de associação espacial (*clusters*) (KREMPI, 2004; BARONIO, VIANCO e RABANAL, 2012).

As técnicas de estatística clássica preveem que seu uso seja vinculado a premissa de aleatoriedade dos dados a serem analisados, portanto, não se recomenda seu uso em estudos de natureza espacial, haja vista que, por princípio, apresentam relações entre si que podem ser traduzidas em regras não aleatórias.

A AEDE procura detectar os padrões e sugerir hipóteses sobre possíveis relações entre as variáveis em análise. Dentre a coleção de técnicas incluídas na AEDE, destacam-se: Histogramas, Diagrama de Espalhamento Moran, *Box Plot*, LISA *Map* e C*luster Map,* indicadores globais de associação espacial: Moran e Geary e indicadores locais de associação espacial: Moran Local e Getis.Local.

6. CARACTERÍSTICAS DOS DADOS ESPACIAIS

6.1. Proximidade espacial

O conceito de proximidade espacial está relacionado com o nível de interdependência entre unidades de interesse (zonas, áreas, regiões, etc) motivado, em maior ou menor medida, pelas interações espaciais, sejam geográficas, sociais, econômicas, de acessibilidade entre áreas de uma cidade ou mesmo conectividade entre regiões (ANSELIN, 2015).

Há diversos tipos de medidas de proximidade, tais como, proximidade física, compartilhamento de fronteiras, distâncias euclidianas, distâncias rodoviárias, ou ainda podem ser representações de natureza econômica/social, como, por exemplo, podem ser consideradas “áreas próximas” aquelas que apresentam rendas médias semelhantes, ou aquelas que apresentam um determinada faixa de fluxo econômico (FISHCER e GETIS, 2010).

A representação formal da proximidade entre cada par de unidades de observação é realizada por meio de matrizes de vizinhança também denominadas matrizes espaciais de proximidade. Essas matrizes compreendem a força de associação espacial de cada localização i com relação a todos os outros lugares j (PÁEZ, 2005; FISHCER e GETIS, 2010).

As matrizes de proximidade W são elaboradas a partir dos elementos w_{ij} que correspondem a medida de proximidade entre as observações das áreas A_i e A_j. Valores diferentes de zero na matriz de proximidade refletem o potencial de interação entre observações. Elementos com valor zero indicam inexistência de interação entre observações. As medidas de proximidade podem representar vizinhos diretos (primeira ordem), vizinhos dos vizinhos (segunda ordem), e assim por diante (VITON, 2010).

A elaboração da matriz de proximidade espaciais deve estar relacionada com o tipo de estudo que se deseja realizar, porque trata-se de uma variável que integra a formulação matemática dos indicadores dos níveis de associação espacial entre áreas (assunto que abordado mais adiante nessa dissertação). Por isso, essa etapa deve ser respaldada por uma abordagem teórica que subsidie a definição de áreas vizinhas. (DARMOFAL, 2015).

Getis e Aldstadt (2009) elencam 4 recomendações para especificação de matrizes de proximidade:

1. Representar a geometria das áreas de tal forma que a quantidade de vizinhos fique na faixa maior que três e menor que sete.
2. Utilizar mais do que 60 unidades espaciais nas análises espaciais.
3. Adotar modelos espaciais simples. Modelos espaciais de primeira e segunda ordem devem ser preferenciais aos modelos de ordem superior.
4. Em geral, é preferível considerar uma quantidade menor de vizinhos.

No exemplo a seguir é apresentado um problema em que se deseja identificar o nível de associação espacial das cinco áreas apresentadas na Figura 4. Para tanto, o pesquisador precisa escolher um critério que defina os vizinhos de cada área.

Nesse exemplo as áreas vizinhas foram definidas como aquelas que compartilham fronteiras, e foram representadas na matriz de proximidade com o valor $w_{ij} = 1$ em contrapartida, os pares de zonas que não são vizinhas foram representadas pelo valor $w_{ij} = 0$, como representado na Tabela 2.

Figura 4 – Áreas

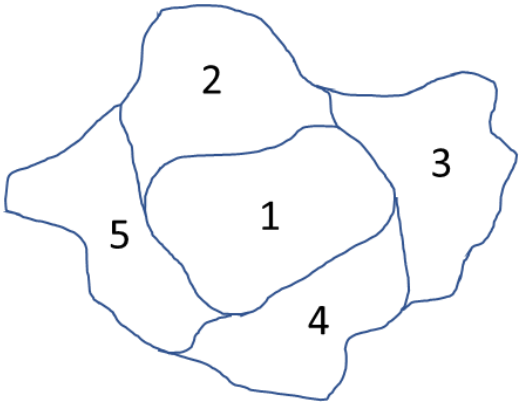


Tabela 2 - Matriz de proximidade

	A1	A2	A3	A4	A5
A1	0	1	1	1	1
A2	1	0	1	0	1
A3	1	1	0	1	0
A4	1	0	1	0	1
A5	1	1	0	1	0

Após a construção da matriz de proximidade, há a necessidade de transformá-la na forma estandardizada, evitando-se problemas de escala. A transformação mais comum é chamada estandardização de linha, em que são divididos os valores de cada elemento da matriz w_{ij} pelo valor da soma de sua respectiva linha, resultando em um valor denominado $\widetilde{w}_{ij}$.

Esse procedimento garante que a soma dos valores de cada linha da matriz de vizinhança resulte no valor igual a 1. A equação geral que contempla a transformação par estandardização é apresentada a seguir (PÁEZ, 2005).

$$\widetilde{w}_{ij} = \frac{w_{ij}}{\max(1, \sum w_{ij})} \tag{2}$$

A representação matemática da matriz de vizinhança estandarizada do exemplo anterior é apresentada na Matriz 1:

Matriz 1 – Representação estandarizada de vizinhança

$$\widetilde{W} = \begin{pmatrix} 0 & 0{,}25 & 0{,}25 & 0{,}25 & 0{,}25 \\ 0{,}33 & 0 & 0{,}33 & 0 & 0{,}33 \\ 0{,}33 & 0{,}33 & 0 & 0{,}33 & 0 \\ 0{,}33 & 0 & 0{,}33 & 0 & 0{,}33 \\ 0{,}33 & 0{,}33 & 0 & 0{,}33 & 0 \end{pmatrix}$$

Há inúmeras formas de expressar as matrizes de proximidade em estudos que consideram a autocorrelação espacial, porém Fishcer e Getis (2010) as organizou em três tipos, denominados: empírico, teórico e topológico, de acordo com as descrições e formulações matemáticas apresentadas a seguir.

Matrizes do ponto de vista descritivo: são estruturas mais flexíveis de representação da proximidade entre pares de zonas, porque o pesquisador pode apontar as características da área de estudo que acreditar serem importantes. Representa associações já existentes dentro do conjunto dos dados a serem analisados.

Matrizes do ponto de vista teórico: Usualmente essas matrizes são concebidas de forma exógena ao sistema estruturadas com base nos modelos de gravidade em que o efeito diminui em função do aumento da distância (Tabela 3).

A aplicação desse tipo de estrutura deve ser utilizada com cautela devido a rigidez da presunção de que o efeito da distância é o mesmo em todas as direções, além de não considerar a influência da topologia. Porém, são úteis em situações onde há dificuldade de se estabelecer uma teoria da estrutura espacial. Câmara *et al.* (2004); Silva (2006); Sweeney (2007); Getis e Aldstadt (2009), Areal, Balcombe e Tiffin (2012) apresentam alguns exemplos de pesos do tipo teórico.

Tabela 3 – Matrizes de pesos espaciais do ponto de vista teórico

Tipo	Formulação matemática e observações
Inverso da distância	$\frac{1}{dn}$ n = fator de calibração. Influência dos vizinhos diminui quanto maior for a distância d.
Inverso da distância	$w_{ij} = exp(-\frac{d_{ij}^2}{s^2})$ Utilizada em situações que apresentam baixa influência a partir de uma determinada distância.
Inverso da distância	$w_{ij} = \frac{1}{\lvert x_i - x_j \rvert}$ Utilizada em situações que apresentam baixa influência a partir de uma determinada distância.
Distância simples	$\widetilde{w}_{ij} = 1;$ se d_{ij}<d Utilizado em problemas e vizinhanças baseadas e distancias, se d_{ij} é a distância entre as zonas i e j, podemos definir e d é uma distância limite pré-definida.

Matrizes do ponto de vista topológico: são matrizes concebidas a partir da constituição geométrica da proximidade espacial. Getis *et al.* (2009), Silva (2006) e Mendonça (2008) apresentam alguns exemplos de pesos do tipo topológico (Tabela 4):

Tabela 4 - Matrizes de pesos espaciais do ponto de vista topológico

Tipo	Formulação matemática e observações
Proporção entre fronteiras compartilhadas e o perímetro	$w_{ij} = \frac{l_{ij}}{l_i}, \quad i \neq j$ $w_{ij} = 0, \quad i = j$ l_{ij} = comprimento da fronteira entre as áreas i e j. l_i = perímetro.
Vizinhança de contiguidade	$w_{ij} = 1,$ se as áreas i e j compartilham fronteiras. $w_{ij} = 0,$ se as áreas não compartilham fronteiras.
Vizinhança de fronteira	$\widetilde{w}_{ij} = \frac{b_{ij}^{\propto}}{d_{ij}^{\beta}}$ b_{ij} = extensão de fronteira da zona i compartilhada com a zona j. $\propto, \beta$ = parâmetros de calibração.
Vizinhança econômica	$w_{ijw} =$ quantidade de trocas comerciais entre áreas i e j e modo w.

6.2. Estacionariedade

De acordo com Lloyd (2007), a estacionariedade é uma propriedade de dados espaciais que se refere a processos que tem propriedades similares em todas as localizações da região de interesse.

Uma estrutura espacial é considerada estacionária quando a média do processo no espaço e a covariância entre os dados não variam sensivelmente de sub-região a sub-região na área de interesse, ou seja, se eles apresentarem um comportamento homogêneo na área de interesse Spatial [s.d.].

A estacionariedade espacial pode apresentar em duas propriedades, a primeira implica em homogeneidade em toda a área a ser estudada, a segunda implica em que os padrões sejam isotrópicos. Uma estrutura espacial é considerada isotrópica se, além de estacionária, a covariância entre os seus dados depender somente da distância entre os pontos e não da direção entre eles, não apresentando viés direcional dos dados.

6.3. Escala espacial

A escala espacial é uma propriedade que se refere aos diferentes níveis de detalhe de representação do mundo real que serão consideradas ou descartadas, a depender do tipo de análise que se deseja fazer.

Deve-se observar que processos que aparentam ser homogêneos em uma escala, podem ser heterogêneos em outra, portanto a escolha da escala a ser utilizada também é um fator crítico da análise de dados espaciais.

Um benefício da correta escolha da escala em estudos de variáveis espaciais é que seu resultado pode auxiliar a delimitar a formulação de hipóteses em análises espaciais sendo, portanto, uma definição fundamental para uma melhor

tomada de decisão e investigação científica. (LLOYD, 2007; OYANA E MARGAI, 2015).

6.4. Problema de fronteira

As associações espaciais entre as unidades territoriais de análise que pertencem a uma determinada área de estudo têm a tendência, de acordo com a primeira lei da Geografia, a serem mais intensas entre zonas vizinhas.

Portanto, existe a chance de que as zonas que se localizam nos limites geográficos da área de estudo estejam mais relacionadas com as zonas externas do que propriamente com as zonas internas (DARMOFAL, 2015; OYANA e MARGAI, 2015).

Essa disposição espacial pode produzir um problema de análise pela falta de dados das zonas externas à área de estudo, porém, Anselin (1988), destaca que esse problema pode ser menos importante nos casos de pesquisas com grandes quantidades de número de observações.

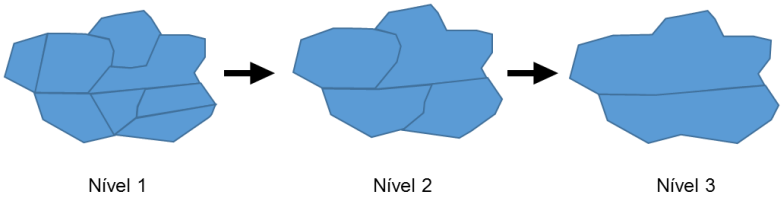
6.5. Problema da unidade aérea modificável

A forma como são dispostas as fronteiras das unidades territoriais de análise, propicia resultados diferentes da aferição de um mesmo fenômeno. A diversidade de interpretações de um mesmo fenômeno decorrente da disposição das fronteiras denomina-se problema da unidade aérea modificável e apresentam efeitos significativos no planejamento urbano e disciplinas correlatas, devido as diversas alternativas de subdivisão da cidade em unidades de análise menores. (CARVALHO ET AL., 2004 e PÁEZ, 2005).

Segundo e Oyana e Margai (2015), há dois efeitos relacionados com o problema da unidade aérea modificável a considerar durante o processo de análise espacial: efeito de escala e efeito de zoneamento.

O **efeito de escala** (**ou "falácia ecológica")** refere-se à tendência que as relações entre variáveis observadas em um determinado nível de agregação (setores censitários, zonas de tráfego, bairros, cidades) podem não ser as mesmas em outro nível de agregação (Figura 5). Portanto, o mesmo conjunto de dados pode resultar em diferentes resultados. Normalmente, quanto maior a unidade de agregação, maior, em média, a correlação entre duas variáveis (CARVALHO *et al.*, 2004).

Figura 5 – Efeito da escala

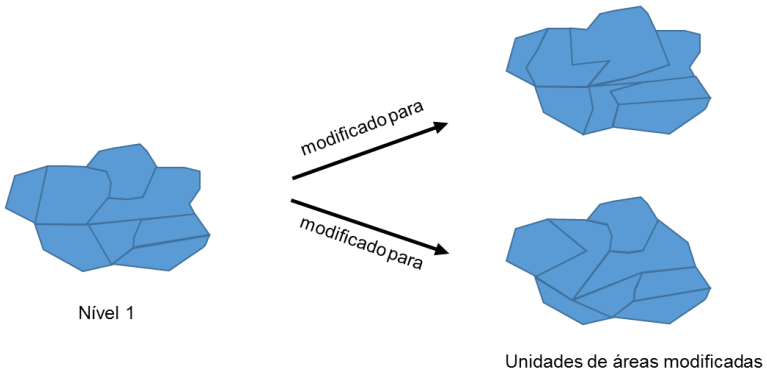


De acordo com Carvalho *et al.* (2004), a escolha da escala depende do problema a ser equacionado e não se pode afirmar que qualquer escala seja a "correta", mas apenas qual dos modelos serve melhor ao que se deseja esclarecer.

O **efeito de zoneamento** refere-se ao tamanho das unidades nas quais os dados são mapeados ou ao problema de partição dos dados espaciais (Figura 6).

Segundo Páez (2005), estudos em que são necessárias agregações em áreas, tais como análises urbanas, deparam-se com o fato de existem diferentes configurações geométricas para agregar áreas de uma determinada região, como apresentado na Figura 6, e cada uma delas pode resultar em diferentes distribuições dos dados em análise.

Figura 6 – Efeito do zoneamento



De acordo com Krempi (2004), a escolha das unidades de coleta de dados é parte importante da análise de dados distribuídos no espaço e devem estar em uma resolução com as menores dimensões possíveis sem prejuízo da quantidade e qualidade dos dados.

6.6. Efeitos espaciais

Para Geoffrey (2007), os efeitos espaciais podem ser divididos em dois tipos: dependência espacial e heterogeneidade espacial, descritos a seguir.

6.6.1. Heterogeneidade espacial

A heterogeneidade consiste na falta de uniformidade do efeito de dependência espacial e/ou das relações entre variáveis em estudo. Esse efeito decorre de particularidades intrínsecas de cada local em que a lógica que define a magnitude de uma determinada variável se altera em cada localização (ANSELIN, 1992; ANSELIN E GETIS, 2007; GEOFFREY, 2007).

Esse efeito pode ser causado por dois motivos, o primeiro refere-se a instabilidade estrutural da variável que pode ser verificada com técnicas econométricas tradicionais e o segundo está associado com a heterocedasticidade

proveniente de omissão de variáveis e erros de especificação (MORENO e VAYÁ, 2002) e (ANSELIN e GETIS, 2007).

6.6.2. Dependência espacial

O conceito mais importante da Análise Espacial foi apresentado por Tobler (1970) com a primeira lei da Geografia, cujo enunciado é descrito a seguir:

"Everything is related to everything else, but near things are more related than distant things"

A maior consequência dessa lei é que a variação que ocorre no espaço não é aleatória e surge devido a existência de uma relação funcional (regra) entre o que ocorre em um ponto determinado do espaço e o que ocorre em outro lugar (MORENO e VAYÁ, 2002).

A dependência espacial, também encontrada na literatura com os nomes de associação espacial, interação espacial, interdependência espacial, é definida como a tendência a que o valor de uma ou mais variáveis associadas a uma determinada localização assemelham-se mais aos valores observados em sua vizinhança do que ao restante das localizações do conjunto amostral (ANSELIN, 1988; HENRIQUE E LOUREIRO, 2004; GETIS, 2007, OYANA E MARGAI, 2015).

A representação matemática formal da dependência espacial pode ser expressa pela equação 3, em que uma observação y associada com uma localização identificada como i, depende das observações nas localidades $j \neq i$ (LESAGE, 1998B).

$$y_i = f(y_j), i = 1, \dots, n \qquad j \neq i \tag{3}$$

Segundo LeSage, 1998b e Paiva e Khan, 2011, existem dois motivos para se esperar que uma amostra de dados observados em um determinado ponto do espaço seja dependente de valores observados em outras localizações. O **primeiro** refere-se a **erros de medida e o segundo é a organização espacial.**

Os **erros de medida** ocorrem porque a delimitação geográfica das unidades espaciais selecionadas para coletar observações é arbitrária e, em alguns casos, pode não refletir exatamente o comportamento do fenômeno observado.

Por outro lado, a **organização espacial** refere-se ao conceito de que os fenômenos observados podem ter o espaço como elemento estruturador nas explicações sobre o comportamento humano e sobre as atividades econômicas. As ciências regionais, por exemplo, são baseadas na premissa que o custo do espaço e do transporte são elementos determinantes para a definição da localização da população e dos mercados (LESAGE, 1998).

Segundo Kelejian e Robinson (1992), Páez (2005), Anselin (2015) e Darmofal (2015), a organização espacial pode ocorrer a partir dos diferentes processos descritos a seguir.

- **Difusão:** adoção gradual de um novo atributo por uma população fixa. Normalmente, a probabilidade de adoção é influenciada pela distância. Por exemplo, a cultura de consumo das pessoas residentes em uma determinada região pode ser influenciada por seus vizinhos contíguos.

- **Transbordamento "Spillover"**: processo que ocorre quando um atributo em alguma localização é função não somente do que ocorre nos seus vizinhos contíguos, mas também da influência de áreas mais distantes, como por exemplo, preços de residências de outros bairros da cidade,

rendimentos salariais em uma cidade em função dos rendimentos das cidades vizinhas etc.

- **Interação espacial**: processo em que ações em uma determinada localização influenciam ações em outras áreas. Esse processo é geralmente resultado de competição, por exemplo, preços estabelecidos por empresas concorrentes em localizações diferentes. Embora este processo seja geralmente conceituado em termos de movimento físico de pessoas ou *commodities*, os fluxos de informações também podem estimular eventos em locais dispersos espacialmente.

- **Segmentação**: a partição de uma região anteriormente homogênea em duas ou mais sub-regiões, cada uma com características claramente únicas é um exemplo de segmentação. Em um contexto urbano, a segmentação espacial pode estar relacionada as economias de aglomeração, estratificação industrial, comercial e residencial e auto seleção racial ou social, entre outros processos.

- **Atribuição**: a variável em análise pode não ser o resultado de qualquer uma das situações anteriores, mas pode ser causalmente vinculada a outra variável. Neste caso, a estrutura espacial da variável dependente herda a estrutura espacial da variável independente à qual está relacionada. Por exemplo, os cidadãos de uma determinada zona podem escolher seus modos de transportes não porque interagem entre si, mas porque outros fatores, tais como renda, tempo de viagem, conforto, valor do tempo, determinam essa escolha.

7. *CLUSTERS* ESPACIAIS E AUTOCORRELAÇÃO ESPACIAL

A dependência espacial é determinada pelos processos de organização espacial mencionados anteriormente. Esses processos estabelecem uma lógica de constituição do espaço, ou seja, a forma como uma determinada variável se distribui no espaço.

Essa lógica de distribuição resulta em padrões espaciais que podem se estender para toda uma área de estudo ou se restringir a algumas subáreas que apresentam semelhanças entre si e diferenças com as demais.

Um dos principais objetivos da análise espacial é a obtenção de padrões espaciais de uma variável que decorrem da presença de dependência espacial. Esses padrões são denominados *clusters* espaciais e são definidos como grupos geograficamente delimitado de ocorrências cujos valores são similares com os valores de seus vizinhos (médias ponderadas dos valores vizinhos) e permitem afirmar serem improváveis de terem ocorrido por acaso (GETIS e ALDSTADT, 2009; ALDSTADT, 2010 e ORD e GETIS, 2012).

A identificação de *cluster*s está relacionada com a quantificação e confirmação da extensão da dependência espacial na área de estudo. Esse procedimento é realizado por meio de um conjunto indicadores denominados índices de autocorrelação espacial, que possibilitam verificar se a ocorrência de um evento em uma determinada localização é estimulada devido a ocorrência de um evento similar em uma localização vizinha ou se ocorrem independentemente do espaço. (KELEJIAN e ROBINSON, 1992 e LLOYD, 2007).

O termo autocorrelação espacial refere-se a correlação dentro de uma mesma variável em diferentes posições do espaço. Os indicadores de autocorrelação espacial são de dois tipos: **Indicadores Globais** e **Indicadores Locais** descritos a seguir (JACQUEZ, 2008).

Os **índices globais** são estatísticas que se referem à estrutura espacial do conjunto de zonas ou áreas de interesse e procuram verificar a existência de *clusters* (dependência espacial) em determinado conjunto de dados. No entanto, esses indicadores não contemplam a heterogeneidade espacial de subdivisões da área de estudo, portanto não possibilitam localizar geograficamente os *clusters* espaciais. Os principais indicadores globais são Moran e Geary (PAIVA, 2011; FORTIN e DALE, 2014).

Os **índices locais** produzem medidas realizadas em uma escala de maior detalhe, com foco em sub-regiões específicas dentro de uma área de estudo, possibilitando identificar a localização da dispersão espacial, aleatoriedade e *clusters* espaciais. Os principais indicadores locais são denominados: Índice de Moran Local e Índice de Getis (BRIGGS, 2016a).

A seguir são apresentados os indicadores globais e locais de autocorrelação espacial, salientando-se que sua interpretação deve considerar os conceitos dos capítulos anteriores referentes a influência da matriz de proximidade, estacionariedade, escala espacial, problema de fronteira e da unidade área modificável, mencionados nos capítulos anteriores.

7.1. Indicador de Moran Global

O índice de Moran é o indicador de autocorrelação global mais utilizado nas pesquisas de estudos espaciais. Esse indicador possibilita verificar a presença de dependência espacial no conjuntos dos dados analisados (ORD E GETIS, 1995 e POULIOU e ELLIOTT, 2009).

O teste de hipótese nula do índice de Moran é baseado na estrutura de covariância dos dados. A presença de dependência espacial implica em covariâncias diferentes de zero entre os valores de uma variável para localizações vizinhas.

Assim, se y_i e y_j são realizações de uma variável y indexada por localizações espaciais $i\ e\ j$, então há dependência espacial se a covariância entre y_i e y_j é diferente de zero, como representado na equação 4 a seguir (FISHCER E GETIS, 2010 e DARMOFAL, 2015).

$$Cov(y_i, y_j) = E(y_i, y_j) - E(y_i)E(y_j) \neq 0 \quad (4)$$

O indicador de Moran caracteriza-se essencialmente como um coeficiente de correlação produto-momento Pearson (coeficiente que mede a correlação entre duas variáveis diferentes) alterado para contemplar o efeito da variável em análise e da matriz de proximidade podendo ser mensurado em diferentes escalas, tais como, nominal, ordinal, intervalos e taxas (WALDHÖR, 1996; FISHCER E GETIS, 2010 e GRIFFITH, 2012).

O cálculo do Índice de Moran é realizado dividindo-se a covariância espacial pelo total da variação nas zonas vizinhas, ambos com relação à média de valores de y. A representação matemática desse índice é apresentada na equação 5 a seguir (PING *et al.*, 2004; OYANA E MARGAI, 2015; BRIGGS, 2016B).

$$I = \left(\frac{n}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}}\right) \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\,(y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (5)$$

Em que:

$I = índice\ de\ Moran$

$y_i = valor\ do\ atributo\ observado\ no\ local\ i$

$y_j = valor\ do\ atributo\ observado\ no\ local\ j$

$\bar{y} = valor\ médio\ do\ atributo\ em\ toda\ a\ área\ de\ estudo$

$w_{ij} = matriz\ de\ vizinhança\ considerada$

$n = quantidade\ de\ áreas$

A seguir é apresentada, a partir do trabalho de Farber (2013), uma explicação pormenorizada do significado de cada um dos componentes da equação de Moran

O primeiro termo da equação de Moran é o numerador descrito como a soma do produto cruzado $\sum_i \sum_j w_{ij} (y_i - \bar{y})(y_j - \bar{y})$ que representa o nível de interação espacial entre as medidas da variável y. A interação espacial é elevada quando os vizinhos se comportam de forma semelhante, ou seja, quando o valor de $(y_i - \bar{y})(y_j - \bar{y})$ for baixo. Os resultados possíveis do primeiro termo podem ser de dois tipos diferentes:

- Tipo 1: os valores de (y_i e y_j) são maiores ou menores do que a média. Nesse caso, o termo contribuirá positivamente na equação de Moran.

- Tipo 2: um dos valores y_i ou y_j é maior que a média e o outo valor y_i ou y_j é menor que a média $\bar{y}$. Nesse caso, o termo irá contribuir negativamente na equação de Moran.

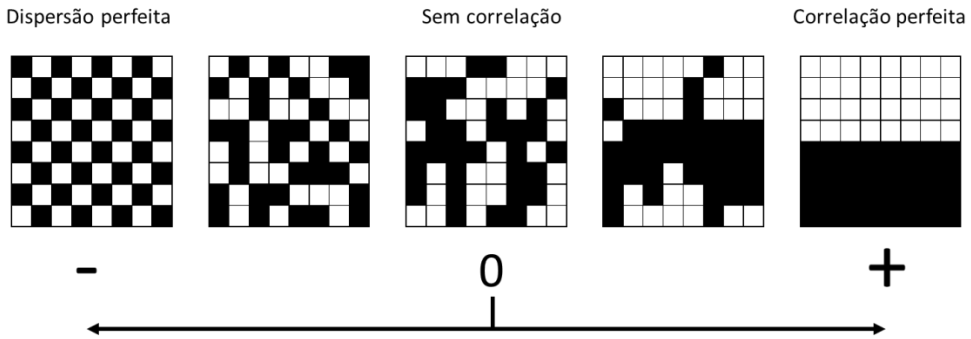
O cálculo do produto é realizado considerando somente os casos em que os pares de localidades i e j são vizinhos entre si. Para selecionar apenas localidades vizinhas é utilizada uma matriz de proximidade w_{ij} (ver capítulo de proximidade espacial), também denominada matriz de pesos ou matriz de vizinhança, que proporciona a medida expressa em valores contínuos ou discretos que indicam o nível de proximidade entre i e j.

Os resultados da soma do produto cruzado $\sum_i \sum_j w_{ij} (y_i - \bar{y})(y_j - \bar{y})$ dependem da escala de valores em que y é medido e, portanto, seus valores podem ser de elevada ou baixa magnitude dependendo da escala de medidas. Para

eliminar o problema de escala, o termo $\left(\frac{n}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}}\right)\left(\frac{1}{\sum_{i=1}^{n}(y_i-\bar{y})^2}\right)$ da equação de Moran tem a finalidade de estandarizar os resultados em uma faixa de valores entre -1 e +1.

A Figura 7 representa esquematicamente os níveis de associação espacial de acordo com os valores do índice Moran Global. Observa-se que os valores do índice de Moran próximos de 1 indicam forte nível de autocorrelação espacial positiva, conhecida como perfeita correlação. Nesse caso, valores atribuídos às áreas de análise adjacentes estão fortemente relacionados. Os valores do índice de Moran próximos de -1 indicam forte nível de negativa autocorrelação espacial, conhecida como perfeita dispersão. Os valores do índice de Moran próximos de 0 indicam a ausência de autocorrelação espacial (FARBER, 2013 e OYANA e MARGAI, 2015).

Figura 7 - Representação esquemática dos valores do Índice de Moran



Fonte: (FARBER, 2013).

7.2. Índice de Geary Global

Segundo Oyana e Margai (2015), outro indicador global de autocorrelação espacial é o índice de Geary (C). Esse indicador determina o grau de associação espacial usando a soma do quadrado da diferença entre os pares de valores de dados como sua medida de covariância.

Segundo Fishcer e Getis (2010), a hipótese nula do indicador de Geary é que as unidades espaciais de análise não diferem umas das outras, isso implica que a expectativa é que não há consistência na diferença entre vizinhos, ou seja, as vezes as diferenças são grandes e as vezes são pequenas. De maneira análoga ao índice de Moran, a matriz de pesos espaciais w_{ij} que procura representar a estrutura de vizinhança entre unidades espaciais de análise de maneira análoga ao indicador de Moran.

O índice de Geary é calculado pela expressão 6 a seguir (HUBERT, GOLLEDGE e COSTANZO, 1981; PING *et al*., 2004).

$$c = \left(\frac{n-1}{2\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}}\right) \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\,(y_i - y_j)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \qquad (6)$$

Em que:

$c = índice\ de\ Geary$

$y_i = valor\ do\ atributo\ considerado\ no\ local\ i$

$y_j = valor\ do\ atributo\ considerado\ no\ local\ j$

$n = quantidade\ de\ áreas$

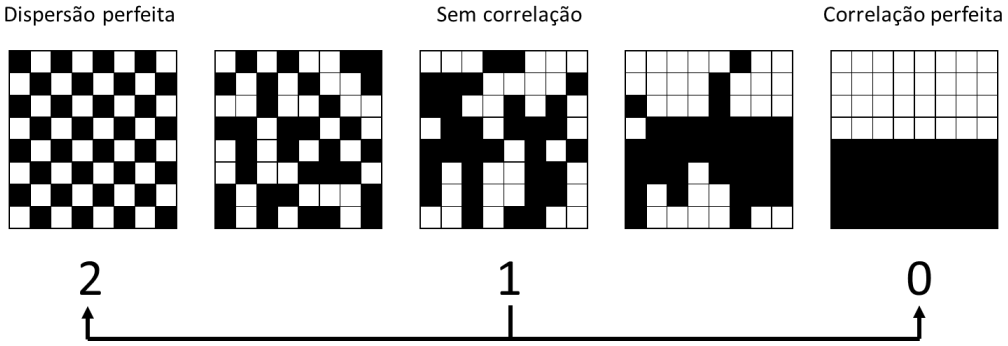
$w_{ij} = matriz\ de\ vizinhança\ considerada$

Observa-se que quanto maior for a semelhança entre os valores de y_i e y_j, menor será o numerador e, consequentemente, menor será o indicador de Geary, portanto valores baixos de c são indicativos da presença de padrões espaciais - *clusters* espaciais (ALDSTADT, 2010).

Os resultados do índice de Geary situam-se em uma faixa de valores entre 0 e 2, sendo que o valores próximos a zero indicam forte autocorrelação espacial, um valor igual a 1 significa um padrão espacial aleatório (ausência de autocorrelação) e uma valor entre 1 e 2 significa autocorrelação espacial negativa,

sendo o valor igual a 2 correspondente da dispersão perfeita (SILVA, 2006). A Figura 8 representa esquematicamente os níveis de associação espacial de acordo com os valores do índice de Geary mencionados anteriormente.

Figura 8 – Representação esquemática de valores do Índice de Geary



Fonte: (FARBER, 2013).

7.2.1. Significância estatística

Após a obtenção do índice de autocorrelação espacial, é necessário verificar se seu valor foi obtido ao acaso, ou seja, outras amostras não se comportam como a amostra analisada ou se realmente reflete a dependência espacial da área de estudo.

Essa verificação é realizada por meio de teste da hipótese nula de que os dados estão dispostos aleatoriamente no espaço, ou seja, que os dados observados não apresentam dependência espacial.

Para testar essa hipótese podem ser utilizados dois tipos de testes de significância estatística dos índices de autocorrelação espacial: o teste paramétrico e o teste não paramétrico. Em ambos os testes são utilizados valores $z - padronizados$ obtidos a partir da subtração do índice global de autocorrelação espacial de seu valor esperado e a divisão do resultado pelo desvio padrão

correspondente, como apresentado na expressão 7 a seguir (HENRIQUE e LOUREIRO, 2004; SILVA, 2006).

$$z = \frac{I - E(I)}{S_{erro(I)}} \tag{7}$$

Em que:

$I = Índice\ global\ de\ autocorelação\ espacial$

$E(I) = esperança\ de\ I$

$S_{erro(I)} = desvio\ padrão\ de\ I$

7.2.2. Teste paramétrico

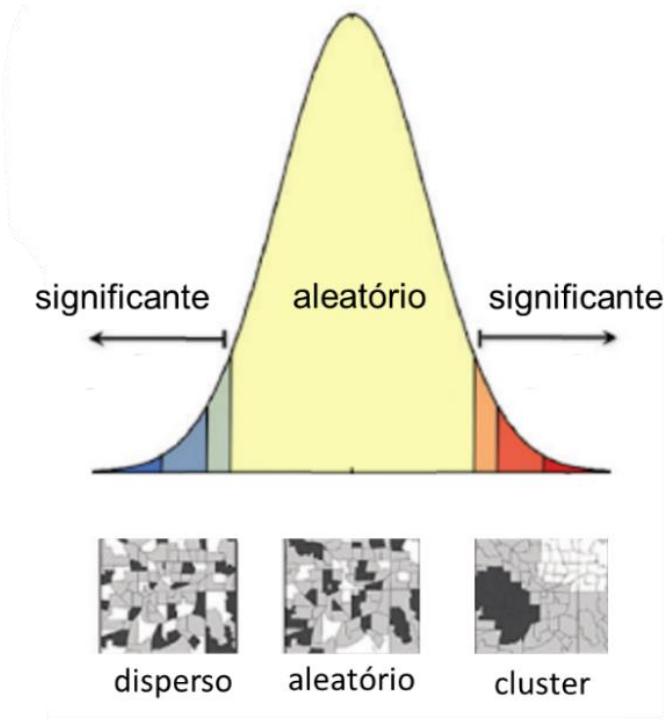
No teste paramétrico, assume-se que cada valor observado da variável em análise é independente e normalmente distribuído. Portanto, o procedimento utilizado para a verificação da significância é realizado comparando-se diretamente o resultado do índice de Moran calculado em um score padronizado z com uma curva de distribuição de probabilidade do tipo normal com $média = 0$ e $variância = 1$.

7.2.3. Teste não paramétrico

O uso do teste não paramétrico, também conhecido como teste de pseudo-significância, é recomendado como uma alternativa ao método paramétrico em situações em que os dados em análise são claramente não-normais (ANSELIN, 1992).

Neste método, o índice se de Moran calculado é comparado a uma distribuição empírica. Para elaborar essa distribuição, os valores originais da variável são diversas vezes permutados aleatoriamente no espaço, fixando-se o valor da zona i e alterando os demais valores entre as zonas restantes. O conjunto de todos os índices de Moran calculados em cada permutação possibilita gerar uma distribuição empírica como apresentado no Gráfico 8 (VASCONCELOS, 2016).

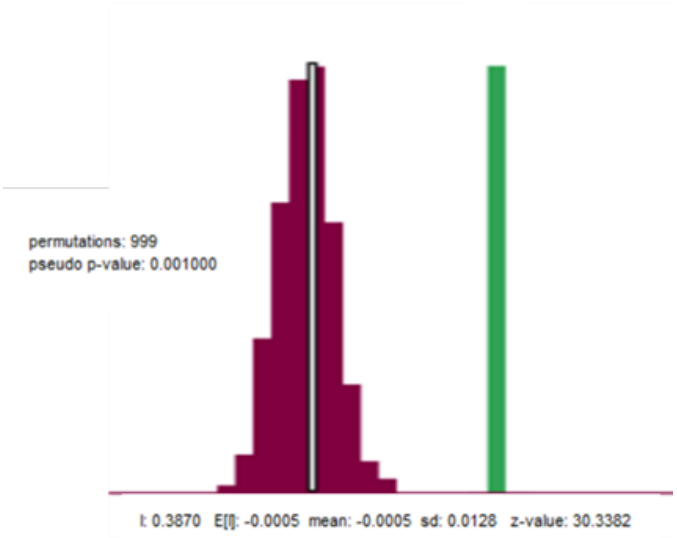
Gráfico 8 – Curva de distribuição empírica



Nessa distribuição, apenas um dos arranjos corresponde ao arranjo original. Caso o valor do índice de Moran original apresente significância estatística ($p = 0{,}05; p = 0{,}01; p = 0{,}001$), então, seu score z (valor padronizado) situa-se nas extremidades ou fora, da curva de distribuição empírica. Nessa situação há indícios confiáveis da presença de *cluster*s (extremidade direita) ou dispersão dos dados (extremidade esquerda). No caso de insignificância estatística, ou seja, aleatoriedade dos dados espaciais, o valor de z deve situar-se entre as extremidades e próximos a zero.

A Gráfico 9 representa um exemplo de distribuição empírica extraída dos dados de emprego da cidade de New York com a utilização do software Geoda. Nesse gráfico é possível observar um valor positivo do índice Moran (I = 0,387), o valor esperado desse mesmo índice na distribuição, o próprio desenho da distribuição, e finalmente o valor original do índice Moran no formato de *z-score* = 30,33.

Gráfico 9 - Distribuição empírica pessoas com emprego em NYC



7.3. Indicador de Moran Local

O indicador local de associação espacial - *Local Indicators of Spatial Association* (LISA), desenvolvido por Anselin (1995), é o mais utilizado em pesquisas científicas e caracteriza-se como uma decomposição do índice de Global de Moran.

Esse indicador produz um valor específico para cada área de análise fornecendo uma indicação de aglomerações significativas das áreas vizinhas (SILVA, 2006). Dessa forma, é possível verificar se o fenômeno se distribui em todo o espaço de forma estacionaria ou se existem bolsões de não-estacionariedade com características próprias diferentes do restante das zonas.

O cálculo desse índice é realizado a partir de desvios da variável e análise em relação à média por meio da expressão 8 (BRIGGS, 2016c).

$$I_i = \frac{(y_i - \bar{y})}{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2} \sum_{j=1}^{n} w_{ij}(y_i - y_j) \qquad (8)$$

Onde:

$I_i = Indicador\ Moran\ Local\ em\ i$

$y_i = valor\ do\ atributo\ observado\ no\ local\ i$

$y_j = valor\ do\ atributo\ observado\ nos\ locais\ j\ vizinhos\ a\ i$

$\bar{y} = valor\ médio\ do\ atributo\ em\ toda\ a\ área\ de\ estudo$

$w_{ij} = matriz\ de\ vizinhança\ considerada$

$n = valor\ amostral$

De acordo com Anselin (1995), analogamente ao indicador de Moran Global, as observações y_i e y_j são computadas em relação à média e a soma sobre cada j ocorre somente para as observações que apresentam vizinhança segundo algum critério adotado pelo pesquisador e representado pela matriz de proximidade.

Quando os valores estão inter-relacionados em padrões espaciais significativos, valores semelhantes (em desvios da média) são encontrados em locais vizinhos (ou seja, autocorrelação espacial positiva). Quando valores diferentes são encontrados em locais vizinhos, é dito que ocorre a associação espacial negativa. A associação zero implica um conjunto de observações espacialmente aleatórias (PÁEZ, 2005).

O indicador de Moran local serve de insumo para a elaboração de análises espaciais com diagrama de dispersão, *Box Map*, *LISA Map* e *Cluster Map* que serão abordados mais adiante.

7.1. Indicador de Getis Local

A estatística G_i foram desenvolvidas para detectar clusters espaciais significantes no nível local quando estatísticas globais não apresentam evidências de associação espacial (ANSELIN, 1995).

Esse indicador procura avaliar a associação espacial de uma variável dentro de uma distância especifica a um ponto. O seu cálculo prevê a soma de todos os valores de uma variável localizadas em um raio d a partir de um ponto i, determinado, no espaço, e a soma total da variável na região de estudo. Esse indicador por ser calculado incluindo ou não o valor de x_i.

A expressão 9 a seguir é utilizada para calcular a estatística G_i incluindo o valore de x_i.

$$G_i(d) = \frac{\sum_{j=1}^{n} w_{ij}\ (d)\ x_j}{\sum_{j=1}^{n} x_j} \tag{9}$$

A expressão 10 a seguir é utilizada para calcular a estatística G_i não incluindo o valore de x_i.

$$G_i(d) = \frac{\sum_{j\neq i}^{n} w_{ij}\ (d)\ x_j}{\sum_{j\neq i}^{n} x_j} \tag{10}$$

Onde:

$G_i(d) = índice$ Getis Local

$w_i = valor\ do\ atributo\ considerado\ no\ local\ i$

$x_j = valor\ da\ observação\ em\ \ j\ não\ normalizada$

$n = número\ de\ observações$

$w_{ij} = elemento\ ij\ da\ matriz\ binária\ W\ com\ 1\ para$
$todos\ os\ locais\ dentro\ de\ uma\ distancia\ d\ a\ partir$
$de\ i\ e\ zeros\ para\ todos\ os\ outros\ locais.$

A estatística $G_i(d)$ possibilita medir a concentração dos valores de uma determinada variável ao redor de uma dada localização. O resultado do cálculo desse indicador pode ser positivo, indicando que há *clusters* de valores altos em

torno do local i, enquanto que valores negativos de $G_i(d)$ indicam valores baixos (PÁEZ, 2005; CASTRO, SAWYER e SINGER, 2007).

Por ser baseada em uma distância de corte d, trata-se de uma estatística muito flexível de associação espacial que trabalham com variáveis positivas e naturais. Essa flexibilidade é limitada pela necessidade de impor uma distância para análise.

Esse indicador não apresenta uma estatística análoga para o cálculo de autocorrelação global e por isso é utilizada com menor frequência do que o índice *LISA*, porém a principal contribuição do índice Getis é que possibilita identificar *clusters* espaciais locais mesmo na ausência de dependência espacial global (DARMOFAL, 2015).

Supõe-se, para fins inferenciais, que esse índice possui distribuição normal. Isso é aceitável quando n (o número de pontos na amostra) é grande e a distância d na equação não é muito pequena (ou seja, não engloba observações dentro do raio) ou muito grande (isto é, como abrange todos as observações na área de estudo).

Se o valor (absoluto) da estatística padronizada for maior do que o de corte em um nível de significância pré-especificado, então a associação espacial positiva ou negativa existe. Os valores positivos da estatística são interpretados como uma aglomeração espacial de valores relativamente altos (mais do que seria esperado por acaso), enquanto os valores negativos representam valores relativamente baixos agrupados.

7.2. Técnicas gráficas

Os indicadores apresentados anteriormente representam medidas quantitativas do nível de autocorrelação espacial de uma determinada variável no espaço, esses valores devem ser plotados em mapas possibilitando a identificação visual de padrões dessa mesma variável para apoiar a interpretação do fenômeno em estudo. Para tanto são utilizadas técnicas gráficas, incluídas em softwares de análise espacial, tais como o Geoda, que possibilitam identificar as zonas com padrões espaciais detectados com indicadores de autocorrelação espacial, testar e identificar as zonas com níveis de significância que possibilitam concluir sobre a aleatoriedade ou organização espacial decorrente de processos espaciais, e finalmente, a identificação de clusters espaciais da variável em análise. Tais técnicas são denominadas de Diagrama de espalhamento, *LISA* Map, *Box Map* e *Cluster Map*. Essas técnicas são descritas de forma pormenorizada a seguir.

7.2.1. Diagrama de espalhamento

O diagrama de espalhamento é uma técnica de representação visual da autocorrelação espacial que tem o objetivo de verificar graficamente a variabilidade de uma determinado fenômeno no espaço, identificar a presença de instabilidades locais da dependência espacial, os locais que apresentam associação espacial, assim como sua a direção e magnitude (ANSELIN, 1995; MENDONÇA, 2008; FARBER, 2017).

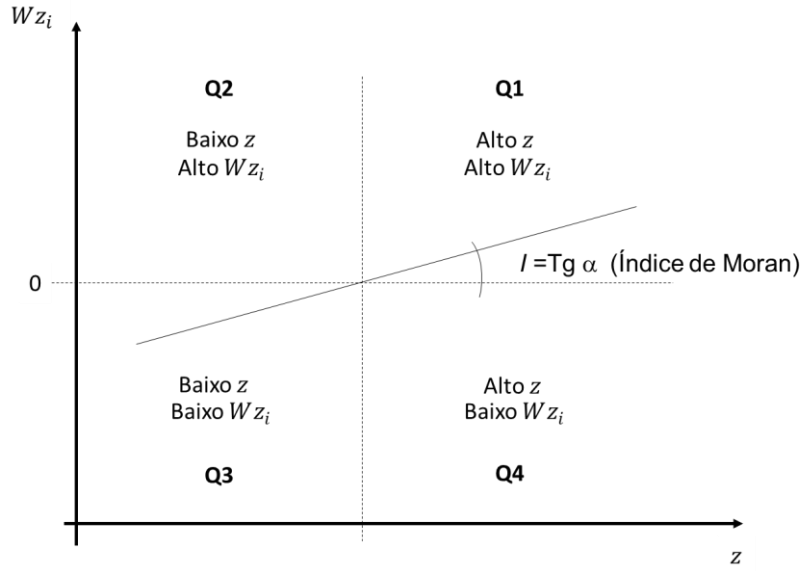
Esse diagrama auxilia na identificação visual de *outliers* e *clusters*. Os *outliers* são localizações que não seguem o mesmo processo de dependência espacial que a maioria das outras observações, enquanto que os *clusters* são identificados quando o valor de uma variável em uma determinada localização tem similaridade tão elevada com seus vizinhos que seria improvável que isso ocorresse de forma aleatória e sem influência do espaço.

A elaboração do diagrama de espalhamento consiste na montagem de gráfico bidimensional, cujos dados são distribuídos em quatro quadrantes. Os valores das abscissas são representados por uma variável y_i localizada no espaço e os valores das ordenadas Wy correspondem a média da mesma variável y obtida de seus vizinhos. Ambos os valores são normalizados em z-*scores* para facilitar a interpretação e categorização do tipo de autocorrelação espacial em *clusters* ou *outliers*. Essas novas variáveis são denominadas respectivamente Wz e z_i.

A análise do diagrama de espalhamento é realizada a partir da identificação do quadrante a que pertence cada ponto do gráfico. Esses quadrantes representam diferentes tipos de associação espacial.

Os pontos encontrados no quadrante Q1 indicam que, em média, para altos valores de z, existem altos valores de Wz_i, e serão chamados de Alto-Alto (High - High). Os pontos do quadrante Q3 indicam que em média, para baixos valores de z, existem baixos valores de Wz_i, e serão chamados de Baixo-Baixo (Low - Low). Os pontos localizados em Q2 e Q4 apontam que, em média, para baixos ou altos valores de z, existem respectivamente, altos ou baixos valores de Wz_i. Os pontos em Q2 de baixo-alto (Low – High) e os pontos em Q4 são chamados de alto-baixo (High - Low). (ANSELIN, 1995).

Gráfico 10 – Quadrantes do Diagrama de Espalhamento Moran



A dependência espacial apresenta maior intensidade à medida que mais pontos se aglomeram no primeiro e terceiro quadrante (Q1 e Q3), formando conjuntos espaciais de dados que se espalham de forma sistemática no espaço. Nesses casos o índice de Moran é positivo, como pode ser observado no Gráfico 11.

Em contraposição, a presença de pontos no segundo e quarto quadrante (Q2 e Q4) indicam a dispersão dos dados no espaço e inexistência de agrupamentos de dados semelhantes, nos quais o fator “espaço’ não influencia diretamente na valoração da variável z. Nesses casos ao índice de Moran é negativo, como observado no Gráfico 12.

Os casos em que os dados se encontram dispersos entre os diferentes quadrantes do diagrama de espalhamento e não seguem uma tendência geral, indicam que o fenômeno em análise não segue padrões e, portanto, é denominado não estacionário. O valor do índice Moran nesses casos é próximo ou igual a zero, como observado no Gráfico 13 (ANSELIN, 1996).

Gráfico 11 – Índice Moran > 0

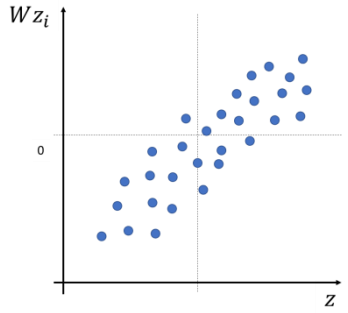


Gráfico 12 – Índice Moran > 0

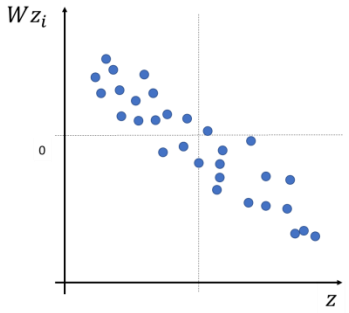
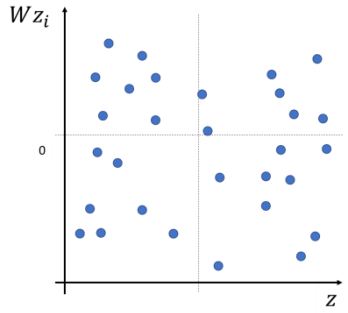
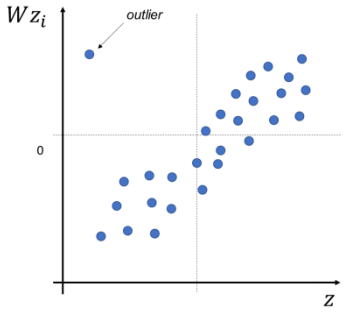


Gráfico 13 – Índice Moran = 0



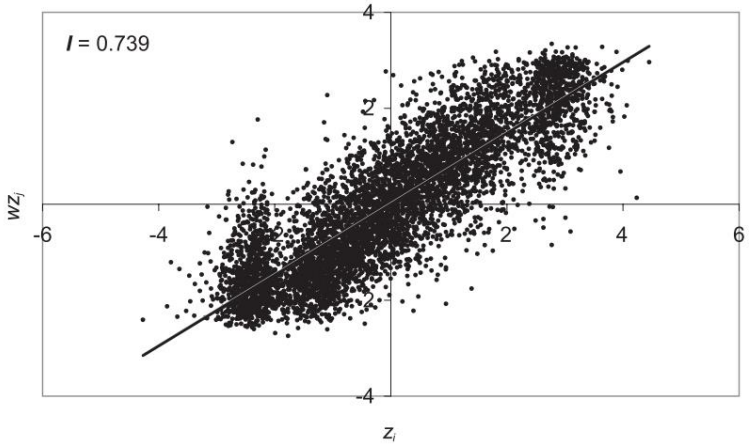
Os dados que apresentam pontos extremos em diagramas de espalhamento, como apresentado no exemplo do Gráfico 14 podem representar apenas um processo que não segue o mesmo processo de dependência espacial do restante dos dados observados e são considerados bolsões de não estacionariedade, especialmente se correspondem a locais espacialmente contínuos ou de fronteira. A presença de *outliers* pode ainda ser causada por problemas de especificação dos pesos espaciais ou devido a escala utilizada na obtenção dos dados (ANSELIN, 1996).

Gráfico 14 - *Outliers*



O Gráfico 15 representa um exemplo de diagrama de espalhamento, onde são observados uma grande quantidade de valores plotados, possibilitando identificar uma tendência predominante entre os quadrantes Q1 e Q3, representando assim, uma tendência de Índice de Moran positivo.

Gráfico 15 - Exemplo de Diagrama de Espalhamento



Quanto menor for a variação dos valores da nuvem de dados, maior será a magnitude do índice e Moran, que corresponde ao valor da inclinação de uma linha de regressão ajustada aos pontos de dispersão (ANSELIN, 2003).

7.2.2. LISA Map

O *LISA Map* é um mapa georreferenciado que possibilita identificar as zonas com correlação local significativamente diferentes do restante dos dados. Nesse tipo de mapa os valores do Indicador Local de Moran são classificados como não significantes ou com significância de 95%, 99%, 99,9%, que correspondem respectivamente a: 1,96 desvios padrões, 2,54 desvios padrões e 3,2 desvios padrões (SANTOS e SOUZA, 2007).

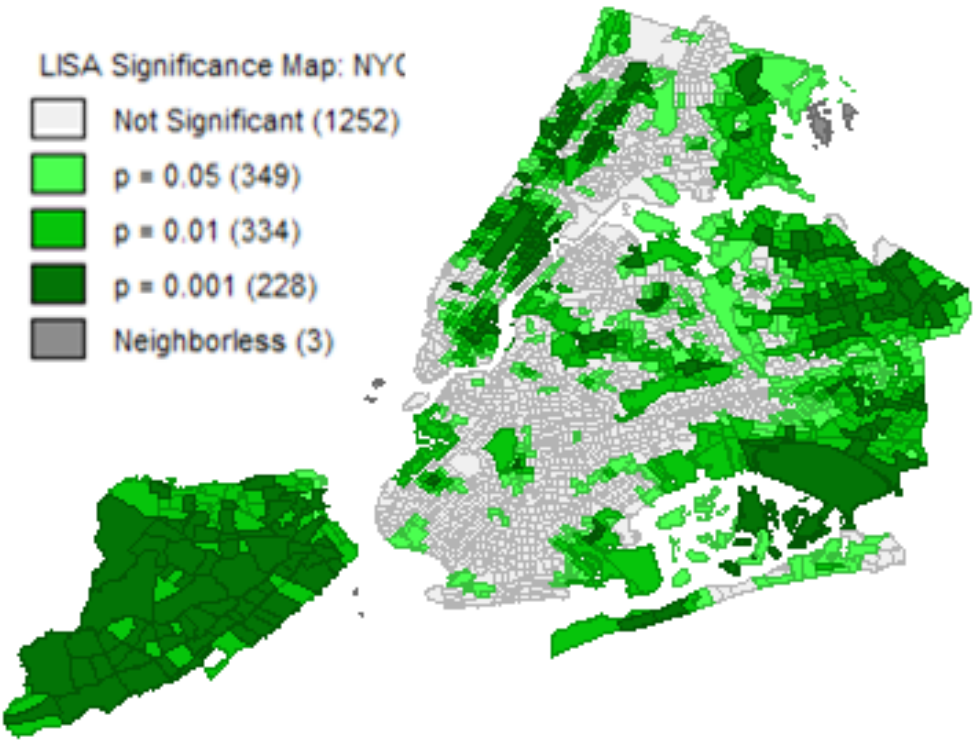
Da mesma forma como foram feitos os testes de significância dos índices globais de autocorrelação, a significância do Indicador Local de Moran deve ser avaliada utilizando-se a hipótese de normalidade ou a simulação de distribuição por meio de permutações aleatórias dos valores da variável viagens produzidas por pessoas, ou seja, os valores do indicador de autocorrelação local original devem ser comparados com uma distribuição estatística formada por valores desse mesmo indicador calculado especificamente a cada permutação espacial dos valores da

variável viagens produzidas. Esse procedimento é realizado com auxílio de softwares de análise espacial como o Geoda que possibilita ao pesquisador, a escolha da quantidade de permutações que deve ser adotada.

As áreas com maior probabilidade e existência de *clusters* espaciais são aquelas que apresentam valores nível de significância $p < 0{,}05$. Os softwares de análise espacial geralmente fornecem do os resultados do *LISA Map* em quatro níveis de significância: $p < 0{,}05$, $p < 0{,}01$, $p < 0{,}001$, $p < 0{,}0001$, além disso, possibilita identificar ás áreas não significantes e os *outliers* (LOPES, 2005).

Segundo (Anselin, 2005), os resultados podem ser sensivelmente diferentes dependendo da quantidade de permutações selecionadas. O exemplo da Figura 9 representa um *LISA Map* da densidade populacional da Região Metropolitana de New York, em que foram utilizadas 10000 permutações com auxílio do software Geoda, possibilitando identificar os diferentes níveis de significância em cada área. Neste gráfico é possível visualizar conjuntos de áreas significantes ($p < 0{,}05$), ou seja, com indícios de que se comportam como *clusters* espaciais.

Figura 9 - *LISA Map* – Densidade Populacional de New York



7.2.3. Cluster Map

O *Cluster Map* é outro mapa temático que aponta os grupos de zonas que apresentam dependência espacial entre si e cujos valores e teste de significância garantem serem gerados por processos espaciais com pequena chance de serem aleatórios. Um conjunto de zonas com dependência espacial pode ser visualizado no *Cluster Map*, assim como os padrões formados por essas zonas.

O *Cluster Map* é elaborado a partir do cruzamento dos do *LISA Map* com o Diagrama de Espalhamento. Os *clusters* correspondem às zonas com valores significativos obtidos no *LISA Map* identificados nos quatro quadrantes do Diagrama de Espalhamento. Os dados cruzados do *LISA Map* com o Diagrama de Espalhamento podem ser classificados da seguinte forma:

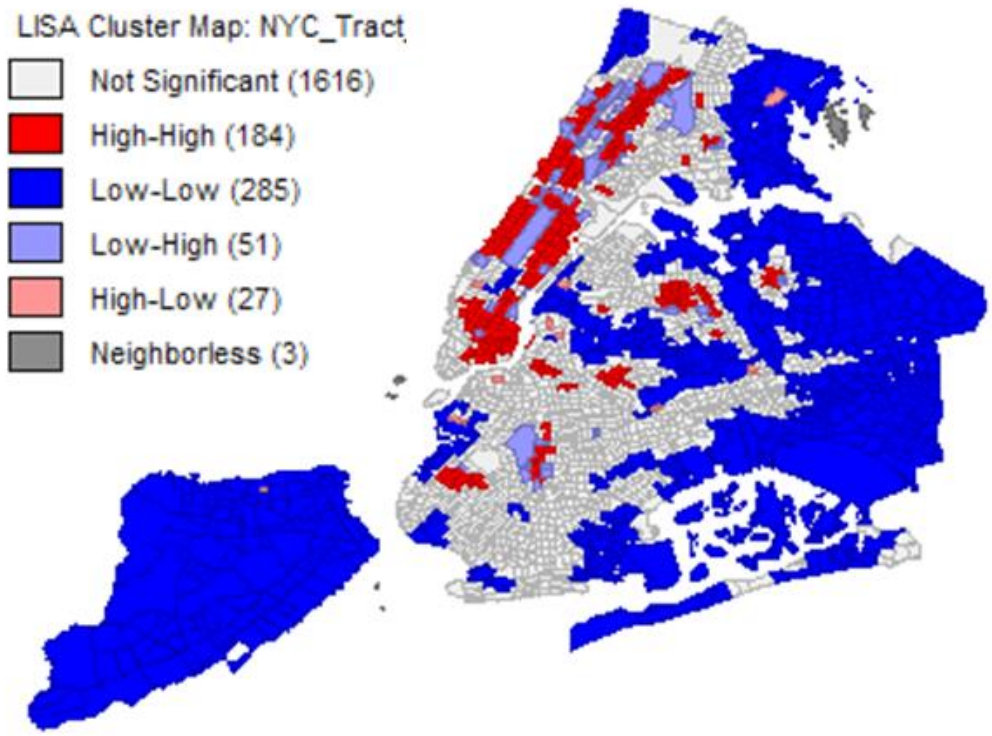
1. *Clusters* Espaciais*:* Valores significativos identificados nos quadrantes Q1 e Q3 do Diagrama de Espalhamento. Sendo que os valores do Q1 correspondem a zonas em que os valores dos Indicadores de Moran Local são altos em cada zona e altos nas médias de seus vizinhos. No caso do Q3, os valores dos Indicadores de Moran Local são baixos em cada zona e baixos nas médias de seus vizinhos.

2. *Outliers* Espaciais: Valores significativos dos quadrantes Q2 e Q4 do Diagrama de Espalhamento. Sendo que as zonas do Q2 correspondem apresentam valores dos Indicadores de Moran Local são baixos em cada zona e altos nas médias de seus vizinhos. No caso do Q4, os valores dos Indicadores de Moran Local são altos em cada zona e baixos nas médias de seus vizinhos.

3. Zonas sem significância: são considerados locais com comportamento aleatório, cujos valores não apresentam influência do espaço.

A Figura 10 apresenta a mesma região anteriormente analisada no *LISA* Map, porém com um exemplo de um *Cluster Map* onde foram identificadas as zonas que formam os *clusters* espaciais em vermelho e azul.

Figura 10 - *Cluster Map* – Densidade Populacional de New York



8. METODOLOGIA

A metodologia apresentada a seguir foi desenvolvida em consonância com os objetivos da presente pesquisa em verificar se *clusters* de zonas homogêneas podem subsidiar a elaboração de modelos de previsão de viagens mais robustos do que aqueles desenvolvidos a partir de metodologias tradicionais.

Inicialmente serão **obtidos os insumos** necessários para a elaboração de modelos de produção de viagens e para a identificação de clusters de zonas com dependência espacial. Esses insumos devem contemplar os dados disponíveis de viagens e socioeconômicos mais recentes na RMC.

Em seguida será verificada a tendência pela qual a variável produções de viagem se espalha no espaço. Essa verificação possibilitará concluir se as produções de viagem seguem uma tendência a se espalhar como uma mancha contínua entre vizinhos de fronteia física ou por outro tipo de organização espacial.

A verificação da tendência de espalhamento das produções de viagem e a decisão sobre a matriz de proximidade mais adequada, serão realizadas por meio de ferramentas de análise exploratória espacial, tais como, indicadores de autocorrelação espacial global e local, testes de significância estatísticas, diagramas de espalhamento, comparações com as médias de vizinhos, mapas LISA e combinações dessas ferramentas de análise espacial para identificar padrões espaciais.

A definição da matriz de proximidade dará subsídios para a nova fase do trabalho que se refere ao cálculo dos indicadores de autocorrelação espacial no nível global e local da RMC. Esses indicadores deverão contemplar todo o espectro das 128 zonas de tráfego e servirão como base para a conclusão sobre a existência da dependência espacial nas zonas de tráfego da RMC. No final dessa fase será calculada a significância estatística do indicador para confirmar se a **autocorrelação espacial no nível global e local** apresentam valores diferentes aos esperados, caso as viagens produzidas por pessoa por zona ocorressem de forma aleatória no espaço da RMC.

Os valores de autocorrelação espacial global e local serão submetidos a **testes de significância** para garantir que tais zonas apresentam produções de viagens que seguem uma determinada lógica de localização espacial e que, portanto, não se caracterizam por um espalhamento espacial aleatório. O resultado das significâncias estatísticas em cada zona de tráfego será apresentado por meio de um mapa denominado LISA MAP, em que são atribuídos os níveis de p-valor para cada zona.

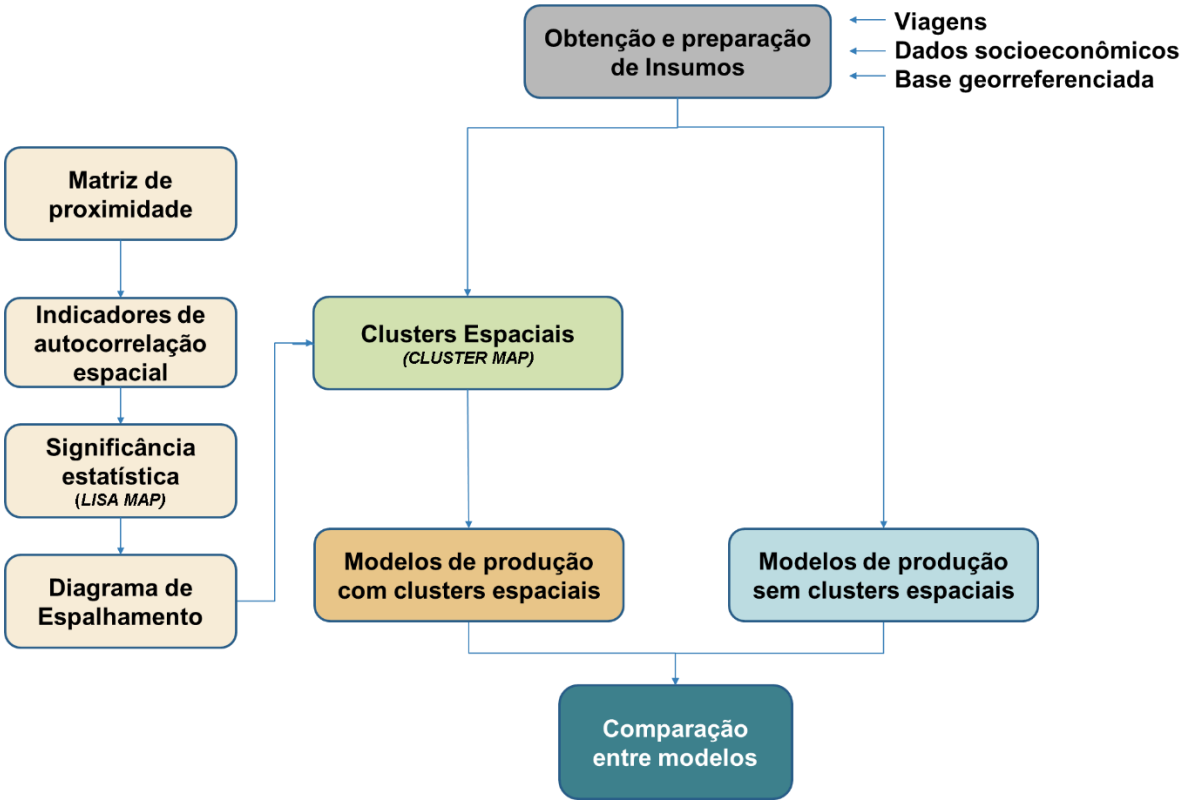
Em seguida, os resultados obtidos do indicador de autocorrelação espacial local serão plotados em um **Diagrama de Espalhamento** onde as viagens produzidas serão classificadas quanto sua relação com seus vizinhos. Esse diagrama possibilita concluir sobre as zonas que apresentam semelhanças ou diferenças com relação aos seus vizinhos.

O cruzamento dos dados do *LISA MAP* com os dados do Diagrama de Espalhamento possibilitará identificar aquelas zonas que são classificadas com dependência espacial local positiva e que apresentam, concomitantemente, significância estatística no nível local.

Após as etapas descritas, serão elaborados dois tipos de modelos. O primeiro será desenvolvido para as zonas de *clusters* e o segundo será baseado nas 128 zonas de tráfego. Para a elaboração de ambos os modelos serão produzidas matrizes de correlação para selecionar as variáveis explicativas mais relacionadas com as viagens produzidas e menos relacionadas entre si. Em seguida, ambos os modelos serão aplicados exclusivamente às zonas de *clusters,* possibilitando realizar comparações de seus desvios com relação aos dados observados em campo.

A Figura 11 representa um esquema de fluxograma que contempla todas as etapas do desenvolvimento do presente trabalho. Nos próximos itens desse capítulo serão apresentadas essas etapas considerando todos seus detalhes.

Figura 11 – Fluxograma das etapas do desenvolvimento



8.1. OBTENÇÃO E PREPARAÇÃO DOS INSUMOS

A primeira fase compreende a obtenção dos **insumos** necessários para o desenvolvimento de modelos de produção de viagens junto a órgãos oficiais e empresas de consultoria de planejamento de transportes.

Esses insumos devem compreender os diferentes bancos de dados relativos às viagens entre origens e destinos que ocorrem na Região Metropolitana de Campinas, assim como os dados socioeconômicos da população. Todos os dados a serem utilizados devem estar associados a unidades territoriais de análise, elaboradas em um nível de agregação que possibilite realizar cálculos estatísticos e o desenvolvimento de modelos de previsão de demanda, especificamente modelos

de produção de viagens. A principal fonte de dados que reúne essas características é a Pesquisa Domiciliar de Origem Destino 2012 realizada na RMC.

Após a obtenção dos insumos da pesquisa, há a necessidade de elaborar um banco de dados georreferenciado e unificado para a elaboração de modelos de produção de viagens. Para cada zona pesquisa, devem ser identificadas as variáveis relativas às características socioeconômicas e de viagens classificadas por motivo de viagem, no horário de pico mais carregado do sistema de transportes, em um dia útil da semana.

O produto da fase de obtenção e preparação dos insumos deve contemplar:

- Base geográfica desenhada em software de GIS, contendo as zonas de tráfego da RMC e chave de relacionamento que possibilita associar banco de dados.

- Banco de dados que contemplam os totais de viagens produzidas em cada zona de tráfego da RMC na hora pico de um dia útil.

- Banco de dados socioeconômicos das pessoas que realizam viagens, georreferenciados com as correspondentes zonas de tráfego da RMC.

8.2. *CLUSTERS* ESPACIAIS

Após a obtenção e preparação dos insumos necessários para a elaboração dos modelos de produção de viagens, procedeu-se a aplicação de técnicas de análise espacial para a identificação de *clusters* de zonas que apresentam dependência espacial com seus vizinhos.

Inicialmente será confirmada a hipótese de dependência espacial na RMC, ou seja, que as viagens produzidas ocorrem de forma não aleatória e seguem padrões relacionados com sua organização no espaço. Para verificar a hipótese de dependência espacial na RMC, será calculado o índice de autocorrelação espacial denominado Indicador Global de Moran, representado por um valor numérico absoluto relativo ao conjunto de todas as zonas de tráfego da RMC.

Previamente ao cálculo desse indicador há a necessidade da definição de uma matriz de proximidade referente a estrutura de vizinhança a ser adotada para cada par de zonas. Para tanto, será verificado por meio de mapas e histogramas de conectividade, como a variável viagens produzidas por pessoa na RMC se distribui no espaço. A depender dessa análise, será adotado um dos três tipos de matrizes de proximidade apresentados na literatura, de acordo com os pontos de vista descritivo, teórico ou topológico.

O próximo passo será a realização do teste da hipótese nula de que os dados são dispostos aleatoriamente no espaço. Para realizar esse teste será verificado o tipo de distribuição dos dados da variável viagens produzidas, seja distribuição normal ou outra distribuição qualquer. Caso essa distribuição apresentar-se como uma curva normal de dados, será adotado o teste paramétrico que compara esse tipo de curva com os dados originais do Indicador Local de Moran.

No caso de os dados reproduzirem outro tipo de distribuição, então será adotado o teste não-paramétrico de permutações aleatórias dos valores da variável nas zonas de tráfego da RMC. Esse teste deve ser realizado com auxílio de ferramentas computacionais que possibilitem a realização de um número elevado de permutações.

Após a confirmação da presença de dependência espacial global e a constatação de significância estatística no conjunto das zonas de tráfego da RMC,

será necessário identificar a localização desse fenômeno na escala das unidades territoriais de análise (zonas de tráfego).

Nessa etapa será utilizado um dos dois indicadores de autocorrelação espacial local, de Moran ou Getis, para cada zona de tráfego. O indicador local de Moran será utilizado caso seja constada a presença de dependência espacial no nível Global da RMC, caso contrário, será utilizado o indicador local de Getis.

O cálculo do indicador local de autocorrelação espacial possibilitará concluir, para cada zona, se a variável em análise (viagens produzidas por pessoa) ocorre aleatoriamente no espaço ou se existem padrões espaciais (dependência espacial) associados ao comportamento dessa variável com relação à suas zonas vizinhas.

Em seguida, será construído um Diagrama de Espalhamento em que as abscissas são representadas pela variável viagens produzidas por pessoa em cada zona e as ordenadas representam os valores médios dos vizinhos da mesma variável para cada zona. As zonas serão identificadas nos quatro quadrantes desse diagrama, de acordo com o tipo associação espacial com seus vizinhos e classificadas em zonas com dependência espacial positiva, negativa e *outliers*.

Particularmente, as zonas de tráfego que apresentarem valores positivos de indicador de autocorrelação espacial (zonas dos quadrantes Q1 e Q3 do Diagrama de Espalhamento) serão selecionadas para as análises posteriores de identificação de dos *clusters* espaciais. As zonas de tráfego que apresentarem valores negativos, serão classificadas nos quadrantes Q2 e Q4 e representarão as zonas com instabilidades da dependência espacial, tais como, ilhas de zonas com valores do indicador de autocorrelação local elevado cercado por vizinhos com valores baixos da mesma variável ou, ao contrário, zonas de autocorrelação baixa cercada por zonas com valores de autocorrelação elevada.

Em seguida, será calculada a significância estatística da dependência espacial no nível local. Para tanto será utilizado o mesmo método empregado para a significância estatística do indicador de autocorrelação global, ou seja, o método não paramétrico de permutações dos valores nas zonas de tráfego da RMC, porém, nesse caso, serão comparados os valores de cada indicador local, que constitui a curva empírica, com os valores originais dos iniciadores de autocorrelação local de cada zona. A apresentação da significância será realizada por meio do *LISA* Map que classifica cada zona de tráfego de acordo com os valore de $p < 0{,}05; p < 0{,}01; p < 0{,}001$ *e* $p < 0{,}0001$. Todas as zonas significantes serão selecionadas e investigadas quanto a possibilidade de constituírem *clusters* espaciais.

A identificação dos *clusters* espaciais de zonas será concluída identificando-se, nos quadrantes Q1 e Q3 do Diagrama de Espalhamento, as zonas que, no Lisa Map, apresentarem um nível de significância $p < 0{,}05$.

8.3. MODELOS DE PRODUÇÃO DE VIAGENS

Após a preparação dos dados, serão desenvolvidos dois modelos de produção de viagens. O primeiro será um modelo baseado nas zonas pertencentes aos clusters espaciais obtidos anteriormente e será denominado “modelo com cluster” o segundo será um modelo geral obtido baseado em todas as zonas da RMC sem distinção e será denominado “modelo sem cluster”.

Cada modelo será representado por uma equação que reproduza, com alguma margem de erro, as viagens observadas em cada zona de tráfego consideradas em sua construção e que ocorrem na hora de pico do sistema, durante um dia típico da semana.

A elaboração dos modelos será realizada por meio da técnica matemática de regressões lineares múltiplas baseada no método dos mínimos quadrados. Esse

processo envolve a seleção da variável resposta e das variáveis explicativas, a verificação das premissas das regressões, a elaboração do modelo, testes de generalização.

8.3.1. Variável resposta

A variável resposta dos modelos será escolhida entre os mais representativos estratos de viagens classificados segundo sua base domiciliar e base não domiciliar e por motivo. A partir dos dados da hora pico mais solicitada na RMC, deverão ser tabulados os estratos mais comumente utilizados em modelos de previsão de demanda: BDT – base domiciliar trabalho, BDE – base domiciliar estudo, BDO – base domiciliar outros e BND – base não domiciliar.

8.3.2. Variáveis explicativas

A variáveis explicativas dos modelos serão escolhidas no banco de dados da pesquisa OD segundo dois critérios. No primeiro deles a variável deve apresentar elevada correlação com as viagens produzidas e no segundo critério, as varáveis devem estar pouco relacionadas entre si para evitar multicolinearidade.

Para realizar essas escolhas serão elaboradas matrizes de correlação Pearson em que serão consideradas altamente correlacionadas as variáveis que apresentarem $r > 0{,}5$ e significantes ao nível de $p < 5\%$. Por outro lado, serão consideradas pouco correlacionadas as variáveis que apresentarem $r < 0{,}5$ e significância ao nível de $p < 5\%$.

8.3.3. Premissas das regressões

As premissas de regressões lineares (linearidade, independência, normalidade, homocedasticidade) serão checadas por meio de gráficos de resíduos

em que as abscissas representam a variável resposta do modelo e as ordenadas representam os resíduos entre dados modelados e observados.

8.3.4. Parâmetros e testes do modelo

Após a seleção das variáveis explicativas e variável resposta, serão desenvolvidas as regressões lineares pelo método dos mínimos quadrados com auxílio do software comercial SPSS, obtendo-se, para ambos os modelos, todos os parâmetros β_n das variáveis explicativas, a constante do modelo, e os indicadores clássicos de avaliação do modelo, tais como, coeficiente de determinação, R^2, $R^2 ajustado$, $teste\ F$ e $teste\ t$. Além disso serão obtidos os valores dos desvios ê para cada zona de tráfego correspondentes aos dados modelados e observados de viagens produzidas e que serão utilizados na última fase dessa pesquisa.

8.4. COMPARAÇÃO ENTRE MODELOS

A foco central desse capítulo corresponde aos cálculos necessários para comparar os modelos com *clusters* e o modelo clássico desenvolvido nos itens anteriores.

Em tese, os modelos clássicos são construídos para reproduzir os valores observados de todas as zonas de uma área de estudo, sem distinção. Portanto, supõem-se que quando aplicados a uma quantidade menor de zonas (que fazem parte do universo que foi considerado para sua elaboração), consigam reproduzir os dados observados da mesma forma.

Por outro lado, os modelos baseados em clusters, como qualquer outro modelo, só podem apresentar resultados restritos às zonas que serviram como base para sua construção e não se aplicam a previsões em conjunto diferente de zonas, como por exemplo, as demais zonas da RMC.

Como ambos os modelos só podem ser submetidos, para efeito de comparação, às mesmas zonas de tráfego, a comparação que se pretende realizar será sobre as zonas de clusters. Portanto, ambos os modelos serão aplicados aos dados das zonas de clusters e seus resultados serão analisados quanto aos desvios entre dados modelados e observados, por meio de tabelas e gráficos comparativos.

9. DESENVOLVIMENTO

O desenvolvimento desse trabalho foi balizado pela metodologia proposta no capítulo anterior e envolveu quatros fases descritas a seguir.

9.1. OBTENÇÃO E PREPARAÇÃO DOS INSUMOS

A obtenção de insumos dessa pesquisa de mestrado, teve como objetivo subsidiar o desenvolvimento dos modelos de produção de viagens da RMC. Tais insumos procederam do banco de dados da Pesquisa Origem - Destino da RMC realizada Companhia Paulista de Trens Metropolitanos – CPTM, entre outubro de 2011 até maio de 2012. Esse banco de dados é parte integrante do documento STM (2012) e foram obtidos junto ao Departamento de Planejamento da Companhia Paulista de Trens Metropolitanos – CPTM.

De acordo com o documento STM (2012), essa pesquisa cobriu 19 municípios da RMC, representados em 185 zonas de tráfego, que correspondem a menor unidade territorial de análise.

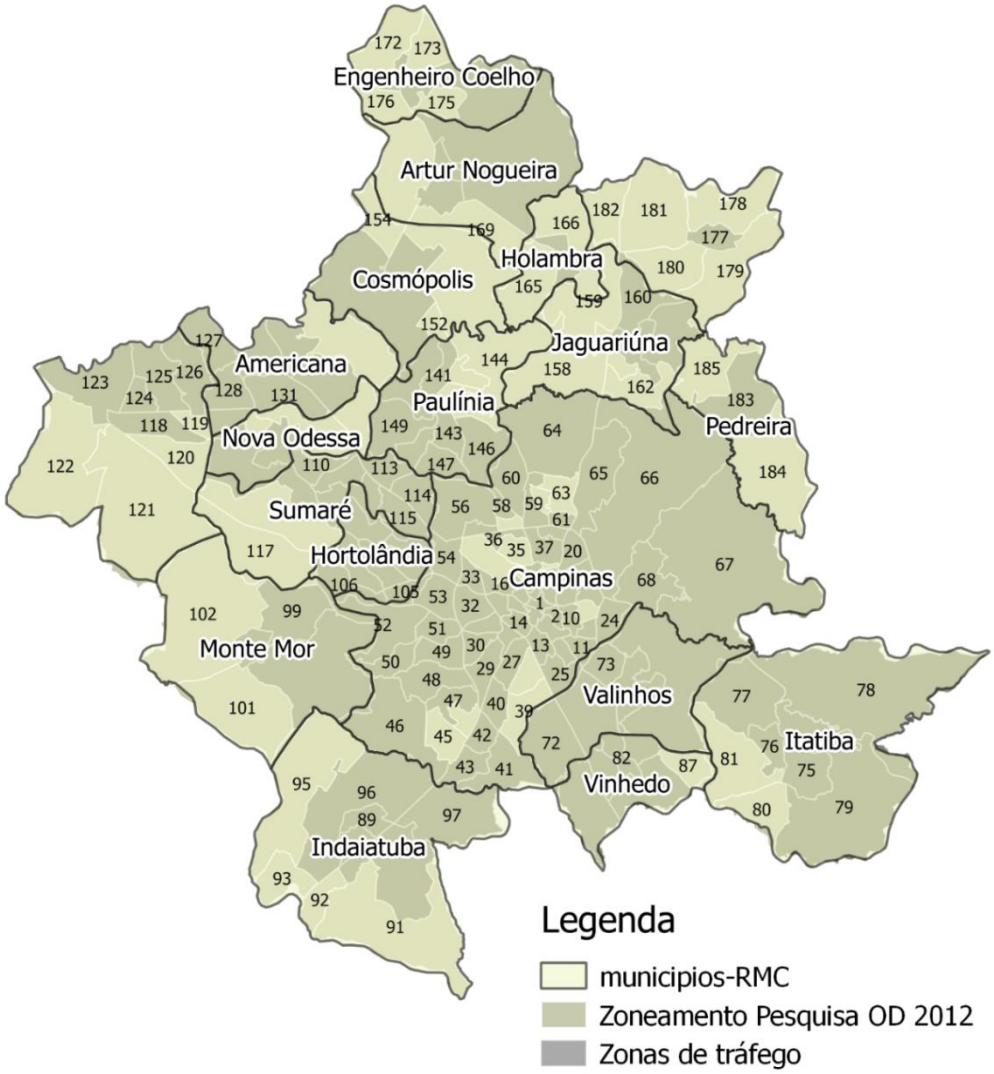
Essas zonas foram concebidas a partir de agregações da base territorial dos setores censitários do IBGE - Instituto Brasileiro de Geografia e Estatística, considerando-se aspectos econômicos, polos geradores, o sistema de transporte, os equipamentos urbanos, as barreiras físicas e áreas vazias.

A partir da definição das zonas de tráfego, foram selecionadas aquelas que seriam contempladas no plano amostral da pesquisa. De acordo com STM (2012), definiu-se que zonas com menos de 515 domicílios, segundo os dados do IBGE, não seriam pesquisadas.

Seguindo esse critério, foram selecionadas 128 zonas (denominadas nessa dissertação como zonas de tráfego) entre 185 zonas previamente planejadas para a pesquisa. No total, essa pesquisa realizada pela CPTM, totalizou 17.551 visitas à domicílios, onde foram realizadas entrevistas por meio de questionários (Anexos 1 a 4).

A Figura 12 a seguir apresenta o zoneamento original de 185 zonas de tráfego e as 128 zonas pesquisadas da OD 2012, assim como a correspondência com os municípios da RMC.

Figura 12 - *Cluster Map* – Densidade Populacional de New York



As principais características do banco de dados resultante da pesquisa são os totais de viagens que ocorrem entre zonas de origem e destino realizadas pela população em suas atividades diárias e, também, suas características socioeconômicas. Todos esses dados estão georreferenciados e são representados em mapas que podem ser visualizados em softwares de geprocessamento, análise estatística e de também de análise espacial.

A preparação dos insumos para a elaboração dos modelos dessa pesquisa envolveu, primeiramente, a tabulação dos dados socioeconômicos e de

viagens, na hora de pico do sistema de transportes da RMC, nas 128 zonas de tráfego da Pesquisa Origem Destino da RMC 2012.

O resultado dessa tabulação é um banco de dados georreferenciado, em que cada linha corresponde a uma zona de tráfego e cada coluna representa os dados socioeconômicos. Os demais dados da pesquisa relacionados a matrículas e empregos foram descartados do banco de dados por serem mais relacionados a atração do que à produção de viagens. Na Tabela 5 é apresentado um extrato desse banco de dados.

Tabela 5 – Extrato do banco de dados tabulado (insumo para modelos e análise espacial)

ZDOM	Cidades	Populacao	Pop_5_17_ sum	Pop18_25_ sum	Frota_A_ sum	Frota_B_ sum	Frota_C_ sum	Frota_D_ sum	Frota_E_ sum	Frota_F_ sum	RendaR$_ sum	P_BDT
1	Campinas	36109,74	7824	10039	1607	6379	9094	4471	2099	1607	3605515	2679
2	Campinas	21689,32	4376	6015	2409	5539	5835	4365	1479	2409	2742951	2100
3	Campinas	20200,59	5540	4653	1251	3647	5549	4336	1670	1251	2454117	2169
4	Campinas	6396,66	561	1149	284	1610	2293	2168	598	284	531324	600
5	Campinas	13881,71	3264	2946	796	2794	4414	2824	2567	796	1518562	1124
6	Campinas	8411,37	1250	1898	678	1701	3610	2527	848	678	710467	428
7	Campinas	11728,97	565	3323	1019	1054	1711	306	0	1019	1265456	978
8	Campinas	28540,70	5442	8010	5755	7237	5902	1601	353	5755	4859664	2923
9	Campinas	23468,32	4340	6026	4085	3178	3438	914	1689	4085	3033729	2773
10	Campinas	24917,73	7315	7493	555	3723	5880	8865	2302	555	1572045	2863
11	Campinas	24129,55	9023	5099	1090	2446	7741	6588	2471	1090	1903008	2700
12	Campinas	19686,92	5405	6522	1210	2857	5032	4520	2235	1210	1484846	1981
13	Campinas	25112,73	4322	6893	1130	4649	8080	4572	2280	1130	2563770	3551
14	Campinas	25197,89	3999	7189	1405	4026	8449	4782	2304	1405	2624399	2517
15	Campinas	24303,96	6169	4177	0	3081	6702	3390	1314	0	2120759	2625
16	Campinas	24568,06	5375	6542	4624	1889	3783	2846	1701	4624	2493653	2524
17	Campinas	4351,52	190	922	889	1023	2098	382	0	889	509211	640
18	Campinas	19759,94	2711	5282	1667	4515	6398	3528	1514	1667	2333565	2698
20	Campinas	26177,24	7840	8706	2901	2045	3660	2454	1563	2901	2866549	2582
21	Campinas	18612,60	5493	4976	289	3778	4138	3016	1189	289	1621010	2281
23	Campinas	16060,55	4494	5179	488	1539	3084	2035	1824	488	1205088	1452
24	Campinas	3807,10	1512	631	203	26	464	113	97	203	660345	585
25	Campinas	6533,54	2852	1277	234	1137	422	782	293	234	642328	938
27	Campinas	43105,04	16640	13395	247	144	6349	4188	1585	247	1284056	3947
28	Campinas	19685,14	4506	4388	122	2021	6368	4164	2449	122	1819234	1626
29	Campinas	39859,27	12323	11037	1852	4067	11904	9900	3454	1852	3073035	5116

A descrição de cada variável do banco de dados é apresentada a seguir.

- **Z_DOM** = zonas de tráfego onde foram realizadas pesquisa domiciliares.
- **Cidades** = nomes dos municípios da RMC relacionados às zonas de tráfego.
- **Populacao** = total de habitantes de cada zona de tráfego.
- **Pop_5_17** = total de habitantes na faixa etária entre 5 e 17 anos de cada zona de tráfego.
- **Pop18_25** = total de habitantes na faixa etária entre 18 e 25 anos de cada zona de tráfego.
- **Frota_A** = total de automóveis dos habitantes da faixa de renda A em cada zona de tráfego.
- **Frota_B** = total de automóveis dos habitantes da faixa de renda B em cada zona de tráfego.
- **Frota_C** = total de automóveis dos habitantes da faixa de renda C em cada zona de tráfego.
- **Frota_D** = total de automóveis dos habitantes da faixa de renda D em cada zona de tráfego.
- **Frota_E** = total de automóveis dos habitantes da faixa de renda E em cada zona de tráfego.
- **Frota_F** = total de automóveis dos habitantes da faixa de renda F em cada zona de tráfego.
- **RendaR$** = total de renda de cada zona de de tráfego.
- **P_BDT** = total de viagens produzidas com base domiciliar trabalho em cada zona de tráfego, na hora pico, em dia útil.

9.2. MODELO DE PRODUÇÃO CLÁSSICO

A partir da obtenção dos insumos e a preparação do banco de dados foi possível elaborar o modelo tradicional baseado nas 128 zonas de tráfego. Esse procedimento envolveu as cinco etapas (Seleção das variáveis do modelo, Elaboração do modelo sem clusters, Identificação de *clusters* espaciais, Desenvolvimento do modelo com clusters, comparação entre modelos) apresentadas a seguir.

9.2.1. Seleção das variáveis do modelo

A primeira etapa para a elaboração do modelo clássico de produção foi a definição da variável resposta e a seleção das variáveis explicativas. A variável resposta denominada nessa dissertação como P_BDT , foi definida como a quantidade de viagens por motivo trabalho produzidas por cada zona na hora pico, em dia útil. Essas viagens correspondem ao extrato de dados com a maior amostra, contando com 70% do total de viagens identificadas na RMC (Tabela 6).

Tabela 6 – Viagens classificadas por motivo

Estrato	Observações	%
base domiciliar trabalho	247.458	70%
base domiciliar escola	46.014	13%
base domiciliar outros	38.785	11%
base não domiciliar	19.983	6%
Total	352.240	100%

A seleção das variáveis explicativas foi realizada verificando-se preliminarmente aquelas mais fortemente correlacionadas com as viagens produzidas P_BDT. Para tanto, foi construída uma matriz de correlações Pearson r com valores que podem variar entre -1 até 1 apresentada na Tabela 7. De acordo com a metodologia, foram selecionadas as variáveis com correlações iguais ou superiores ao critério $r \geq 0{,}5$ e com nível de significância $p < 5\%$.

Tabela 7 – Correlações entre variáveis explicativas e viagens produzidas

	Populacao	Pop_5_17	Pop18_25	Frota_A	Frota_B	Frota_C	Frota_D	Frota_E	Frota_F	RendaR$
P_BDT	,872**	,772**	,839**	,176*	,454**	,774**	,745**	,598**	,176*	,761**

** $p < 1\%$; * $p < 5\%$.

Observa-se na tabela 7 que as variáveis destacadas em verde $Populacao, Pop\_5\_17, Pop\_18\_25, Frota\_C,\ Frota\_D, Frota\_E$ *e* $RendaR\$$ atendem aos critérios de correlação *Pearson* e de significância estatística, portanto foram pré-selecionadas como candidatas para compor o modelo. As demais variáveis $Frota\_A$, , $Frota\_B$ e , $Frota\_F$ apresentam correlação inferior ao estabelecido no critério de correlação e foram descartadas.

Posteriormente à pré-seleção de variáveis, foram identificadas aquelas com correlação baixa com as demais, evitando-se a presença de multicolinearidade, garantindo que cada variável represente uma parcela da explicação da variável resposta. Para tanto foi elaborada uma nova matriz de correlações apresentada na Tabela 8, onde buscou-se atender ao mesmo critério do valor do coeficiente Pearson $r < 0{,}5$ e nível de significância $p < 5\%$.

Tabela 8 – Correlações Pearson e níveis de significância - variáveis explicativas

	Populacao	Pop_5_17	Pop18_25	Frota_C	Frota_D	Frota_E	RendaR$
Populacao	1	,943**	,957**	,725**	,707**	,495**	,760**
Pop_5_17	,943**	1	,917**	,555**	,640**	,420**	,569**
Pop18_25	,957**	,917**	1	,670**	,700**	,496**	,709**
Frota_C	,725**	,555**	,670**	1	,826**	,734**	,727**
Frota_D	,707**	,640**	,700**	,826**	1	,829**	,519**
Frota_E	,495**	,420**	,496**	,734**	,829**	1	,373**
RendaR$	,760**	,569**	,709**	,727**	,519**	,373**	1
** Correlação significante no nível de p = 1%							
* Correlação significante no nível de p = 5%							

Observa-se que somente a variável , $Frota\_E$ apresentou baixas correlações com as demais variáveis pré-selecionadas, portanto destaca-se como uma variável para compor o modelo. Adicionalmente, a variável $Populacao$. apresenta o menor valor de correlação com a variável $Frota\_E$, portanto, de acordo

com os critérios estabelecido, as variáveis selecionadas para compor o modelo clássico de produção de viagens correspondentes à todas as 128 zonas de tráfego foram: $Populacao$ e $Frota\_E$.

9.2.2. Elaboração do modelo clássico

O modelo de produção de viagens foi elaborado a partir de regressões lineares múltiplas baseadas no método dos mínimos quadrados, considerando-se as variáveis selecionadas anteriormente, variáveis explicativas: $Populacao$ e $Frota\_E$ e variável resposta: P_BDT correspondente às viagens produzidas na hora pico em um dia útil em cada uma das 128 zonas pesquisadas.

O *software* estatístico SPSS foi utilizado para a elaboração das regressões lineares, cujos parâmetros: $\hat{\beta}_0$, $\hat{\beta}_{\text{Populacao}}$, $\hat{\beta}_{\text{Frota\_E}}$ são apresentados a seguir, juntamente com os principais testes estatísticos de interpretação do modelo: Coeficiente de Determinação R^2 e $R^2 ajustado$, $Teste\ F$ e $Teste\ t$.

Na Tabela 9 são apresentados os valores do Coeficiente de Determinação resultantes do modelo. Por se tratar de um modelo multivariado em que a entrada de mais de uma variável no modelo aumenta artificialmente o R^2, optou-se por analisar o $R^2 ajustado$.

Tabela 9 – Coeficiente de Determinação do modelo

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,893[a]	,797	,794	671,243

a. Predictors: (Constant), Frota_E_sum, Populacao

b. Dependent Variable: P_BDT

O resultado apresentado do $R^2 ajustado$ na Tabela 9 aponta que 79,4% da variação dos valores da variável P_BDT é atribuída às variáveis explicativas adotadas nesse modelo.

Em seguida foi realizado o $Teste\ F$, para análise da significância conjunta das variáveis explicativas, propiciando verificar a hipótese nula de que todos os coeficientes têm valor zero, ou seja, que pelo menos uma variável é significativa.

Analisando a saída produzida no software SPSS da Tabela 10, verifica-se que o $Teste\ F$ apresentou significância inferior a 0,05, indicando que a hipótese nula é rejeitada ao nível de 95% de significância.

Tabela 10 – Teste F

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	221331806,0	2	110665903,0	245,614	,000[b]
	Residual	56320966,74	125	450567,734		
	Total	277652772,7	127			

a. Dependent Variable: P_BDT

b. Predictors: (Constant), Frota_E_sum, Populacao

Após a constatação de que o modelo pode ser aplicado, foi realizado o teste dos parâmetros individuais do modelo de regressão como objetivo de verificar se cada variável e a constante do modelo são significativamente diferentes de zero, ou seja, se realmente podem fazer parte do modelo.

A Tabela 11 apresenta a saída de resultados desse teste onde é possível inferir que os coeficientes das variáveis explicativas $Populacao$ e $Frota\_E\_sum$ apresentam significância abaixo de $p < 5\%$, confirmando que essas variáveis influenciam no comportamento da variável resposta e, portanto, podem ser utilizadas no modelo.

Tabela 11 – Coeficientes das variáveis explicativas e teste t

Model		Unstandardized Coefficients		Standardized Coefficients	t
		B	Std. Error	Beta	
1	(Constant)	204,201	99,252		2,057
	Populacao	,067	,004	,762	16,451
	Frota_E_sum	,255	,053	,221	4,774

a. Dependent Variable: P_BDT

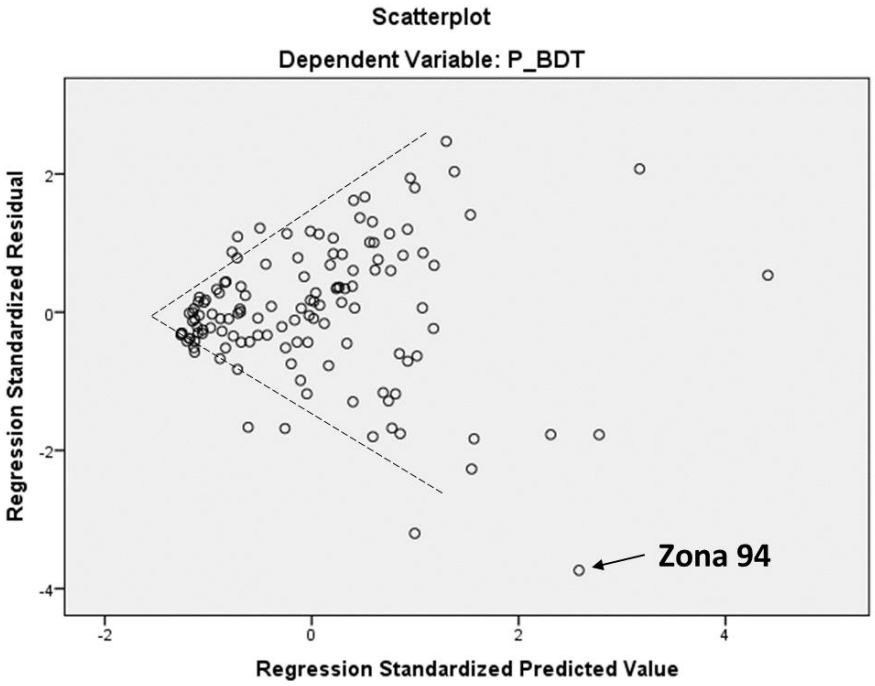
Os coeficientes $\hat{\beta}_{0\text{Populacao}}$ e $\hat{\beta}_{\text{Frota\_E}}$ e a constante obtidos pelas na Tabela 11 são apresentados na equação 13 representativa do modelo de produção de viagens para todas as 128 zonas pesquisadas da RMC.

$$P_{BDT} = 204{,}201 + 0{,}067 * Populacao + 0{,}255 * Frota\_E\_sum \qquad (11)$$

Após a verificação de validade do modelo, procedeu-se a verificação ao atendimento às quatro premissas das regressões lineares: linearidade, independência, normalidade, homocedasticidade.

Para tanto, foi elaborado o gráfico bidimensional, apresentado na Gráfico 16, em que os valores modelados das produções de viagens foram representados no eixo das abscissas e os valores dos desvios foram representeados no eixo das ordenadas.

Gráfico 16 – Distribuição de desvios



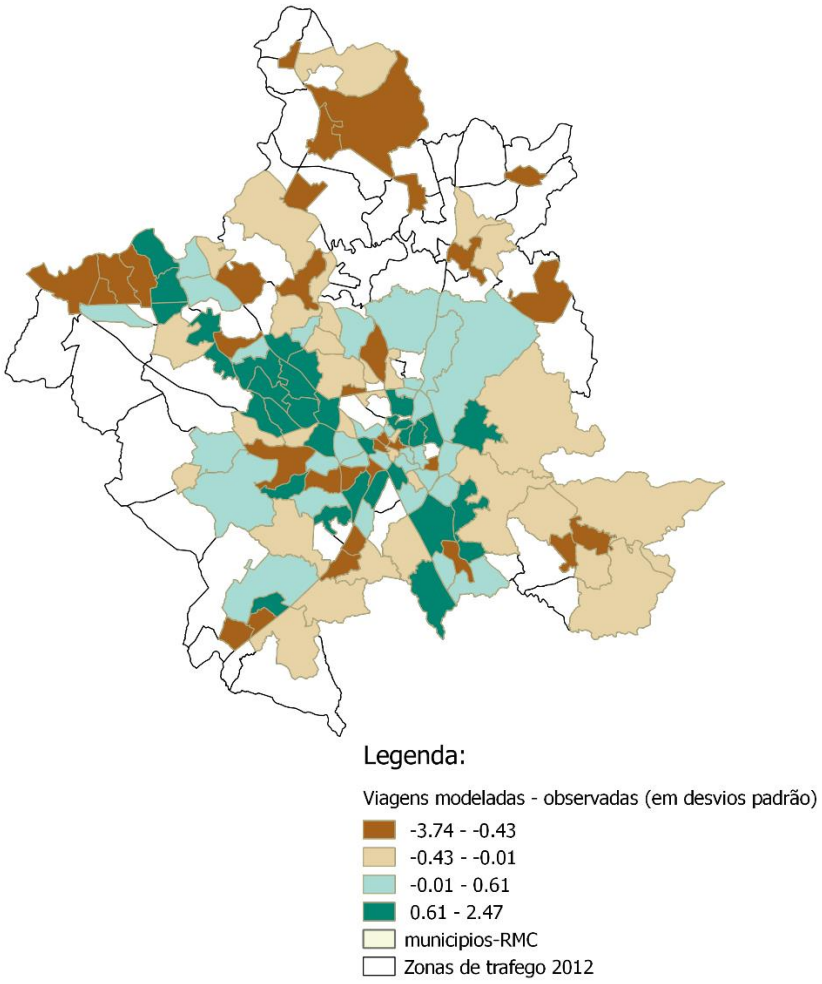
Observa-se que o modelo não apresenta variabilidade constante em todo o espectro dos dados e não reproduz as viagens com a mesma acurácia em todas as zonas de tráfego, sendo possível identificar desvios com magnitude elevada que comprometem o uso do modelo, principalmente à medida que a quantidade de viagens aumenta, o que representa violação da **premissa de homocedasticidade** dos dados.

Um exemplo desse fenômeno é a zona 94, que apresenta 2832 viagens observadas e o modelo resultou em 5347, ou seja, um desvio de 2515, que representa uma diferença de 47%.

A verificação da **premissa de independência** dos desvios entre dados modelados e observados, foi gerado um mapa (Figura 13) em que tais desvios são localizados em cada zona de tráfego. Observa-se que há vários conjuntos de zonas que seus desvios seguem, em muitas zonas de tráfego, uma

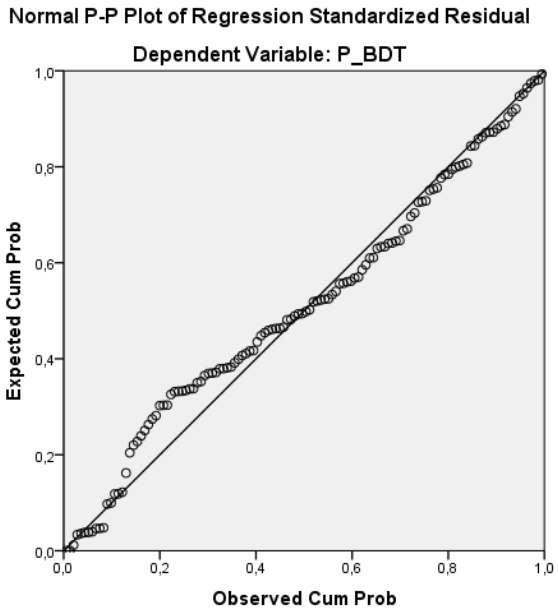
tendência de valores semelhantes com seus vizinhos contíguos, revelando um indício de violação da independência dos desvios.

Figura 13 – Distribuição espacial dos desvios



Para verificar a **premissa de normalidade** dos desvios foi gerado um gráfico de probabilidade normal no software SPSS representado na Gráfico 17, em que as abscissas correspondem a probabilidade observada acumulada e as ordenadas a probabilidade acumulada que se observaria se a distribuição fosse uma curva normal. Como os pontos representados nesse gráfico estão na sua maioria próximos a diagonal principal, pode-se concluir que os desvios entre os dados modelados e observados se caracterizam por uma distribuição aproximadamente normal.

Gráfico 17 – Gráfico de probabilidade normal



Por fim, para verificar a **premissa de linearidade** das variáveis explicativas, foram gerados gráficos de correlação parcial entre as variáveis explicativas e a variável dependente (Gráfico 18 e Gráfico 19), os quais possibilitam concluir que ambas as variáveis explicativas $Populacao\ e\ Frota\_E\_sum$ apresentam comportamento linear com a variável P_BDT.

Gráfico 18- Regressão parcial Pop

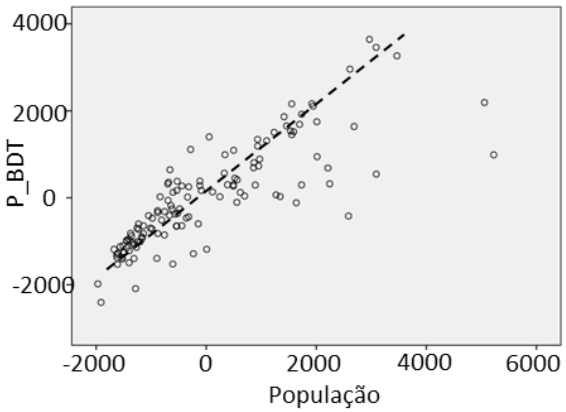
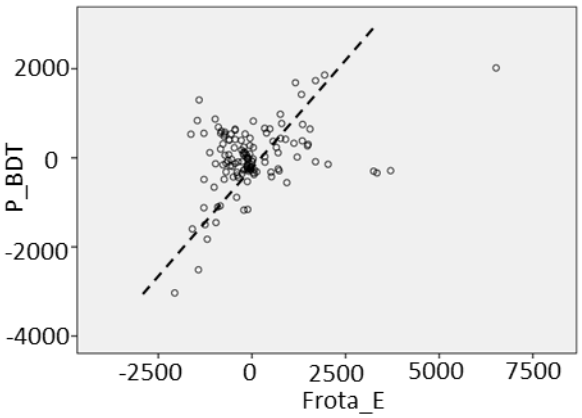


Gráfico 19- Regressão parcial Frota_E



9.3. *CLUSTERS* ESPACIAIS

A próxima fase envolveu a identificação dos *clusters* espaciais que correspondem a zonas com dependência espacial, ou seja, grupos de zonas formados segundo padrões espaciais de produção de viagens na RMC.

Para atingir esse objetivo foram realizados os seguintes procedimentos de análise espacial:

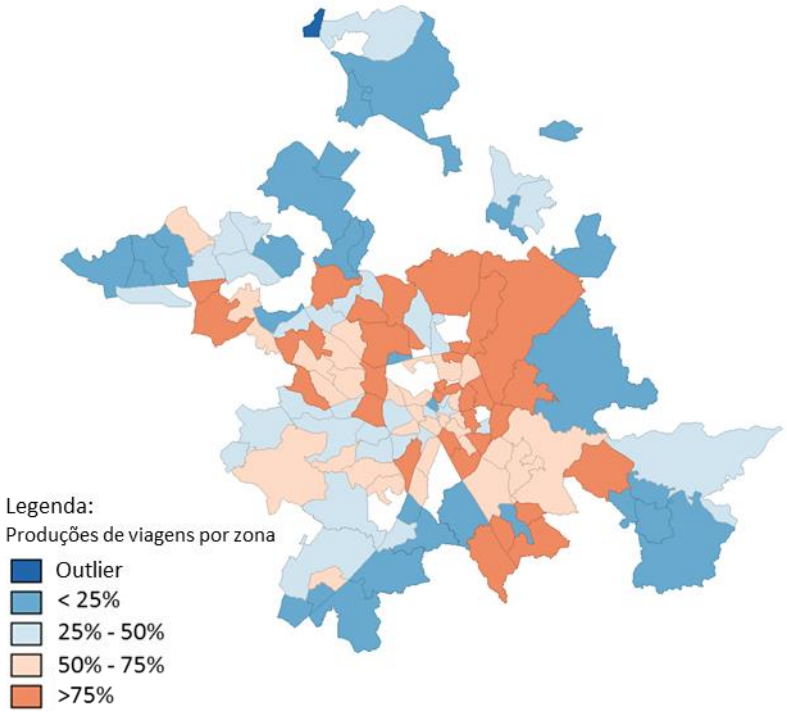
- Construção de uma matriz de proximidade.
- Cálculo de autocorrelação espacial global e a correspondente significância estatística.
- Elaboração do Diagrama de Espalhamento.
- Cálculo do teste de significância estatística local e sua representação gráfica no LISA Map.
- Identificação dos *clusters* espaciais.

9.3.1. Matriz de Proximidade Espacial

Segundo a metodologia, foi definida a estrutura de vizinhança mais adequada para a análise espacial dos dados dessa pesquisa e que será utilizada no cálculo da autocorrelação espacial. Esta estrutura foi materializada em uma matriz de proximidade definida a partir de uma análise prévia da distribuição das produções de viagem por pessoa no espaço da RMC.

Para auxiliar na definição desta estrutura foi plotado um mapa *box plot*, como mostrado na Figura 14, que possibilita analisar a tendência de distribuição espacial das viagens produzidas por zona de acordo por faixas de valores. A sua representação é feira ordenando-se os valores em ordem crescente dividindo a distribuição dos dados em quatro partes iguais, A primeira parte contém valores 25% menores, a segunda parte contém o intervalo de valores entre os 25% e 50% menores, a terceira parte contém o intervalo e valores entre os 50% e 75% maiores e a última parte contém os valores 75%maiores.

Figura 14 – *Box plot* das produções de viagem por pessoa por zona na RMC



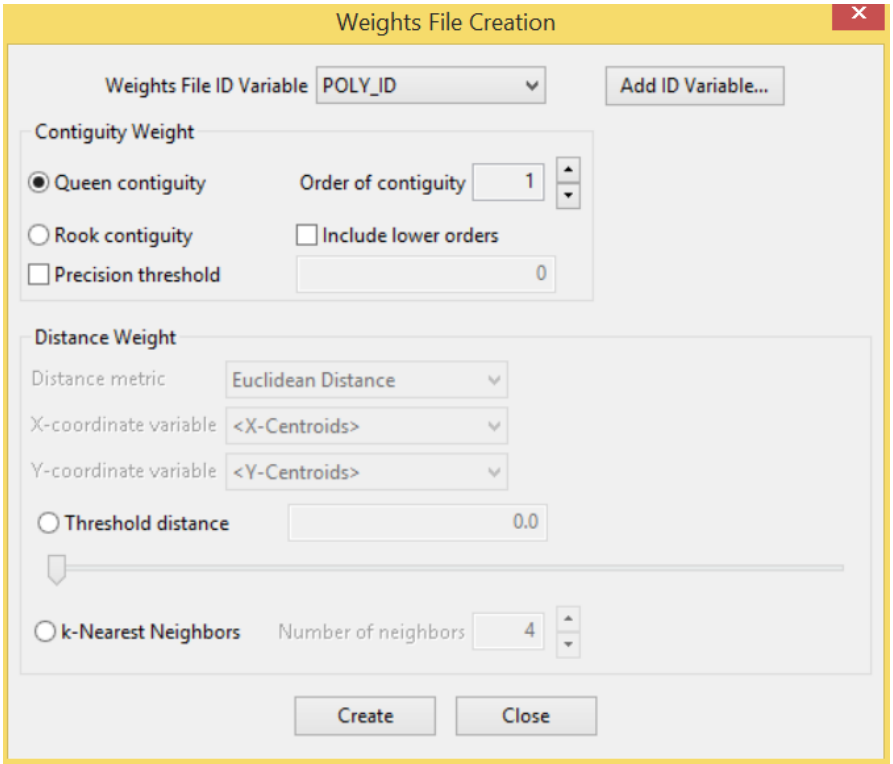
Observa-se que, de forma geral, há uma tendência dos valores com magnitudes semelhantes se aglomerarem em zonas adjacentes. Isso pode ser atribuído a vários fatores, desde influência das variáveis socioeconômicas que determinam as viagens até características locais que raramente são observadas em pesquisas de transpores, tais como violência urbana, iluminação, condições das calçadas, topografia, acessibilidade ao transporte coletivo, permeabilidade de linhas de ônibus urbano e combinações conjuntas desses fatores.

A partir da evidência de que a associação espacial das produções de viagens ocorre com maior intensidade entre zonas contiguas, foi possível adotar uma estrutura de matriz de proximidade do ponto de vista topológico que caracterize como vizinhas aquelas zonas que compartilham fronteiras físicas de qualquer comprimento em qualquer direção.

A medida de proximidade escolhida é representada por uma matriz de i zonas por j zonas, onde $i = j$ é a diagonal principal e que suas células seguem o critério dicotômico de primeira ordem, onde $w_{ij} = 1$ para i e j vizinhos físicos contíguos e $w_{ij} = 0$ para i e j não vizinhos.

Para elaborar esta matriz de proximidade foi utilizado o software Geoda que possibilita selecionar o tipo de matriz de vizinhança a ser adotado (cfr. Figura 15). No presente estudo foi adotada a matriz de contiguidade em que são considerados vizinhas todas as zonas que compartilham em qualquer direção, suas fronteiras físicas formadas por segmentos ou vértices (*queen contiguity*). Adicionalmente, foi adotado apenas vizinhos de primeira ordem (*order of contiguiity* = 1). A seguir é presentada a tela do programa Geoda em que são definidos os parâmetros mencionados anteriormente.

Figura 15 – Configuração da Matriz de Vizinhança – software Geoda



Após a configuração da matriz de vizinhança, foram realizados alguns testes para verificar se esta matriz atende às recomendações de especificação recomendadas no capítulo 6.1 – Proximidade Espacial.

Esta verificação foi realizada com auxílio do histograma de conectividade apresentado no Gráfico 20, que representa, no eixo x, o número de vizinhos e, no eixo y, a frequências de zonas com determinado número de vizinhos. Este histograma possibilitou identificar a quantidade zonas que apresentam 1, 2, 3, 4, 5, 6, 7, 8 e 9 vizinhos.

Na Gráfico 20 são identificadas as zonas pela quantidade de vizinhos (N). Observa-se que as zonas com áreas menores apresentam uma quantidade de vizinhos maior do que as zonas com grandes áreas. Especialmente esse fenômeno pode ser observado nas regiões mais centrais onde as áreas das zonas são menores do que as zonas das regiões mais afastadas do centro da RMC.

Gráfico 20– Histograma de conectividade

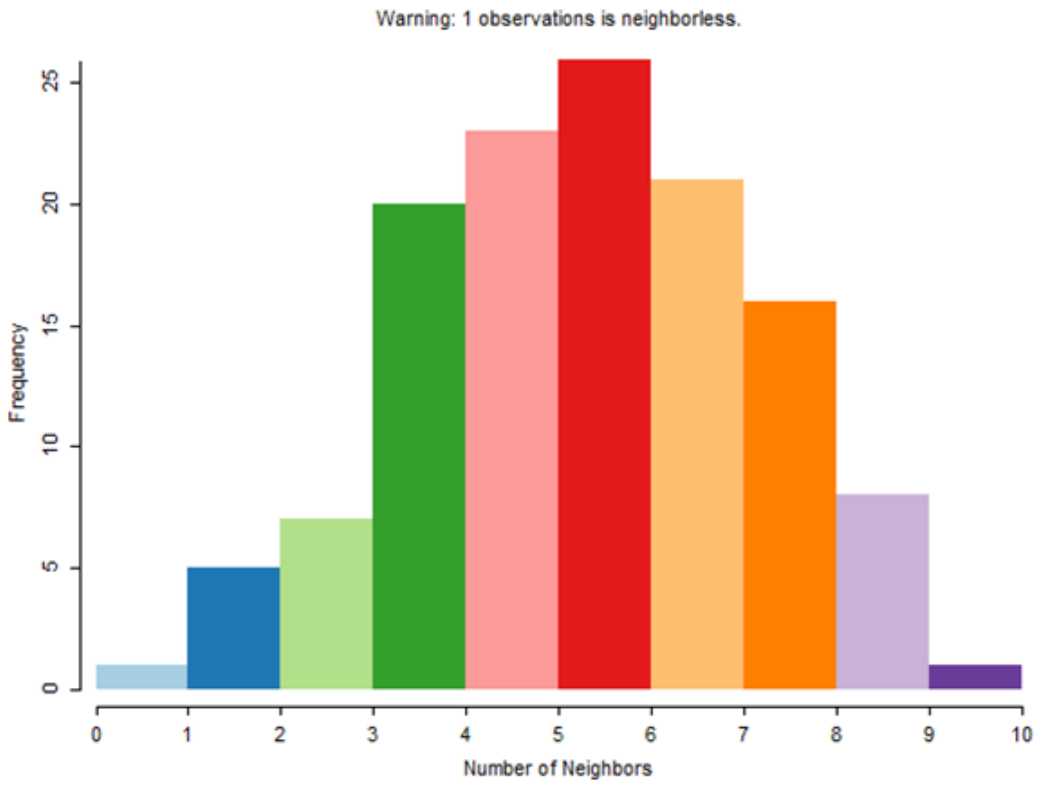


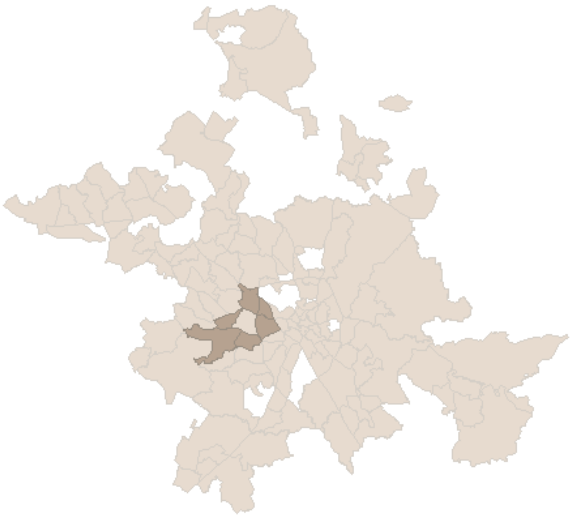
Gráfico 21 – Quantidade de vizinhos por zona

A partir da análise do histograma, foi verificado o atendimento às premissas mencionadas anteriormente.

- Premissa 1: A quantidade de vizinhos deve ser maior que 3 e menor que 7. Essa premissa foi atendida em 70% dos casos.
- Premissa 2: Utilizar mais de 60 unidades espaciais. Essa premissa foi atendida na integralidade, pois foram consideradas 128 zonas.
- Premissa 3: Utilizar especificações de primeira ou segunda ordem. Essa premissa foi atendida, pois foi utilizada matriz de proximidade de primeira ordem com vizinhos contíguos como apresentada no exemplo da Gráfico 22.

- Premissa 4: Considerar quantidade menor de vizinhos. Essa premissa foi atendida por considerar no máximo 10 vizinhos e a maioria das zonas (80%) tem até 7 vizinhos.

Gráfico 22 – Exemplo vizinhos de primeira ordem



9.3.2. Indicador de Moran Global

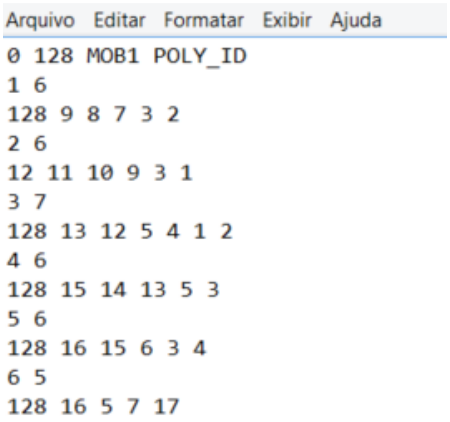
Após a adoção criteriosa da matriz de proximidade, procedeu-se o cálculo do indicador de autocorrelação espacial de Moran, também denominado Indicador Global de Moran (I), com o objetivo de caracterizar a dependência espacial das viagens produzidas em cada zona da RMC

A variável que se deseja testar a dependência espacial corresponde às viagens produzidas por zona na RMC $(PBDT)$. No entanto, como essa variável depende da magnitude das áreas das zonas, pode produzir resultados indesejados, tais como, efeitos de escala (falácia ecologia) e efeito de zoneamento, ambos abordados no item 6.5 desta dissertação. Portanto, optou-se por considerar que a variável $PBDT$ fosse ponderada pela população da respectiva zona $\left(\frac{P_{BDT}}{pessoa}\right)$, eliminando o efeito de suas dimensões.

Para a realização dos cálculos desse indicar procedeu-se construção de um banco de dados no software Geoda, associando-se a variável $\left(\frac{P_{BDT}}{pessoa}\right)$ a cada zona de tráfego. Em seguida, foi imputada de forma automática, a matriz de proximidade adotada no capítulo anterior no formato descrito a seguir.

A matriz de proximidade é representada por meio de uma estrutura de texto descrita abaixo e apresentada na Figura 16. A primeira linha do arquivo contém o cabeçalho da estrutura de dados e contempla quatro valores, 0 (exigência interna do software), número de observações (128 zonas de tráfego), o nome do arquivo georreferenciado em formato *shape* (MOB1), nome da variável ID que identifica a zona de tráfego (POLY_ID). Depois dessa linha a estrutura é idêntica, portanto para cada observação, há duas linhas de informação, sendo que a primeira contem a identificação da zona e sua respectiva quantidade de vizinhos e a segunda contem a identificação das zonas vizinhas.

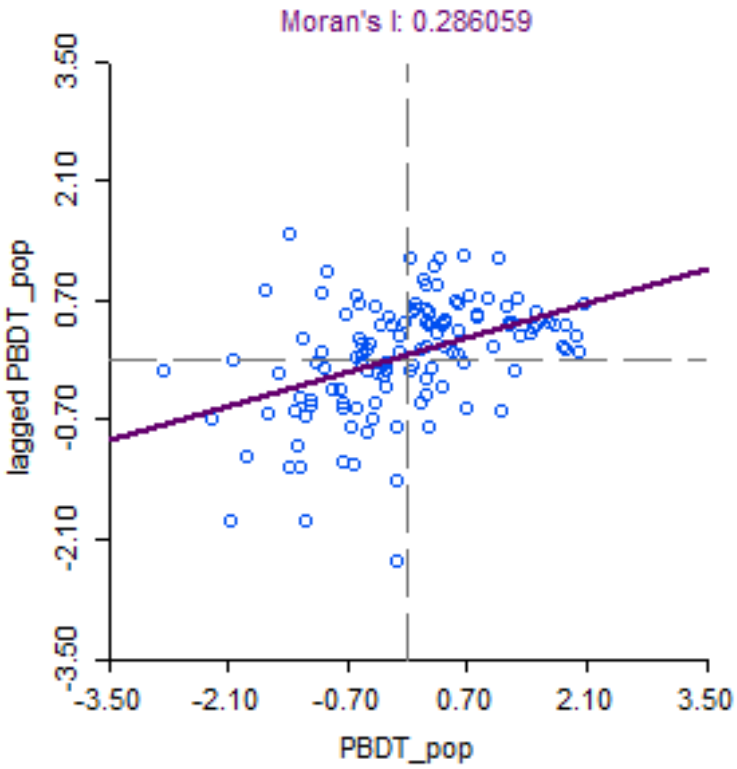
Figura 16 – Matriz de vizinhança (arquivo de entrada)



Os cálculos do indicador de Moran Global são realizados pelo software Geoda utilizando as mesmas formulações apresentadas no capítulo 7.1 e resultaram em um valor $\mathrm{I} = 0{,}2861$. O valor positivo entre 0 e +1 indica correlação direta e é um indício da presença da dependência espacial no conjunto dos dados da variável analisada e da possível existência de *clusters* de zonas na RMC.

Após calcular o valor do Indicador Global e Moran, foi elaborado um Diagrama de Espalhamento com o objetivo de classificar as zonas em quatro tipos de autocorrelação espacial de acordo com o quadrantes a que pertencem, sendo que os valores positivos são apresentados nos quadrantes Q1 (valores altos da variável e dos vizinhos) e Q3 (valores baixos da variável e dos vizinhos) e os valores negativos são apresentados nos quadrantes Q2 (valores baixos da variável e valores altos dos vizinhos) e Q4 (valores altos da variável e valores baixos dos vizinhos). Todos os valores foram apresentados na Gráfico 23.

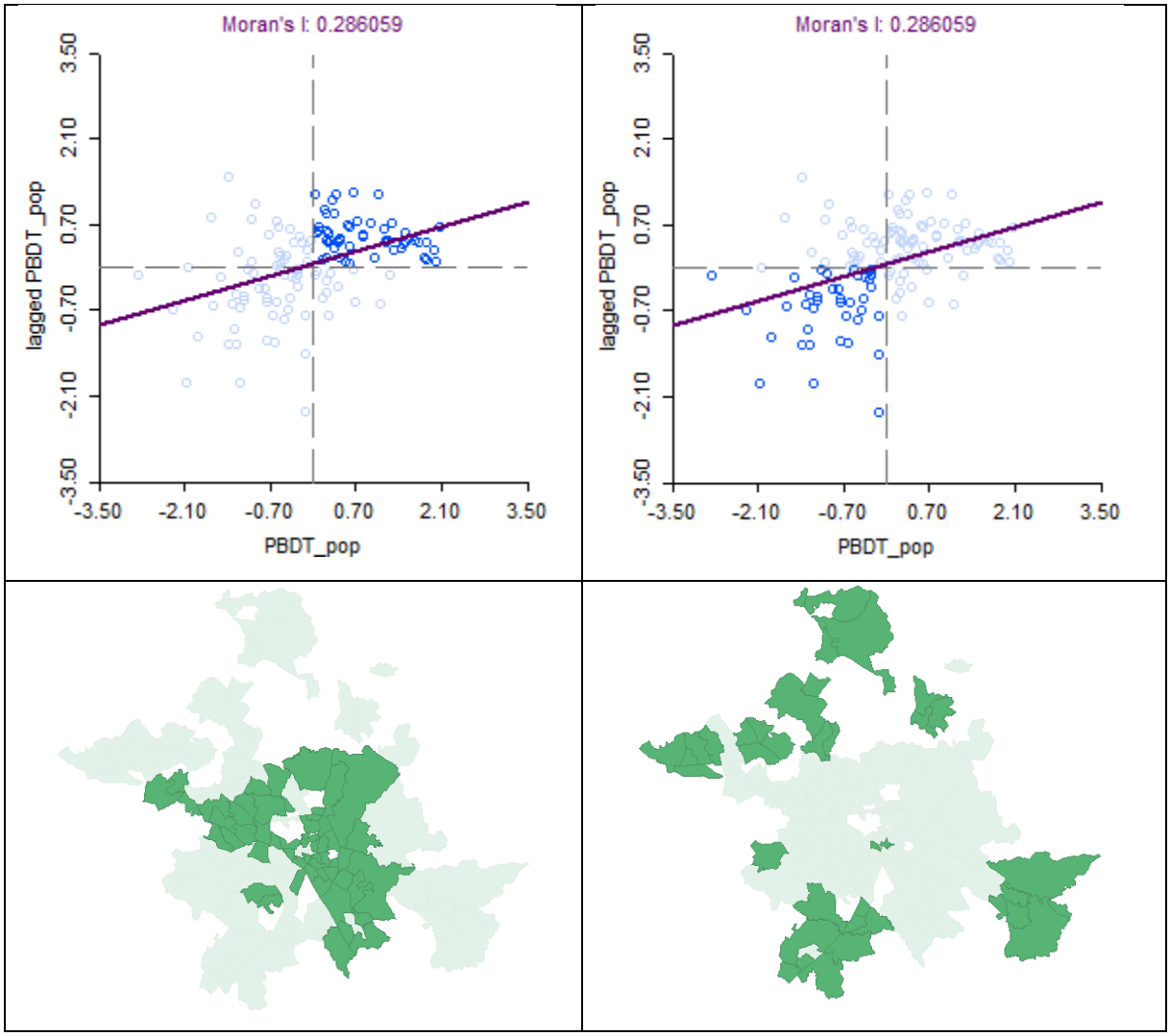
Gráfico 23 – Diagrama de Espalhamento Moran



O Diagrama de Espalhamento possibilita verificar que 54 zonas (42,1% do total) se encontram nos quadrantes Q1 e localizam-se na porção central da RMC; no quadrante Q3 foram classificadas 38 zonas (29,6% do total) localizadas principalmente nas cidades mais no perímetro da RMC. As zonas pertencentes aos quadrantes Q1 e Q3 apresentam dependência espacial positiva relação a seus

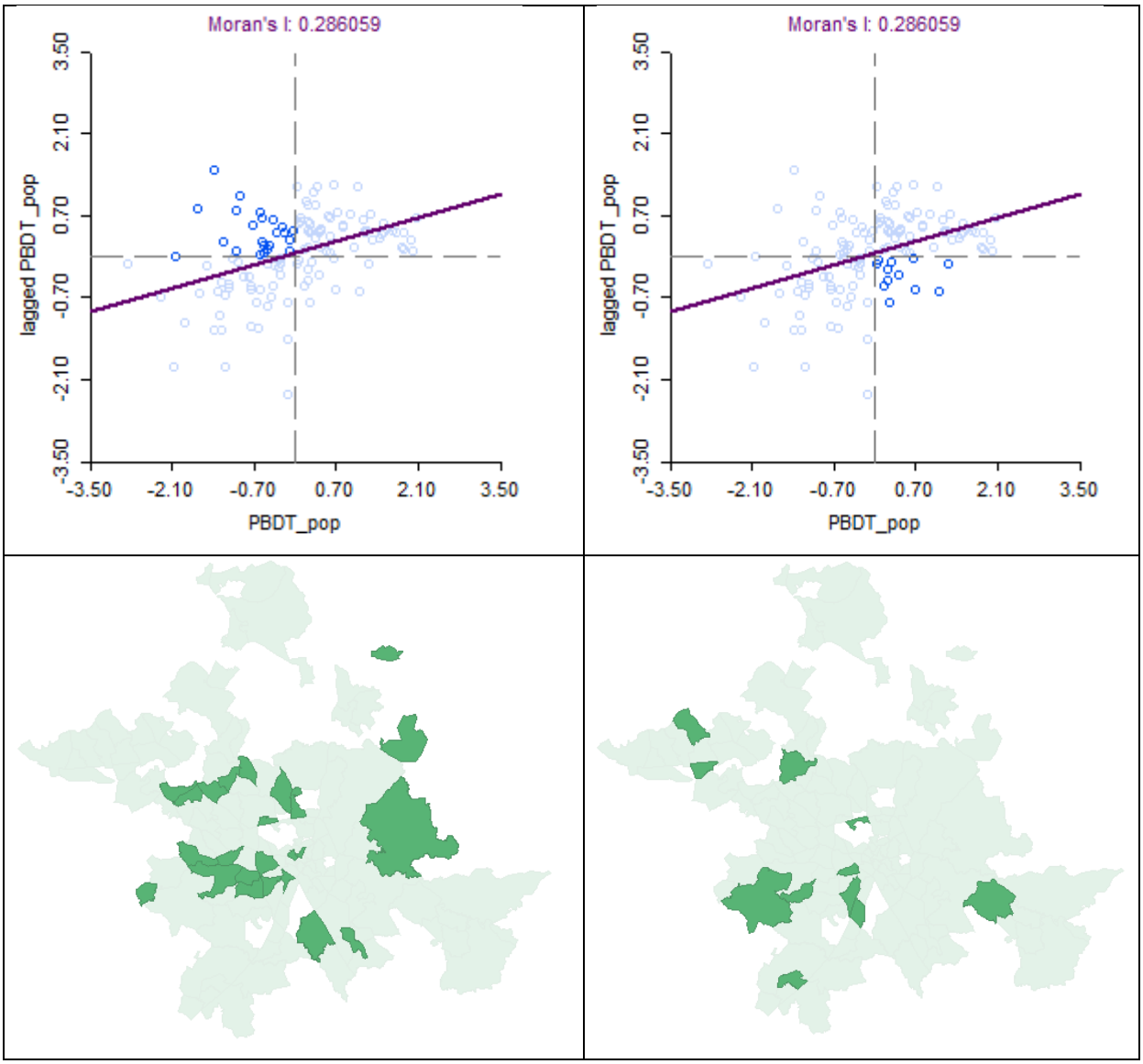
vizinhos e são candidatas a uma investigação mais aprofundada para detecção de clusters espaciais. Na Figura 17 seguir são apresentados diagramas de espalhamento em que cada ponto corresponde a uma zona de tráfego da RMC. A identificação espacial das zonas selecionadas dos quadrantes 1 e 3 podem ser observadas nos seus correspondentes mapas temáticas bidimensionais.

Figura 17 – Diagramas de Espalhamento e Mapas temáticos – Quadrantes Q1 e Q3



Na Figura 18 podem ser observadas 24 zonas pertencentes ao quadrante Q2 e 12 zonas no quadrante Q4. As zonas classificadas nos quadrantes Q2 e Q4 não parecem apresentar padrões de distribuição espacial, e encontram-se mais dispersas na RMC do que as demais zonas.

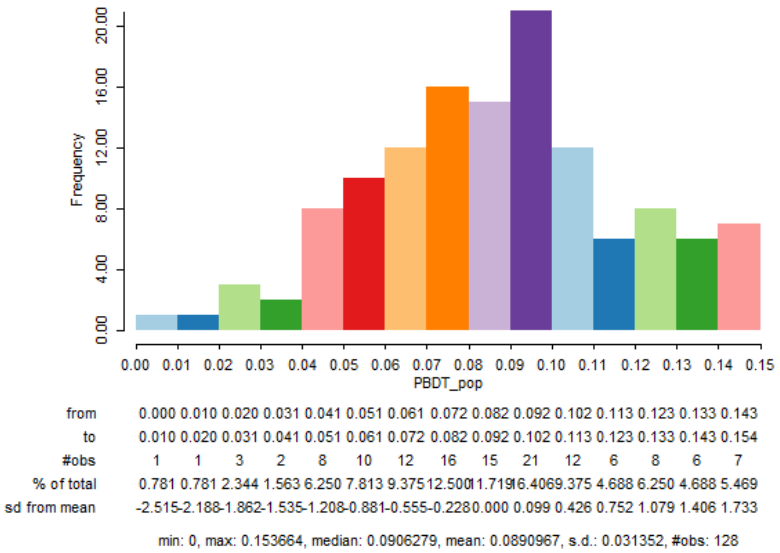
Figura 18 – Diagramas de Espalhamento e Mapas temáticos – Quadrantes Q2 e Q4



Em seguida, o Indicador Global de Moran foi submetido à verificação da sua validade estatística por meio do teste da hipótese nula de que seu valor foi obtido ao acaso, ou seja, que um novo cálculo realizado com outra amostra poderia chegar a resultados diferentes. Essa verificação compreende o teste de significância estatística em que a hipótese nula enuncia que os dados são dispostos aleatoriamente no espaço sem a presença de dependência espacial.

De acordo com os itens 7.2.2 e 7.2.3 apresentados anteriormente, os testes podem ser paramétricos ou não paramétricos, sendo necessário, portanto, a verificação da normalidade dos dados da variável em estudo. No Gráfico 24 é apresentado um histograma que representa a frequências de zonas em cada faixa de valores da variável $\left(\frac{P_{BDT}}{pessoa}\right)$. Exemplo: existem 12 zonas com valores entre 0,06 e 0,07 de viagens produzidas por pessoa.

Gráfico 24 – Distribuição de frequências de $\left(\frac{P_{BDT}}{pessoa}\right)$

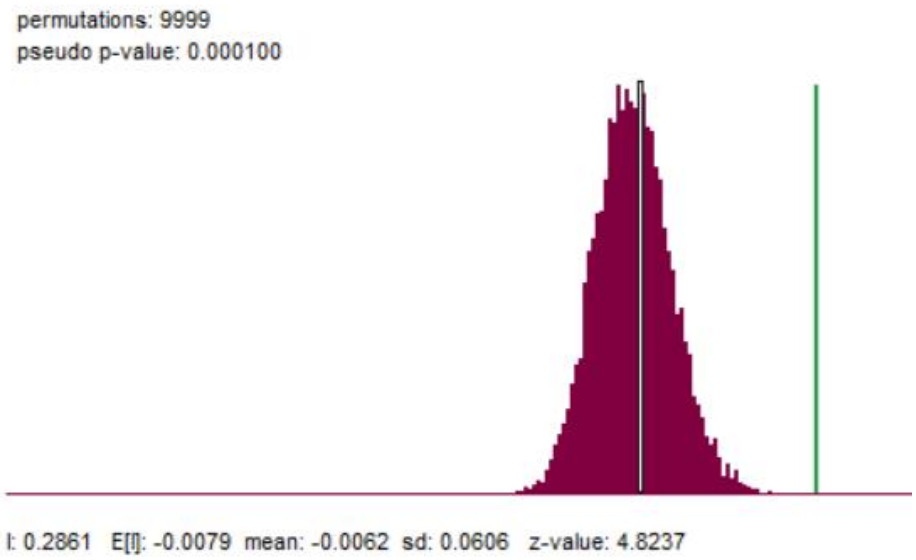


Este histograma possibilita inferir que os dados não compreendem uma distribuição normal perfeita, sendo recomendável a aplicação de testes não paramétricos para a verificação da significância estatística.

Neste trabalho foi adotado o teste não paramétrico de pseudo-significância em que foram comparados os dados originais da variável P_BDT com uma distribuição empírica elaborada a partir de novos valores do Índice de Moran, calculados para permutações espaciais aleatórias dos valores de $\left(\frac{P_{BDT}}{pessoa}\right)$.

Para realizar esse teste fez-se 9999 permutações aleatórias e em cada uma delas foi calculado um novo valor do Indicador Global de Moran. O conjunto desses indicadores possibilitou compor a distribuição empírica apresentada na Gráfico 25.

Gráfico 25 - Referência empírica da distribuição Moran sob a hipótese nula



Observa-se que juntamente com a distribuição empírica, foram extraídos alguns resultados que compreendem:

- $\mathrm{permutation}$ = número de permutações adotadas para a realização do teste de significância estatística do Indicador Global de Moran.
- $\mathrm{pseudo}\ p-value$ = significância estatística do teste.
- $\mathrm{E}[I]$ = média teórica esperada caso os dados fossem representados por uma distribuição normal.
- Mean = média de referência da distribuição empírica obtida das permutações de valores de $\left(\frac{P_{BDT}}{pessoa}\right)$. entre as zonas da RMC.

- $\mathrm{s.d.}$ = desvio padrão da distribuição empírica obtida das permutações de valores de $\left(\frac{P_{BDT}}{pessoa}\right)$. entre as zonas da RMC.
- $z - value$ = valor do Indicador Global de Moran estandarizado obtido dos dados da variável $\left(\frac{P_{BDT}}{pessoa}\right)$.nas 128 zonas da RMC.

O valor de $\mathrm{pseudo}\, p - value = 0{,}0001$ indica que nenhum dos valores simulados nas 9999 permutações é maior do que o valor de $z - value = 4{,}8237$. Este valor encontra-se fora da curva empírica e, portanto, confirma que o valor do Indicador Global de Moran $\mathrm{I} = 0{,}2861$ não é resultante da aleatoriedade e, portanto, possibilita concluir com segurança a existência de uma organização espacial, ou seja, existe dependência espacial da variável $\left(\frac{P_{BDT}}{pessoa}\right)$ entre as zonas da RMC.

9.3.3. Indicador de Moran Local

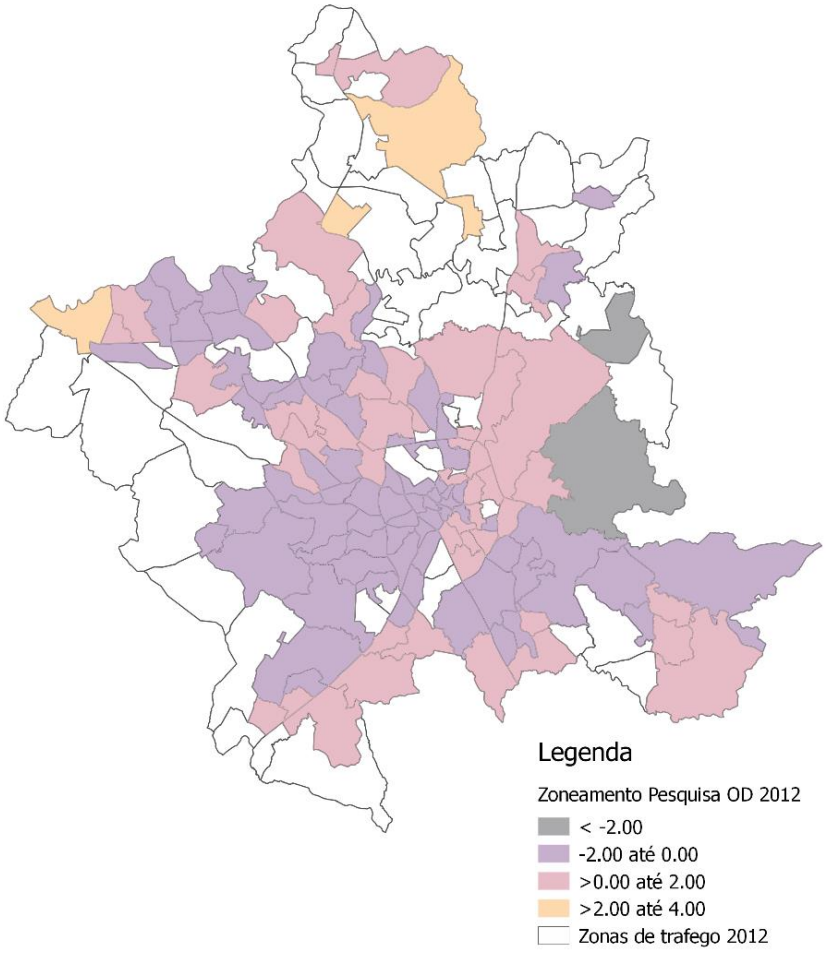
A constatação da presença de dependência espacial da variável $\left(\frac{P_{BDT}}{pessoa}\right)$ na RMC por meio do Indicador de Moran Global, enseja a necessidade de identificar localmente os valores individuais da dependência espacial em cada zona com relação aos seus vizinhos, propiciando identificar diferentes regimes de associação espacial não detectados no indicador global.

O cálculo da dependência espacial local foi realizado utilizando-se as formulações recomendadas na literatura como exposto no item 7.3 dessa dissertação envolvendo a utilização do valores da variável $\left(\frac{P_{BDT}}{pessoa}\right)$ em cada zona de tráfego e sua relação com os valores encontrados em seus vizinhos, sendo utilizado o mesmo conceito de vizinhança adotado anteriormente no cálculo do Indicador de Moran Global, ou seja, uma matriz de proximidade de vizinhos contíguos. Esses cálculos foram realizados com auxílio do software Geoda e Quantum Gis. Portanto,

foi necessário calcular os resultados em cada zona, extraí-los do Geoda, e finalmente, transformá-los para introduzi-los no Quantum Gis.

Na Figura 19 é apresentada um mapa contendo uma visualização de todos os resultados dos indicadores de Moran Local alocados a cada uma das 128 zonas de tráfego da RMC. Neste mapa apresenta-se os desvios padrão da variável $\left(\frac{P_{BDT}}{pessoa}\right)$

Figura 19 – Indicadores de Moran Local para todas as zonas da RMC



O mapa anterior possibilitou identificar padrões espaciais em que zonas com Indicadores de Moran Local semelhantes parecem estar mais próximas fisicamente umas das outras, apontando indícios de dependência espacial também

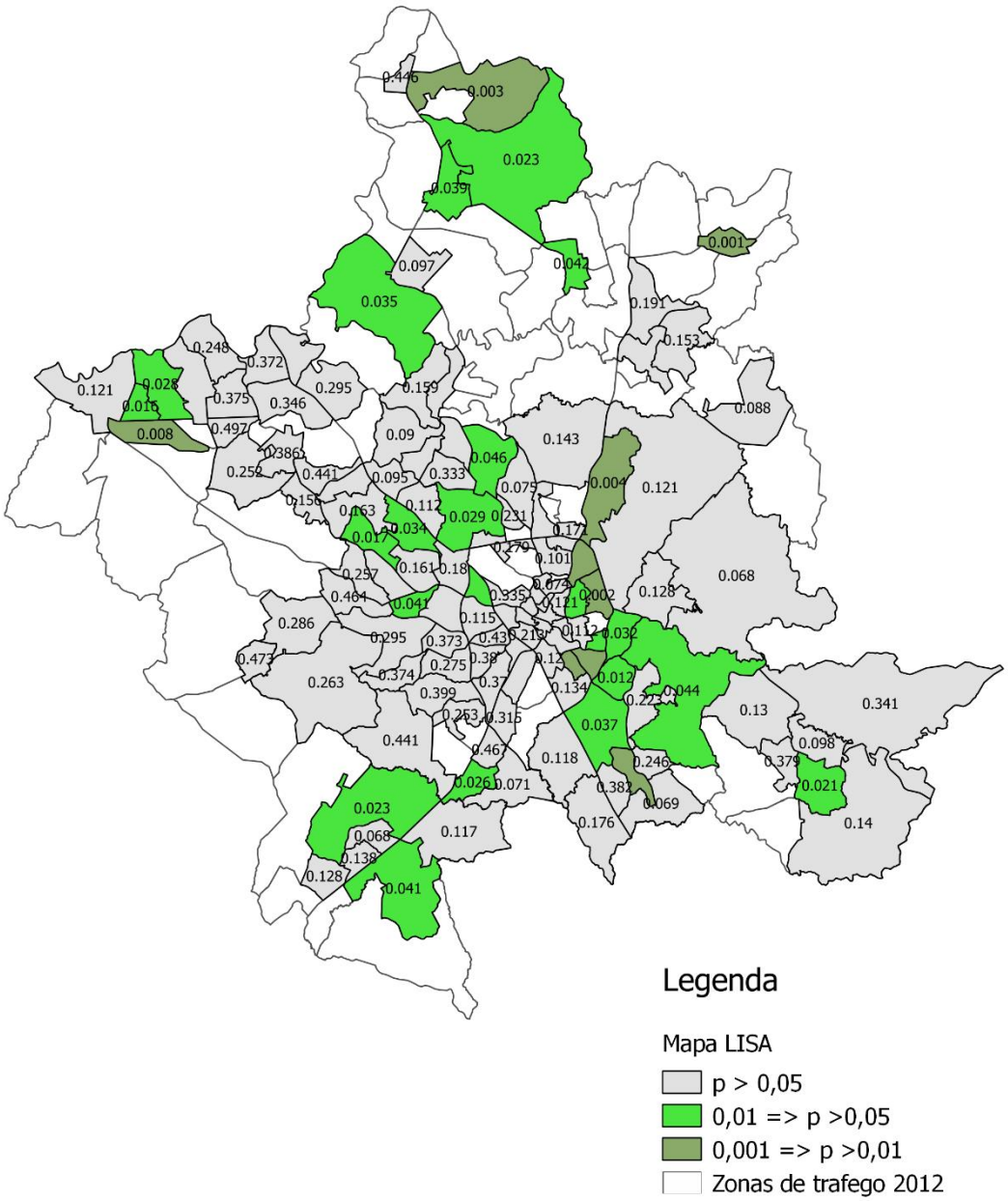
no nível local. Para confirmar essa hipótese foi realizado um teste de significância utilizando-se o mesmo método não paramétrico de permutações empregado anteriormente durante o cálculo do Indicador Global de Moran.

O resultado do Teste de Significância Local é plotado em um mapa coroplético denominado *LISA Map* apresentado na Figura 20. Nessa figura são apontados os valores do Indicador Local de Moran em cada zona, e seus respectivos níveis de significância. Os resultados correspondem à quantidade de permutações adotadas. Na presente dissertação foram adotadas 99999 permutações da variável $\left(\frac{P_{BDT}}{pessoa}\right)$ entre as 128 zonas pesquisadas da RMC.

Observa-se na mesma Figura 20, que após as permutações, 31 zonas (aproximadamente 24% de todas as zonas de tráfego) apresentaram significância p < 5%, sendo que 23 zonas possuem p entre 1% e 5%; 6 zonas com p situado em uma faixa entre 0,1% e 1%; 2 zonas com $p < 1$, e finalmente, 96 zonas com resultados não significantes do Indicador de Moran Local.

A zonas selecionadas no LISA Map serão utilizadas na próxima fase do trabalho para a identificação dos *clusters* espaciais.

Figura 20 – LISA *Map*– Valores da significância por zona



9.3.4. Clusters espaciais

A última fase do processo de identificação dos clusters de dependência espacial da variável $\left(\frac{P_{BDT}}{pessoa}\right)$ nas 128 zonas pesquisadas da RMC envolveu o cruzamento simples entre as zonas com Indicadores Locais de Moran significantes, obtidas no LISA *Map* e as zonas dos quadrantes Q1 e Q3 do Diagrama de Espalhamento. Este cruzamento de dados teve o objetivo de identificar as zonas com mais significância estatística, dentre aquelas que se comportam com padrões semelhantes ao de seus vizinhos contíguos. Portanto, os clusters resultantes da análise espacial dos dados do LISA *Map,* correspondentes aos dados plotados no Diagrama de Espalhamento, são representados por 15 zonas significativas no quadrante Q1 apresentadas no Gráfico 26 e Figura 21 e 12 zonas significativas no quadrante Q3 apresentadas no Gráfico 27 e Figura 22.

Gráfico 26– Diagrama de Espalhamento

Zonas de Cluster (Quadrante Q1)

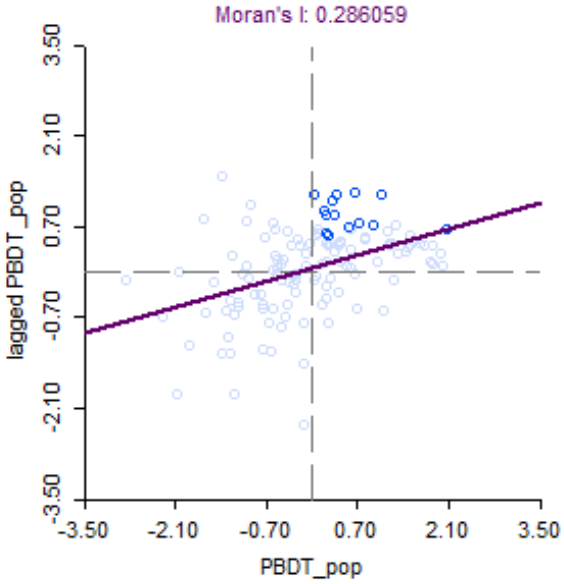


Figura 21 – Mapa RMC

Zonas de Cluster (Quadrante Q1)

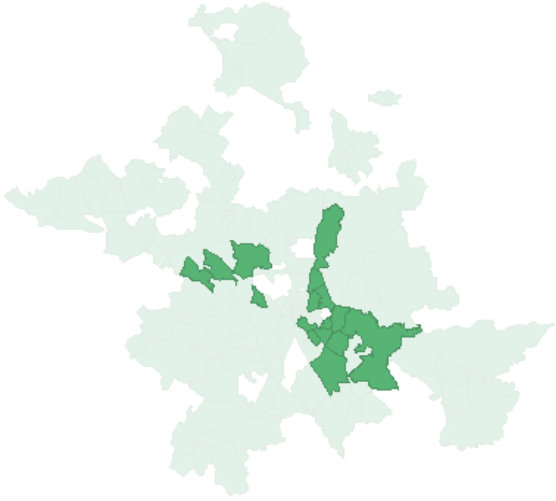


Gráfico 27 – Diagrama de Espalhamento

Zonas de Cluster (Quadrante Q3)

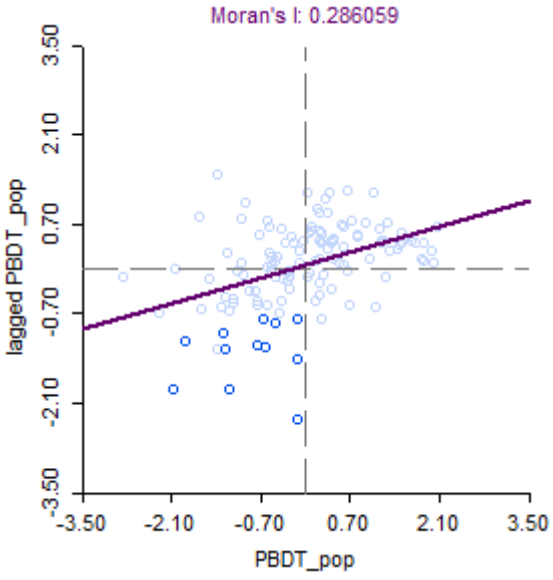
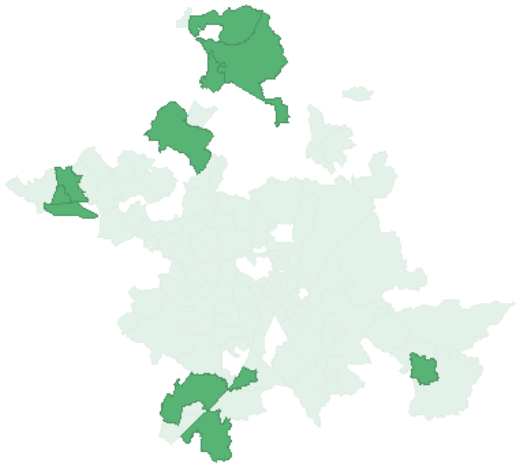


Figura 22 – Mapa RMC

Zonas de Cluster (Quadrante Q3)



9.4. MODELO DE PRODUÇÃO COM *CLUSTERS* ESPACIAIS

A presente etapa do desenvolvimento do trabalho contemplou a elaboração do modelo de produção de viagens de regressões múltiplas, baseado nos clusters espaciais definidos no capítulo anterior.

9.4.1. Seleção das variáveis do modelo

A escolha das variáveis explicativas do modelo envolveu a construção de de uma matriz de correlação Pearson entre cada variável explicativa e a variável resposta P_BDT. O critério para seleção das variáveis candidatas a integrar o modelo foram os mesmos utilizados na elaboração do modelo clássico. Portanto, foram construídas duas matrizes de correlação, sendo que a primeira foi utilizada para definir as variáveis mais correlacionadas com a variável resposta P_BDT e a segunda serviu de base para a definição das variáveis explicativas com menor colinearidade.

As variáveis que atenderam a esses critérios foram destacadas na cor amarelo da matriz da Tabela 12 seguir. As demais variáveis que não atenderam aos critérios de seleção foram descartadas das análises subsequentes. Portanto, as variáveis $Populacao, Pop\_5\_17, Pop\_18\_25, Frota\_C, e\ RendaR\$$, foram pré-selecionadas como candidatas para compor o modelo por apresentarem correlação superior a 0,5 com a variável resposta P_BDT e $p - value < 0{,}05$. As demais variáveis $Frota\_A, Frota\_B, Frota\_D, Frota\_E\ e\ Frota\_F$ apresentam correlação inferior ao estabelecido no critério de correlação, além de não apresentarem significância estatística e foram descartadas.

Tabela 12 - Correlações entre variáveis explicativas e viagens produzidas

	Populacao	Pop_5_17	Pop18_25	Frota_A	Frota_B	Frota_C	Frota_D	Frota_E	Frota_F	RendaR$
P_BDT	,990**	,934**	,935**	0,097	0,408	,558*	0,367	0,071	0,097	,740**

Em seguia foi elaborada a matriz de correlação entre as variáveis pré-selecionadas para identificar aquelas que apresentavam menores correlações com as demais, evitando elaborar um modelo com variáveis correlacionadas que explicam a mesma porção da variabilidade da variável resposta P_BDT . Essa matriz é apresentada na Tabela 13.

Tabela 13 - Correlações Pearson e níveis de significância - variáveis explicativas

	Populacao	Pop_5_17	Pop18_25	Frota_C	RendaR$
Populacao	1	,951**	,965**	,544*	,737**
Pop_5_17	,951**	1	,948**	0,404	,540*
Pop18_25	,965**	,948**	1	0,437	,647**
Frota_C	,544*	0,404	0,437	1	,504*
RendaR$	,737**	,540*	,647**	,504*	1
** Correlação significante no nível de p = 1%					
* Correlação significante no nível de p = 5%					

Observa-se que as variáveis que apresentam as menores correlações com as demais são Pop_5_17 e $Frota\_C$, e seus respectivos valores das correlações Pearson são $R = 0{,}404$ e $R = 0{,}437$. Além disso, a correlação entre essas variáveis situa-se em um nível menor que 0,5 e, portanto, é possível considera-las conjuntamente para fazer parte do modelo. Portanto, as variáveis explicativas testadas no modelo foram: Pop_5_17 e $Frota\_C$.

9.4.2. Elaboração do modelo clássico

Com auxílio do software SPSS foi possível realizar as regressões lineares pelo método dos mínimos quadrados, que resultou nos parâmetros: $\hat{\beta}_0$, $\hat{\beta}_{\text{Pop\_5\_17}}$, $\hat{\beta}_{\text{Frota\_C}}$. Assim como no modelo clássico, foram calculados os principais indicadores para interpretação dos parâmetros calculados dos resultados obtidos: Coeficiente de Determinação R^2 e $R^2 ajustado$, $Teste\ F$ e $Teste\ t$.

A Tabela 14 representa os resultados dos coeficientes de determinação do modelo. Observa-se que o $R^2 ajustado$ não apresentou uma influência muito

significativa com relação ao R^2 porém, ainda sim, por recomendação da literatura foi utilizado para minimizar os efeitos de inflação do R^2 . O $R^2 ajustado = 0{,}912$ corresponde a um valor elevado dessa medida e indica que a variabilidade dos valores de P_BDT, podendo-se afirmar que a variabilidade dessa variável pode ser atribuída à variabilidade das duas variáveis explicativas Pop_5_17 e Frota_C em conjunto, cada uma correspondendo a uma parcela da explicação de P_BDT.

Tabela 14 – Coeficiente de Determinação (R^2 e $R^2 ajustado$)

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,955[a]	,912	,898	418,672

a. Predictors: (Constant), Frota_C_sum, Pop_5_17_sum

b. Dependent Variable: P_BDT

Na Tabela 15 são apresentados os resultados do teste F, confirmando que seu valor atende aos critérios do F crítico.

Tabela 15 – Teste F

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	23556619,78	2	11778309,89	67,195	,000[b]
	Residual	2278723,073	13	175286,390		
	Total	25835342,86	15			

a. Dependent Variable: P_BDT

b. Predictors: (Constant), Frota_C_sum, Pop_5_17_sum

A tabela 16 apresenta os resultados da constante $\hat{\beta}_0$ e dos parâmetros $\hat{\beta}_{\text{Pop\_5\_17}}$, $\hat{\beta}_{\text{Frota\_C}}$ do modelo. Observa-se que ambas as variáveis apresentam valores positivos confirmando a lógica de acréscimo de uma unidade de P_BDT devido a acréscimos das variáveis explicativas e apresentam $p < 5\%$, satisfazendo ao critério de significância estatística dos parâmetros e garantindo que ambos têm influência na variável resposta e fazem parte do modelo.

Tabela 16 - Coeficientes das variáveis explicativas e teste t

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	277,771	217,808		1,275	,225
	Pop_5_17_sum	,192	,020	,847	9,407	,000
	Frota_C_sum	,150	,062	,216	2,402	,032

a. Dependent Variable: P_BDT

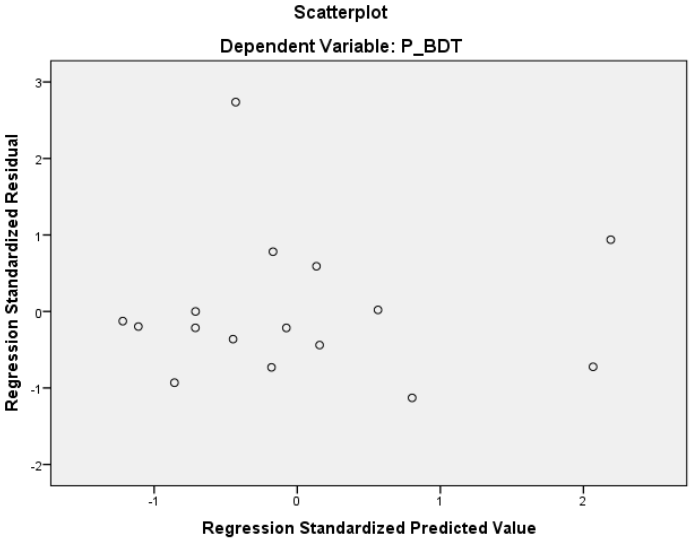
O modelo com cluster apresenta a forma funcional da expressão 12.

$$P_{BDT} = 277{,}771 + 0{,}192 * Pop_{5_{17}} + 0{,}150 * Frota\_C \qquad (12)$$

Em seguida procedeu-se a verificação do atendimento às premissas de linearidade, independência, normalidade e homocedasticidade.

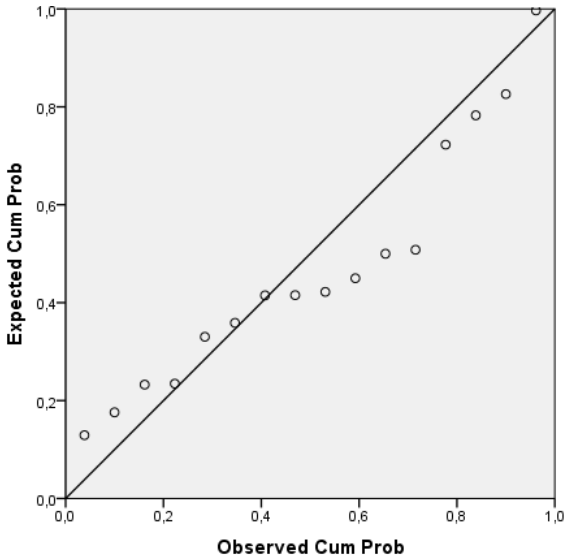
A verificação do atendimento à premissa de homocedasticidade e independência dos resíduos foi realizada por meio do gráfico de resíduos estandardizados da Gráfico 28. Observa-se que os dados se encontram completamente dispersos sem nenhuma evidência de tendência ou padrões não aleatórios, garantindo o atendimento ao critério de homocedasticidade em toda a extensão dos dados, ou seja, a média que se aumenta os totais de viagens por zona, não ocorre um aumento ou diminuição dos desvios entre dados modelados e observados.

Gráfico 28 – Gráfico de desvios estandarizados



A verificação da **premissa de normalidade** dos desvios foi realizada por meio de uma análise de um gráfico de probabilidade normal elaborado no software SPSS e representado no Gráfico 29. Observa-se que os dados não aderem completamente à diagonal principal do gráfico, podendo-se deduzir que não violam a premissa de normalidade dos resíduos.

Gráfico 29 – Gráfico de probabilidade normal



A última premissa a ser testada, **de linearidade** das variáveis explicativas com relação a variável resposta foi realizada comparando-se individualmente a correlação das mesmas. Para tanto, foram gerados gráficos de correlação parcial entre as variáveis explicativas e a variável dependente (Gráfico 30 e Gráfico 31). Observa-se que nas duas variáveis há uma forte tendência de linearidade com relação a variável resposta, atendendo completamente essa premissa.

Gráfico 30– Gráfico de probabilidade normal

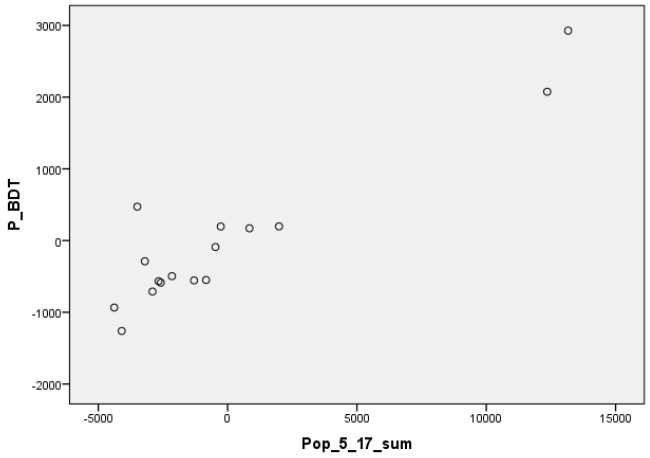
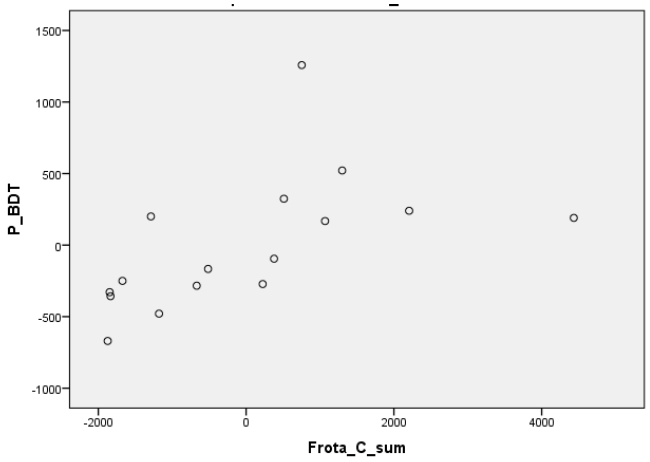


Gráfico 31- Regressão parcial Frota_C_sum



10. COMPARAÇÃO DOS RESULTADOS DOS MODELOS

A última fase do trabalho compreendeu a comparação entre o modelo clássico baseado em todas as zonas da RMC apresentado na equação 13 e o modelo elaborado a partir da identificação de *clusters* espaciais, apresentado na equação 14.

$$P\_BDT\_128 = 204{,}201 + 0{,}067 * Populacao + 0{,}255 * Frota\_E\_sum \quad (13)$$

$$P\_BDT\_16 = 277{,}771 + 0{,}192 * Pop\_5\_17 + 0{,}150 * Frota\_C \quad (14)$$

A comparação dos resultados dos modelos foi realizada aplicando-se as equações 16 e 17 no mesmo espectro de dados. Considerando que os modelos só têm validade para o conjunto de dados utilizados na sua elaboração, a única comparação possível foi submeter ambos os modelos ao conjunto de dados das zonas de *clusters* 1 que contém 16 zonas de tráfego.

Os resultados obtidos foram plotados na Tabela 17 em que são identificadas as zonas de tráfego, os valores de produções de viagens observadas, os valores do modelo clássico e de clusters aplicados nas zonas, finalmente, os desvios entre dados modelados e observados de ambos os modelos.

Dessa tabela é possível observar que o modelo baseado em clusters apresentou uma média geral praticamente idêntica à média dos dados observados, enquanto que, o modelo clássico apresentou uma diferença de 15%. Os totais de viagens produzidas por zonas também apresentam dados que comprovam que o modelo baseado em clusters apresenta diferença de apenas 0,11%. Observa-se ainda que o modelo clássico apresenta um desvio total uma vez e meia maior do que os desvios do modelo baseado em clusters.

Analisando-se os dados observados frente aos dados modelados de ambos os modelos foi possível elabora a tabela 17.

Tabela 17 – Comparações entre modelos

Zonas	P_BDT obs	P_BDT_128_C1	P_BDT_16_C1	Desvios (V1 - V0)	Ddesvios (V2 - V0)
9	2.773	2.207	1.627	566	1.146
11	2.700	2.451	3.171	249	471
12	1.981	2.093	2.070	112	89
20	2.582	2.357	2.332	225	250
21	2.281	1.754	1.953	527	328
23	1.452	1.745	1.603	293	151
24	585	484	638	101	53
33	1.185	1.244	1.273	59	88
56	1.635	1.714	1.939	79	304
65	1.276	1.027	1.274	249	2
70	701	765	1.090	64	389
71	2.882	2.322	2.870	560	12
73	2.178	1.835	2.359	343	181
108	4.454	3.249	4.748	1.205	294
115	5.305	3.651	4.903	1.654	402
146	692	574	774	118	82
Total	34.662	29.472	34.625	6.404	4.243
Média	2.166	1.842	2.164	400	265

O Gráfico 32 e o Gráfico 33 apresentam possibilitam comparar respectivamente os dados modelados e observados de ambos os modelos. Observa-se que o modelo baseado em clusters apresenta maior aderência aos dados observados em toda a extensão do gráfico, enquanto que o modelo baseado nas 128 zonas de tráfego apresenta um padrão de afastamento dos dados observados (para baixo) à medida em que os valores das viagens produzidas crescem. Portanto é possível concluir que a aplicação do modelo baseado em cluster resultou em uma melhor adesão aos dados observados quando comparado com o modelo baseado nas 128 zonas da RMC.

Gráfico 32– Modelo cluster vs observado

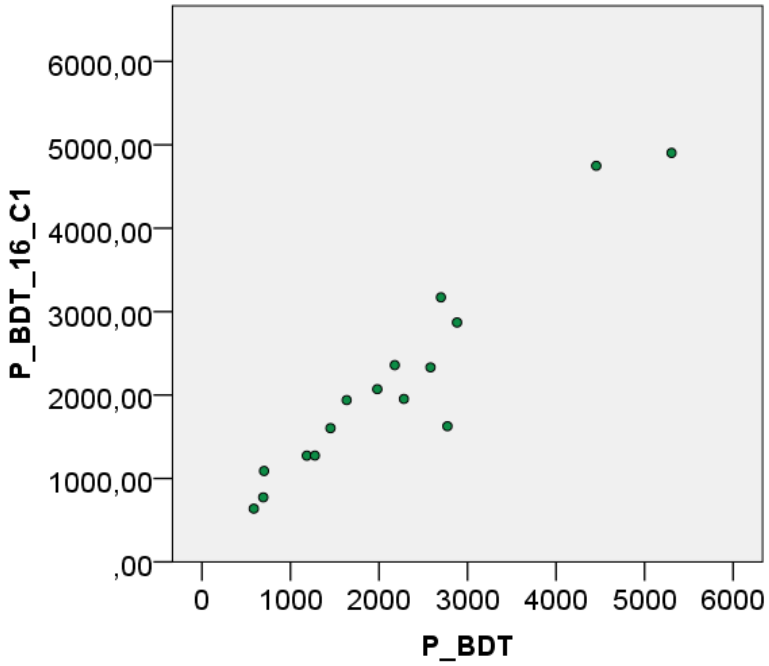
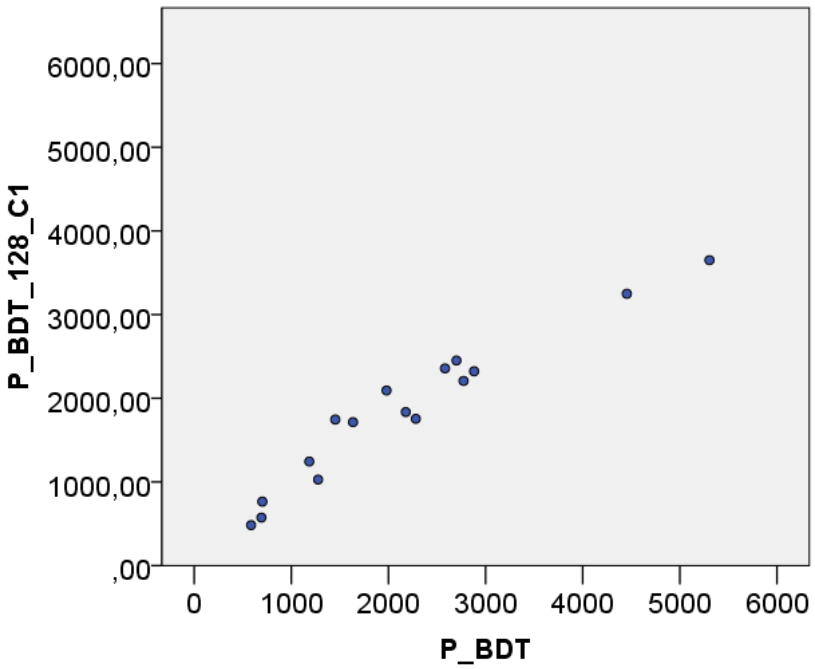


Gráfico 33 – Modelo clássico vs observado



11. Conclusões e Recomendações

A aplicação de técnicas de análise espacial revelaram que a organização espacial das produções de viagens na RMC ocorrem de acordo com padrões identificáveis de dependência espacial e que o desenvolvimento de modelos baseados em conjuntos de zonas que apresentam essa características resultam em desvios menos discrepantes entre dados modelados e observados do que os modelos clássicos, além de atenderem mais adequadamente às premissas de regressões lineares quando comparados com os resultados obtidos pela aplicação do modelo clássico.

Outra conclusão importante é que a organização espacial mencionada anteriormente não pode ser captada pelos modelos tradicionais, pois consideram que os efeitos observados em qualquer zona são os mesmos em qualquer ponto do espaço, ou seja, procuram representar o comportamento médio de todas as zonas de tráfego de uma determinada área de estudo.

Portanto, quanto maior o nível de complexidade e heterogeneidade entre as zonas de uma determinada área de estudo, menos recomendável será o uso de técnicas tradicionais.

Destaca-se ainda, que os resultados encontrados no modelo baseado em dependência espacial são diretamente influenciados pela escolha da matriz de proximidade. A liberdade dessa escolha possibilita testes diversos até encontrar a melhor representação da maneira como as zonas estão relacionadas, porém seu uso deve sempre estar embasado criteriosamente pela teoria.

Para futuras pesquisas sugere-se os seguintes temas:

- o estudo de obtenção de *cluster*s por análise multivariada, o qual vem sendo estudado na Universidade de Chicago.

- a análise de *cluster*s espaciais para obtenção de zoneamento homogêneo, o qual servirá de apoio ou substituirá os métodos empíricos que procuram identificar zonas com homogeneidade de acordo com conjuntos de variáveis.

- a comparação entre análise de *cluster*s tradicional com análise de *cluster*s considerando a estrutura espacial.

- um estudo de *cluster*s espaciais com atrações de viagens calculadas com as técnicas convencionais.

- investigar as causas que implicam na formação dos *cluster*s espaciais.

Referências Bibliográficas

Advances in Spatial Science. [s.l: s.n.].

AGRESTI, A.; FINLAY, B. **Métodos Estatísticos para as Ciências Sociais**. 4. ed. Porto Alegre: Penso, 2012.

ALDSTADT, J. Spatial Clustering. *In*: FISCHER, M. M.; GETIS, A. (Eds.). . **Handbook of applied spatial analysis**. London: Springer, 2010. p. 801.

ANDY, F. **Descobrindo a Estatistica Usando o SPSS**. 2th. ed. [s.l.] Artmed Editora Ltda, 2009.

ANSELIN, L. **Spatial Econometrics: Methods and Models**. [s.l: s.n.].

___. SPATIAL DATA ANALYSIS WITH GIS : AN INTRODUCTION TO APPLICATION IN THE SOCIAL SCIENCES Technical Report 92-10. **Systems Research**, n. August, 1992.

___. Local indicators of spatial association — LISA. **Geographical Analysis**, v. 27, n. 2, p. 93–115, 1995.

___. **The Moran Scatterplot as an ESDA tool to assess local instability in spatial associationSpatial analytical perspectives on GIS**, 1996.

ANSELIN, L. An introduction to EDA with GeoDa. **Spatial Analysis Laboratory (SAL). Department of …**, n. Figure 1, p. 1–20, 2003.

ANSELIN, L. Exploring Spatial Data with GeoDa: A Workbook. **Geography**, p. 244, 2005.

___. Thirty years of spatial econometrics. **Papers in Regional Science**, v. 89, n. 1, p. 3–25, 2010.

___. **Spatial Weights**. Disponível em: <https://geodacenter.asu.edu/spatial-weights-4>. Acesso em: 23 jul. 2015.

ANSELIN, L.; GETIS, A. Spatial Statistical Analysis and Geographic

Information Systems. *In*: **Advances in Spatial Science**. [s.l.] Springer, 2007. .

AREAL, F. J.; BALCOMBE, K.; TIFFIN, R. Integrating spatial dependence into Stochastic Frontier Analysis. **Australian Journal of Agricultural and Resource Economics**, v. 56, n. 4, p. 521–541, 2012.

BARONIO, A.; VIANCO, A.; RABANAL, C. Una Introducción a La Econometría Espacial. p. 33, 2012.

BOYCE, D.; WILLIAMS, H. **Forecasting Urban Travel**. Northampton, MA, USA: Edward Elgar Publishing Limited, 2015.

BRAGA, A. *et al.* Dependência espacial utilizando análise de dados de área aplicada na Mesorregião Metropolitana de Belo Horizonte por meio do Indicador Econômico. **Simpósio Nacional de Probabilidade e Estatística**, 2010.

BRIGGS, R. **Spatial Autocorrelation Concepts**. Disponível em: <http://www.utdallas.edu/~briggs/>.

___. **Global Measures of Spatial Autocorrelation**. Disponível em: <http://www.utdallas.edu/~briggs/>.

___. **Local Measures of Spatial Autocorrelation**. Disponível em: <http://www.utdallas.edu/~briggs/>.

BRUTON, M. **Introduction to transportation planning**. [s.l: s.n.].

BURTON, I. The Quantitative Revolution and Theoretical Geography. **The Canadian Geographer/Le Géographe canadien**, v. 7, n. 4, p. 151–162, 1963.

CÂMARA, G. *et al.* Cap 1 - Análise Espacial e Groprocessamento. **Análise Espacial de Dados Geográficos**, p. 26, 2004.

CARVALHO, M. S. *et al.* Cap 5 - Análise de Dados de Área. *In*: EMBRAPA (Ed.). . **Análise Espacial de Dados Geográficos**. Planaltina, DF: [s.n.]. v. 0p. 209.

CASTRO, M. C. DE; SAWYER, D. O.; SINGER, B. H. Spatial patterns of

malaria in the Amazon: Implications for surveillance and targeted interventions. **Health and Place**, v. 13, n. 2, p. 368–380, 2007.

DARMOFAL, D. **Spatial Analysis for the Social Sciences**. New York: Cambridge University Press, 2015.

FARBER, S. **GEOG 3020 Lecture 23-4 Spatial Autocorrelation**. Disponível em: <https://www.youtube.com/watch?v=O2W9IbVlUbE>.

___. **GEOG 3020 Lecture 23-9 Spatial Autocorrelation**. Disponível em: <https://www.youtube.com/watch?v=vhuh1v3UPvA>. Acesso em: 9 out. 2017.

FÁVERO, L. P. *et al.* **Análise de Dados - Modelagem Multivariada para Tomada de Decisões**. 10. ed. São Paulo: Elsevier Ltd, 2009.

FISHCER, M. M.; GETIS, A. **Handbook of applied spatial analysis**. London: [s.n.].

FRAME, B. *et al.* **Progress in Spatial Analysis**. [s.l: s.n.].

FURNISH, P.; WIGNALL, D. Making the most of models: Using models to develop more effective transport policies and strategies. **Concrete International**, v. 2, p. 1–46, 2009.

GEARY, R. The contiguity ratio and statistical maping. **The Incorporated Statistician**, v. 5, n. 3, p. 115–145., 1954.

GETIS, A. Reflections on spatial autocorrelation. **Regional Science and Urban Economics**, v. 37, n. 4, p. 491–496, 2007.

___. A history of the concept of spatial autocorrelation: A geographer's perspective. **Geographical Analysis**, v. 40, n. 3, p. 297–309, 2008.

___. Spatial Weights Matrices. v. 41, p. 404–410, 2009.

GETIS, A.; ALDSTADT, J. Constructing the Spatial Weights Matrix Using a Local Statistic. **Geographical Analysis**, 2009.

GOODCHILD, M. *et al.* Integrating GIS and spatial data analysis: problems and possibilities. **International Journal of Geographical Information Systems**, v. 6, n. 5, p. 407–423, 1992.

GRIFFITH, D. A. Spatial statistics: A quantitative geographer's perspective. **Spatial Statistics**, v. 1, p. 3–15, 2012.

HAIR, JOSEPH, F. **Multivariate Data Analysis**. 6th. ed. [s.l: s.n.].

HENRIQUE, C. S. Diagnóstico espacial da mobilidade e da acessibilidade dos usuários do sistema integrado de transporte de Fortaleza. v. Mestrado, p. 165, 2004.

HUBERT, L. J.; GOLLEDGE, R. G.; COSTANZO, C. M. Generalized Procedures for Evaluating Spatial Autocorrelation. **Geographical Analysis**, v. 13, n. 3, p. 224–233, 1981.

JACQUEZ, G. M. 22 Spatial Cluster Analysis. p. 395–416, 2008.

KELEJIAN, H. H.; ROBINSON, D. P. Spatial autocorrelation. **Regional Science and Urban Economics**, v. 22, n. 3, p. 317–331, 1992.

KENNEDY, P. **A Guide to Econometrics**. [s.l: s.n.]. v. 2

KREMPI, A. P. **Explorando recursos de estatística espacial para análise da acessibilidade da cidade de bauru**. [s.l: s.n.].

KUHN, W. **Spatial Concepts**. Disponível em: <http://spatial.ucsb.edu/spatial-concepts/>.

LAMEGO, M. O IBGE e a geografia quantitativa brasileira: v. 3, 2014.

LESAGE, J. P. Spatial econometrics. **A companion to theoretical econometrics**, p. 273, 1998.

LLOYD, C. D. **Local Models for Spatial Analysis**. 2. ed. New York: CRC Press, 2007.

LOPES, S. B. **Efeitos da dependência espacial em modelos de previsão de demanda por transporte**. [s.l.] Universidade de São paulo, 2005.

LOPES, S.; BRONDINO, N.; SILVA, A. DA. GIS-Based Analytical Tools for Transport Planning: Spatial Regression Models for Transportation Demand Forecast. **ISPRS International Journal of Geo-Information**, v. 3, n. 2, p. 565–583, 2014.

MARÔCO, J. Analise estatistica com o SPSS Statistics. *In*: **Analise e Gestao da Informacao**. [s.l: s.n.]. p. 990.

MARTIN, D. J. Spatial representation: the social scientist's perspective. **Geographical Information Systems, Volume 1: Principles and Technical Issues**, v. 1, p. 71–80, 1999.

MCNALLY, M. G. The Four Step Model. **Transportation**, n. UCI-ITS-AS-WP-00-51, p. 35–52, 2007.

MENDONÇA, A. C. DE. **Desenvolvimento de um Modelo de Previsão da Demanda de Passageiros do Transporte Rodoviário Interestadual Utilizando Regressão com Efeitos Espaciais Locais**. [s.l: s.n.].

MORAN, P. A. . The Interpretation of Statistical Maps. **Journal of the Royal Statistic Society**, v. 10, p. 243–251, 1948.

MORENO, R.; VAYÁ, E. Econometría espacial; nuevas técnicas para el análisis regional: una aplicación a las regiones europeas. **Investigaciones Regionales**, v. 1, p. 83–106, 2002.

NIEMEIER, D. A.; SONG, B. Urban Travel Demand Modeling Debbie. *In*: **Transportation Planning Handbook**. 3. ed. Washington, D.C.: Institute of Transportation Engineers, 2009. p. 1067.

ORD, J. K.; GETIS, A. Local Spatial Autocorrelation Statistics: Distributional Issues and an Application. **Geographical Analysis**, v. 27, n. 4, p. 286–306, 1995.

ORD, J. K.; GETIS, A. Local spatial heteroscedasticity (LOSH). **Annals of Regional Science**, v. 48, n. 2, p. 529–539, 2012.

ORTUZAR, J. DE D.; WILLUMSEN, L. G. **Modelling Transport**. 4th. ed. [s.l.] John Wiley and Sons, Ltd., 2011.

OYANA, T. J.; MARGAI, F. **Spatial Analysis: Statistics, Visualization, and Computational Methods**. New York: CRC Press, 2015.

PÁEZ. Spatial statistics for urban analysis: A review of techniques with examples. **GeoJournal**, v. 61, n. 1, p. 53, 2005.

PAIVA, W. DE L.; KHAN, A. S. **DEPENDÊNCIA ESPACIAL E EMPREGO FORMAL: O que é possível afirmar para indústria cearense?** [s.l.] Universidade Federal do Ceará saeed@ufc.br, 2011.

PING, J. L. *et al.* Exploring spatial dependence of cotton yield using global and local autocorrelation statistics. **Field Crops Research**, v. 89, n. 2–3, p. 219–236, 2004.

POULIOU, T.; ELLIOTT, S. J. An exploratory spatial analysis of overweight and obesity in Canada. **Preventive Medicine**, v. 48, n. 4, p. 362–367, 2009.

ROCHA, S. **ANÁLISE DE GERAÇÃO DE VIAGENS URBANAS POR TRANSPORTE COLETIVO ATRAVÉS DE TÉCNICAS DE GEOESTATÍSTICA**. [s.l.] Universidade Federal da Bahia, 2014.

SANTOS, L. DOS. **Análise dos Acidentes de Trânsito do Município de São Carlos Utilizando Sistema de Informações Geográficas – SIG e Ferramentas de Estatística Espacial**. [s.l.] Universidade Federal de São Carlos, 2006.

SANTOS, M. S.; SOUZA, W. V. Análise Espacial de Dados de Áreas. *In*: **Introdução à Estatistica Espacial para a Saúde Pública**. Brasiília: Ministério da Saúde, 2007. v. 3p. 124.

SILVA, A. R. Universidade De Brasília. 2006.

SOUZA, L. A. DE. **ANÁLISE ESPACIAL EM MODELOS DE GERAÇÃO DE VIAGENS Luiz**. [s.l.] UFRJ, 2013.

Spatial Autocorrelation. Disponível em: <https://geodacenter.asu.edu/drupal_files/2010-27.pdf>. Acesso em: 21 maio. 2015.

STM. **Pesquisa Origem-Destino 2011 Região Metropolitana de Campinas - Síntese dos Resultados Pesquisas Domiciliar e Cordon Line**. [s.l.] Secretaria dos Transportes Metropolitanos - STM, 2012. Disponível em: <http://www.stm.sp.gov.br/odrmc/images/stories/ODRMC_2011_sintese.pdf>.

SWEENEY, S.; I, L. M. **Introduction to Spatial Analysis**. Disponível em: <http://www.geog.ucsb.edu/~sweeney/g172/G172_lattice_notes.pdf>. Acesso em: 26 mar. 2015.

TOBLER, W. R. A computer movie simulating urban growth in the Detroit region. **Economic Geography**, v. 26, p. 234–240, 1970.

TOBLER, W. R. A Computer movie simulating urban growth in the Detroit region. **Economic Geography, Vol. 46, Supplement: Poroceeding. Internationa Geographical**, v. 13, n. 332, p. 462–465, 1970.

VASCONCELOS, V. **Autocorrelação espacial**. Disponível em: <https://www.youtube.com/watch?v=twibfxJpOrY>.

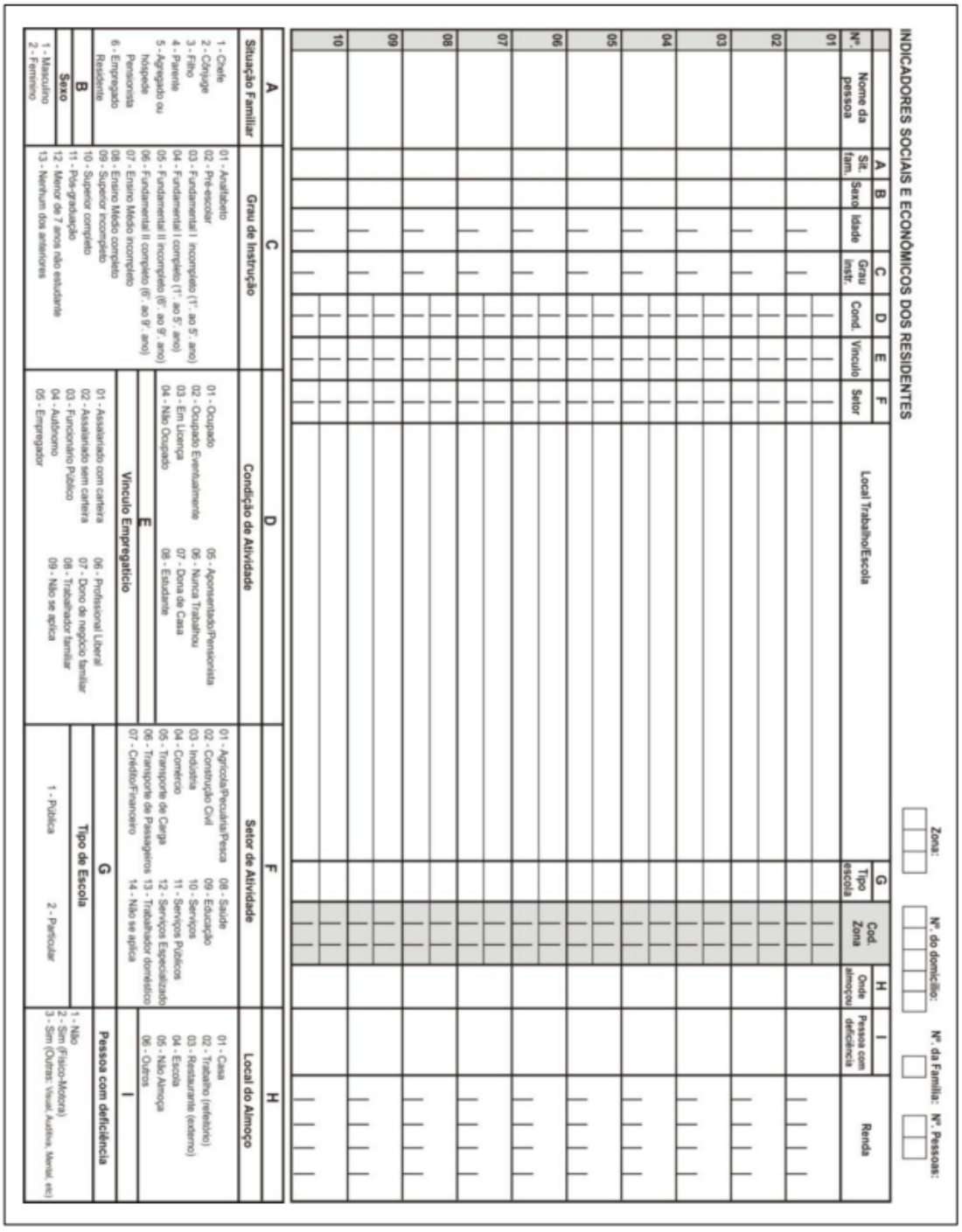
VITON, P. A. Notes on Spatial Econometric Models. **City and regional planning**, v. 870, n. 3, p. 9–10, 2010.

WALDHÖR, T. The spatial autocorrelation coefficient Moran's I under heteroscedasticity. **Statistics in medicine**, v. 15, n. 7–9, p. 887–892, 1996.

WEINER, E. **Urban Transportation Planning in the United States**. [s.l: s.n.].

___. **Urban Transportation Planning in the United States**. [s.l: s.n.].

ANEXO 1: Questionário de domicílios



INDICADORES SOCIAIS E ECONÔMICOS DOS RESIDENTES

Zona: ☐☐ Nº. do domicílio: ☐☐☐☐ Nº. da Família: ☐ Nº. Pessoas: ☐☐

Nº.	Nome da pessoa	A Sit. fam.	B Sexo	Idade	C Grau instr.	D Cond.	E Vínculo	F Setor	Local Trabalho/Escola	G Tipo escola	Cod. Zona	H Onde almoçou	I Pessoa com deficiência	Renda
01														
02														
03														
04														
05														
06														
07														
08														
09														
10														

A — Situação Familiar

1 - Chefe
2 - Cônjuge
3 - Filho
4 - Parente
5 - Agregado ou hóspede
6 - Empregado Residente
7 - Pensionista

B — Sexo

1 - Masculino
2 - Feminino

C — Grau de Instrução

01 - Analfabeto
02 - Pré-escolar
03 - Fundamental I incompleto (1ª. ao 5º. ano)
04 - Fundamental I completo (1ª. ao 5º. ano)
05 - Fundamental II incompleto (6º. ao 9º. ano)
06 - Fundamental II completo (6º. ao 9º. ano)
07 - Ensino Médio incompleto
08 - Ensino Médio completo
09 - Superior incompleto
10 - Superior completo
11 - Pós-graduação
12 - Menor de 7 anos não estudante
13 - Nenhum dos anteriores

D — Condição da Atividade

01 - Ocupado
02 - Ocupado Eventualmente
03 - Em Licença
04 - Não Ocupado
05 - Aposentado/Pensionista
06 - Nunca Trabalhou
07 - Dona de Casa
08 - Estudante

E — Vínculo Empregatício

01 - Assalariado com carteira
02 - Assalariado sem carteira
03 - Funcionário Público
04 - Autônomo
05 - Empregador
06 - Profissional Liberal
07 - Dono de negócio familiar
08 - Trabalhador familiar
09 - Não se aplica

F — Setor de Atividade

01 - Agrícola/Pecuária/Pesca
02 - Construção Civil
03 - Indústria
04 - Comércio
05 - Transporte de Carga
06 - Transporte de Passageiros
07 - Crédito/Financeiro
08 - Saúde
09 - Educação
10 - Serviços
11 - Serviços Públicos
12 - Serviços Especializados
13 - Trabalhador doméstico
14 - Não se aplica

G — Tipo de Escola

1 - Pública
2 - Particular

H — Local do Almoço

01 - Casa
02 - Trabalho (refeitório)
03 - Restaurante (externo)
04 - Escola
05 - Não Almoça
06 - Outros

I — Pessoa com deficiência

1 - Não
2 - Sim (Físico-Motora)
3 - Sim (Outras: Visual, Auditiva, Mental, etc)

ANEXO 2: Questionário socioeconômico

Pesquisa Origem-Destino 2011
Região Metropolitana de Campinas

RMC OD 2011

ENDEREÇO: _____

Zona: | Nº do Domicílio: | Nº da Família: | Pesquisador: | Telefone do domicílio (se houver)

1ª. visita: ____/____/____	2ª. visita: ____/____/____	3ª. visita: ____/____/____
Dia da Semana: _____	Dia da Semana: _____	Dia da Semana: _____
Hora: _____	Hora: _____	Hora: _____
Início da entrev.: ☐ Sim ☐ Não	Início da entrev.: ☐ Sim ☐ Não	Início da entrev.: ☐ Sim ☐ Não

Resultado do Domicílio

1 - Não encontrado
2 - Em obras
3 - Desocupado
4 - Lazer
5 - Recusa
6 - Ninguém em casa
7 - Ausência do responsável
8 - Mudança atividade
9 - Completo com viagens
10 - Completo sem viagens
11 - Incompleto

Nº do Domicílio:
Substituição:

Observações: _____

Supervisor nº ☐☐☐ Data: ____/____/____ Assinatura: _____

Codificador nº ☐☐☐ Data: ____/____/____ Assinatura: _____

Quantidade de Famílias ☐

Tipo de Domicílio
1. Particular
2. Coletivo
☐

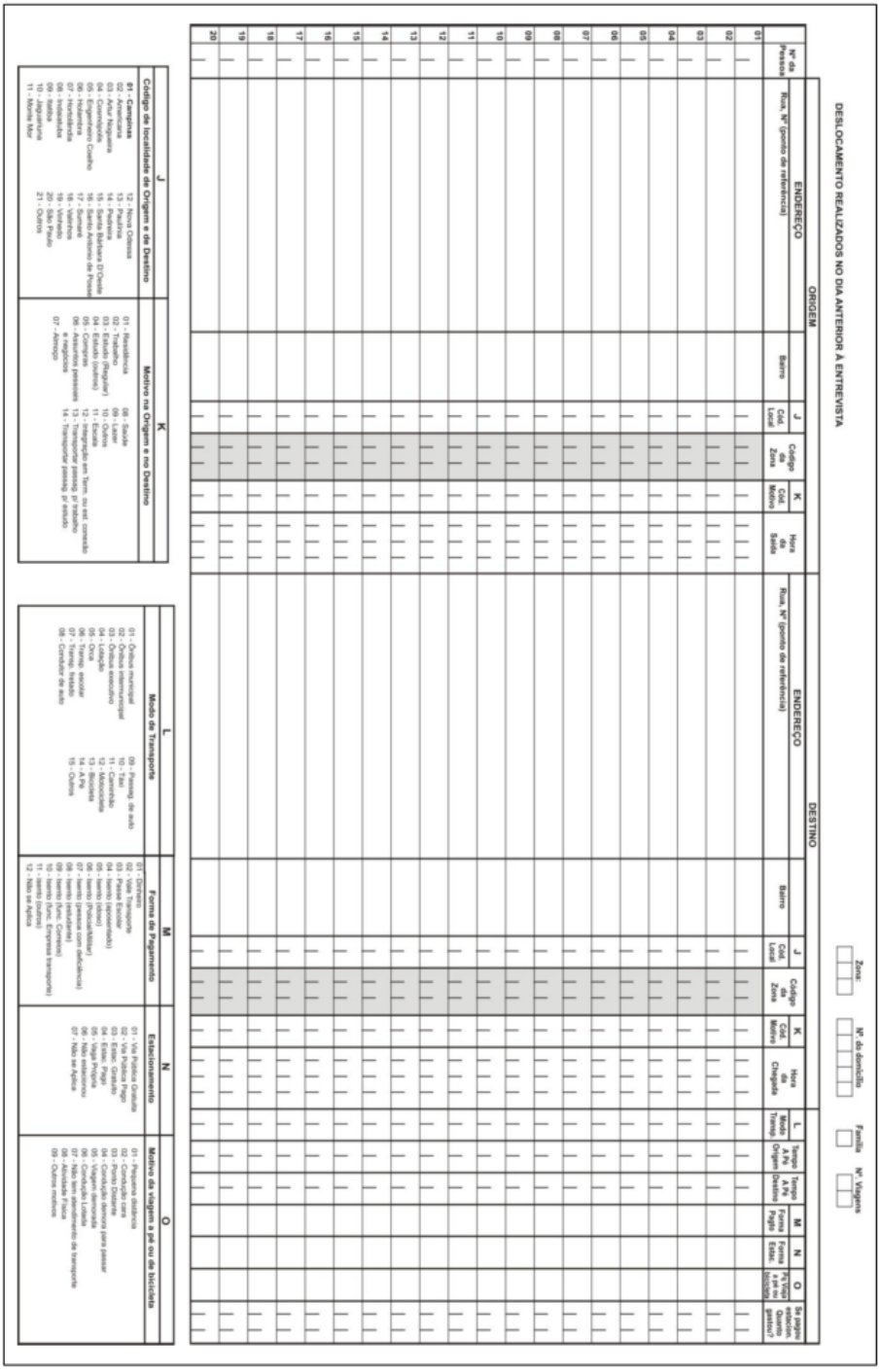
Condição de Moradia
1. Própria
2. Alugada
3. Cedida
4. Invadida
☐

Ind. Conforto Familiar	
Televisão em cores	
Rádio	
Banheiro	
Automóvel	
Empregada Mensalista	
Máquina de Lavar	
Videocassete e/ou DVD	
Geladeira	
Freezer (aparelho independente ou parte da geladeira)	

Ano Fabricação Autos
☐☐
☐☐
☐☐
☐☐

REALIZAÇÃO: OFICINA ENGENHEIROS CONSULTORES ASSOCIADOS LTDA.

ANEXO 3: Questionário de viagens



ANEXO 4: Viagens base domiciliar motivo trabalho hora pico manhã

Zonas	Produção BDT	Produção BDE+BDO
1	2.680	942
2	2.023	711
3	1.952	980
4	739	248
5	1.188	417
6	716	344
7	938	330
8	3.407	1.197
9	2.813	1.181
10	2.753	908
11	2.709	939
12	1.988	1.172
13	2.590	1.000
14	2.121	1.020
15	2.634	946
16	2.769	1.069
17	630	233
18	2.691	1.351
19	58	21
20	2.652	712
21	2.388	766
22	260	96
23	1.576	757
24	588	44
25	749	367
26	40	15
27	3.739	1.876
28	1.989	998
29	4.379	1.470
30	3.404	1.196
31	3.172	1.764
32	2.021	772
33	1.056	380
34	125	46
35	295	108
36	641	354
37	4.255	1.786
38	532	234
39	274	101
40	804	544
41	18	7
42	236	104
43	3.160	693
44	111	41
45	110	40
46	93	34
47	7.532	1.482
48	7.749	1.700
49	2.857	1.659
50	3.638	881

Zonas	Produção BDT	Produção BDE+BDO
51	1.028	169
52	1.749	856
53	1.230	542
54	3.505	1.176
55	1.078	178
56	1.405	441
57	82	30
58	767	255
59	701	750
60	1.440	975
61	577	254
62	60	22
63	3	1
64	1.055	357
65	1.225	621
66	706	260
67	76	28
68	2.454	1.081
69	3.408	1.093
70	699	257
71	3.053	865
72	478	162
73	2.186	382
74	3.167	872
75	776	244
76	2.140	718
77	227	83
78	947	348
79	273	100
80	299	110
81	152	56
82	2.007	644
83	655	222
84	1.597	502
85	945	416
86	1.753	514
87	377	139
88	2.175	717
89	3.433	922
90	1.076	545
91	340	125
92	15	6
93	236	87
94	3.066	1.705
95	256	94
96	3.486	1.095
97	113	42
98	1.420	416
99	1.620	793
100	1.026	347

s	Produção BDT	Produção BDE+BDO
101	38	14
102	181	67
103	3.831	928
104	2.960	815
105	2.502	840
106	3.348	1.236
107	1.315	372
108	4.316	1.159
109	2.734	1.148
110	2.149	520
111	2.730	661
112	1.858	584
113	2.988	656
114	3.298	870
115	5.323	905
116	130	48
117	314	115
118	2.565	752
119	51	19
120	0	0
121	60	22
122	108	40
123	123	45
124	2.135	824
125	1.681	929
126	5.104	1.971
127	3.233	1.037
128	5.170	1.389
129	2.691	652
130	2.420	776
131	1.952	553
132	1.898	375
133	1.003	315
134	148	55
135	3.347	810
136	401	135
137	243	89
138	218	80
139	82	30
140	2.121	1.172
141	39	14
142		
143	1.453	482
144	38	14
145	124	46
146	651	220
147	243	82
148	2.387	801
149	370	125
150	2.633	578

Zonas	Produção BDT	Produção BDE+BDO
151	227	83
152	18	7
153	85	31
154	28	10
155	231	85
156	1.497	572
157	16	6
158	142	52
159	182	67
160	405	137
161	578	196
162	28	10
163	616	301
164	78	29
165	248	91
166	220	81
167	1.231	443
168	114	42
169	74	27
170	135	50
171	64	21
172	135	50
173	96	35
174	185	68
175	80	29
176	0	0
177	605	231
178	48	18
179	177	65
180	85	31
181	146	54
182	27	10
183	1.808	512
184	95	35
185	56	20

ANEXO 5: Método dos Mínimos Quadrados

A equação 18 a seguir representa a formalização matemática da minimização dos desvios mencionados entre $\hat{y}_i$ e y_i.

$$\min \sum (y_i - \hat{y}_i)^2 \qquad (15)$$

Em que:

$y_i =$ valor observado da variável resposta para a i-ésima observação

$\hat{y}_i =$ valor estimado da vari´vel dependente para a i-ésima observação

$$\hat{b}_1 = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2} = \frac{cov(x,y)}{Var\ (x)} \qquad (16)$$

$$\hat{b}_0 = \bar{y} - b_1\bar{x} \qquad (17)$$

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad (18)$$

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} \qquad (19)$$

Em que:

$\hat{b}_1 =$ *valor da variável independente para a* $i-$ *ésima observação*

$\hat{b}_0 =$ *intecepto da equação de regressão*

$x_i =$ *valor da variável independente para a* $i-$ *ésima observação*

$y_i =$ *valor da variável dependente para a* $i-$ *ésima observação*

$\bar{x} =$ *valor médio da variável independente*

$\bar{y} = valor\ médio\ da\ variável\ dependente$

$n = número\ total\ de\ observações$

$cov(x, y) = covariância\ de\ x\ e\ y =$

$Var\ (x) = variância\ de\ x$

Para modelos com duas variáveis explicativas a estimativa de b_1 é descrita conforme equação 23 a seguir:

$$\hat{b}_1 = \frac{cov(X_1, Y)Var(X_2) - cov(X_2, Y)cov(\mathrm{X}_1, X_2)}{Var(X_1)Var(X_2) - [cov(X_1, X_2)]} \quad (20)$$

ANEXO 6: Coeficiente de Determinação

Segundo Fávero *et al*. (2009), a principal medida que mede a capacidade explicativa do modelo é expressa pelo parâmetro denominado coeficiente de determinação R^2. Este parâmetro informa o quanto da variância de y pode ser creditada ao modelo de regressão e pode ser interpretado como a redução proporcional do erro pelo uso da equação de previsão linear em vez da média $\bar{y}$ (AGRESTI e FINLAY, 2012).

O cálculo do coeficiente de determinação é realizado por meio da expressão que representa a divisão entre a soma dos quadrados da regressão e a soma total dos quadrados, conforme apresentado na equação 24 a seguir.

$$R^2 = \frac{SQR}{SQ\mathrm{T}} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} \quad (21)$$

Para Marôco (2011) o valor de R^2 correspondente a um ajustamento adequado depende da análise que se deseja fazer. Em ciências exatas, valores acima de 0,9 são aceitáveis, em ciências sociais aceitam-se valores acima de 0,5.

Segundo Agresti e Finlay (2012), o R^2 tem 5 propriedades:

1. Seus valores estão entre -1 e 1.
2. O valor mínimo possível para SQU é 0, nesse caso R^2 = 1.
3. Quando a inclinação da reta dos mínimos quadrados é b = 0, o intersepto a é igua a $\bar{y}$.
4. Quanto mais próximo de 1 for o valor de R^2, mais efetiva é a linha dos mínimos quadrados comparada com a média.
5. O R^2 não depende das unidades de mensuração e ele assume o mesmo valor tanto quando x prevê y, como quanto y prevê x.