UNIVERSIDADE DE SÃO PAULO
ESCOLA DE ENGENHARIA DE SÃO CARLOS

ANABELE LINDNER

Análise desagregada de dados de demanda por transportes através de modelagem
geoestatística e tradicional

São Carlos

2015

ANABELE LINDNER

Análise desagregada de dados de demanda por transportes através de modelagem
geoestatística e tradicional

Dissertação apresentada à Escola de Engenharia de São Carlos como parte dos requisitos para a obtenção do título de Mestre em Ciências, Programa de Pós-graduação em Engenharia de Transportes.

Área de concentração: Planejamento e Operação de Sistemas de Transporte.

Orientadora: Prof. Dra. Cira Souza Pitombo

São Carlos

2015

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

L735a Lindner, Anabele
Análise desagregada de dados de demanda por
transportes através de modelagem geoestatística e
tradicional / Anabele Lindner; orientadora Cira Souza
Pitombo. São Carlos, 2015.

Dissertação (Mestrado) - Programa de Pós-Graduação
em Engenharia de Transportes e Área de Concentração em
Planejamento e Operação de Sistemas de Transporte --
Escola de Engenharia de São Carlos da Universidade de
São Paulo, 2015.

1. Geoestatística. 2. Regressão logística. 3.
Demanda por transportes. I. Título.

FOLHA DE JULGAMENTO

Candidata: Engenheira **ANABELE LINDNER**.

Título da dissertação: "Análise desagregada de dados de demanda por transportes através de modelagem geoestatística e tradicional".

Data da defesa: 23/02/2015

Comissão Julgadora:

Resultado:

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

Aprovada

Prof. Associado **Antônio Néelson Rodrigues da Silva**
(Escola de Engenharia de São Carlos/EESC)

APROVADA

Dra. **Ana Rita Gonçalves Neves Lopes Salgueiro**
(Universidade Federal do Ceará/UFC)

Aprovada

Coordenador do Programa de Pós-Graduação em Engenharia de Transportes:
Prof. Associado **Paulo César Lima Segantine**

Presidente da Comissão de Pós-Graduação:
Prof. Associado **Paulo César Lima Segantine**

“Se projetas alguma coisa, ela te sairá bem
e a luz brilhará em teus caminhos”.

Jó, 22:28

AGRADECIMENTOS

Agradeço primeiramente a Deus, pelas oportunidades concebidas em minha vida e à força contemplada na jornada.

Agradeço à minha família, especialmente aos meus pais, pelos princípios ensinados, os quais permitiram que eu me tornasse uma pessoa perseverante.

Agradeço ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pelo apoio à pesquisa.

Agradeço à professora, Cira Souza Pitombo, que se prontificou a me orientar e sempre foi muito solícita e disponível.

Agradeço às amizades que se edificaram nesta etapa. Tenho uma imensa gratidão pelos amigos Felipe Costa Bethonico, Lucas Assirati e Samille Santos Rocha, não apenas pelo companheirismo, mas também pela preocupação, solidariedade e apoio nos momentos difíceis deste Mestrado.

Agradeço ao amicíssimo Diego Fernandes Neris, que me encorajou a fazer esta Pós Graduação e me inspira com o seu sucesso.

Agradeço a Márcia Pereira de Andrade, professora fundamental na minha iniciação aos estudos da área de Engenharia de Transportes desde o período da Graduação.

RESUMO

LINDNER, A. **Análise desagregada de dados de demanda por transportes através de modelagem geoestatística e tradicional**. 2015. 106p. Dissertação (Mestrado em Ciências) - Departamento de Engenharia de Transportes, Escola de Engenharia de São Carlos, Universidade de São Paulo, São Carlos, 2015.

O conhecimento do padrão de deslocamento populacional bem como a estimativa de demanda por transportes são de fundamental importância para a tomada de decisões relativas ao planejamento urbano e de transportes. Em geral, a obtenção destas informações é realizada por modelos tradicionais como o modelo quatro etapas. Entretanto, modelos clássicos não levam em conta a dependência espacial das variáveis. A Geoestatística, valendo-se da utilização de variáveis regionalizadas, apresenta-se como uma ferramenta auxiliar capaz de modelar informações espaciais. Este trabalho tem por objetivo estimar dados desagregados de demanda por transportes através de modelagem geoestatística e tradicional. Neste estudo, a modelagem tradicional e a geoestatística puderam ser comparadas por meio de um banco de dados referente à pesquisa Origem/Destino da Região Metropolitana de São Paulo, realizada em 2007. A abordagem tradicional se baseou em um modelo de regressão enquanto que a abordagem geoestatística consistiu na estimação espacial de variáveis com base na modelagem de semivariogramas e Krigagem. Ao final do trabalho, foi possível realizar a comparação dos resultados da abordagem tradicional e geoestatística em coordenadas de valores conhecidos. Os resultados indicaram que a modelagem tradicional apontou uma taxa de acertos de 96% pelo modelo de Regressão Logística Múltipla adotada para a variável dicotômica de *preferência por modo motorizado* (variável objeto de estudo). A abordagem tradicional baseou-se na calibração de um modelo por meio de outras oito variáveis. Entretanto, a modelagem geoestatística, utilizando-se apenas das coordenadas geográficas domiciliares, resultou em 67% de taxa de acertos de previsão da variável. Isso demonstrou que, apesar de possuir menor taxa de acertos, a modelagem geoestatística, por utilizar menor número de informações para previsão

da variável, teve um resultado satisfatório e demonstra-se promissora na área de planejamento de transportes, sobretudo considerando sua habilidade de estimação em outras coordenadas geográficas além das amostradas.

Palavras-chave: Geoestatística, regressão logística, demanda por transportes.

ABSTRACT

LINDNER, A. **Disaggregated data analysis on transportation demand through traditional and geostatistical modeling**. 2015. 106p. Thesis (Master of Science) – Engineering School of São Carlos, University of São Paulo. São Carlos, 2015.

The comprehension of population displacement patterns and travel demand forecasting is crucial on making decisions related to urban transportation planning. In order to obtain this information, classic models like the sequential Four-step model are applied. However, classic models do not consider spatial location in their approach. Geostatistics is displayed as a suitable complementary instrument able to model spatial information. This work intends to forecast disaggregated data on transportation demand through traditional and geostatistical modeling. The present study compares the results from classic approach and Geostatistics through an Origin-Destination Survey dataset, carried out in São Paulo Metropolitan Area in 2007. The classic approach was based on regression models whereas Geostatistics consisted in variable spatial estimation by semivariograms modeling and Kriging. At the end of the study, a comparison between regression and geostatistical analysis was conducted through results of prediction in locations where the values of the variable are known. Results indicated that classic modeling had a 96% hit rate by Multiple Logistic Regression adopted for the dummy variable *preference for motorized travel mode* (object of study variable). Classic modeling was based on a training model using other eight predictor variables. Meanwhile, Geostatistics, using only residential geographical coordinate, resulted in a 67% hit rate for predicting the variable object of study. This demonstrates that, even though Geostatistics had lower hit rate compared to Multiple Logistic Regression, it had satisfactory outcome and proves to be a promising approach in transport planning, given that it considered less information to predict the variable, especially considering its ability of estimating in other geographical coordinates in addition to those sampled.

Keywords: Geostatistics, logistic regression, demand for transportation

LISTA DE FIGURAS

Figura 2.1: Gráfico de dispersão para variável dependente com apenas duas categorias.....	35
Figura 2.2: Curva S de probabilidade para a variável dependente binária.....	36
Figura 3.1: Etapa 1) para processo de obtenção do semivariograma.....	46
Figura 3.2: Etapa 2) para processo de obtenção do semivariograma.....	46
Figura 3.3: Etapa 3) para processo de obtenção do semivariograma.....	46
Figura 3.4: Etapa 4) para processo de obtenção do semivariograma.....	47
Figura 3.5: Parâmetros para interpretação gráfica do semivariograma.....	48
Figura 3.6: Representação de uma variável espacial isotrópica e anisotrópica.....	49
Figura 3.7: Exemplos de melhores ajustes para semivariogramas experimentais... 50	
Figura 3.8: Fluxograma para abordagem básica de calibração e validação de modelos.....	51
Figura 4.1: Região Metropolitana de São Paulo e Zonas de Tráfego.....	59
Figura 4.2: Representação dos domicílios apurados na Pesquisa O/D.....	60
Figura 4.3: Esquema simplificado do método proposto.....	62
Figura 4.4: Processo de calibração/validação.....	67
Figura 4.5: Representação da interpolação dos valores de coordenadas.....	70
Figura 5.1: Histogramas das variáveis de demanda por transporte.....	74

Figura 5.2: Semivariogramas experimentais nas direções: omnidirecional, (0° , -90°), (15° , -75°), (30° , -60°) e (45° , -45°) para <i>preferência por modo motorizado</i>	77
Figura 5.3: Semivariogramas experimentais nas direções (60° , -30°) e (75° , -15°) para <i>preferência por modo motorizado</i>	78
Figura 5.4: Amplitudes e razões das direções estudadas e elipsóide de anisotropia geométrica.....	79
Figura 5.5: Semivariogramas ajustados para a direção principal (-15°) e direção ortogonal (75°).....	79
Figura 5.6: Mapa krigado para a variável <i>preferência por modo motorizado</i>	80
Figura 5.7: Mapa da taxa de acertos para a abordagem geoestatística (validação pela média dos nós adjacentes).....	83
Figura 5.8: Mapa da taxa de acertos para a abordagem geoestatística (validação cruzada).....	84
Figura 5.9: Mapa da taxa de acertos por município para a RLM.....	87
Figura 5.10: Gráficos de comparação de taxas de acertos entre a modelagem Geoestatística (média e validação cruzada) e a modelagem por RLM.....	89
Figura 5.11: Proporção de registros por município da RMSP.....	90

LISTA DE TABELAS

Tabela 3.1: Tipos de dados e problemas em análise espacial.....	40
Tabela 4.1: Informações dos municípios da RMSP	61
Tabela 4.2: Informações utilizadas da Pesquisa O/D RMSP 2007.....	64
Tabela 4.3: Medidas descritivas para as variáveis quantitativas.....	65
Tabela 4.4: Percentual de observações da variável <i>critério de renda</i>	65
Tabela 4.5: Percentual de observações da variável <i>preferência por modo motorizado</i>	65
Tabela 4.6: Medidas estatísticas avaliadas nas modelagens.....	71
Tabela 5.1: Parâmetros para cálculo dos semivariogramas experimentais.....	76
Tabela 5.2: Resultados de validação pela média dos nós adjacentes (modelagem geoestatística).....	81
Tabela 5.3: Resultados da validação cruzada (modelagem geoestatística).....	82
Tabela 5.4: Resultados para qualidade do ajuste da RLM.....	86
Tabela 5.5: Resultados de validação para RLM.....	87

LISTA DE ANEXOS E APÊNDICE

Anexo 1: Mapa da Região Metropolitana de São Paulo 105

Apêndice 1: Resultados de taxa de acertos para a validação por município..... 106

SUMÁRIO

1. INTRODUÇÃO.....	25
1.1 PROBLEMA DE PESQUISA.....	26
1.2 JUSTIFICATIVA.....	26
1.3 OBJETIVO.....	27
1.3.1 Objetivos específicos.....	27
1.4 ESTRUTURA DO TRABALHO.....	27
2. MODELAGEM DE DEMANDA POR TRANSPORTES.....	29
2.1 CONSIDERAÇÕES INICIAIS.....	29
2.2 MODELO QUATRO ETAPAS.....	31
2.3 ESCOLHA MODAL.....	33
2.3.1 Regressão Logística.....	35
2.3.1.1 Regressão Logística Simples.....	36
2.3.1.2 Regressão Logística Múltipla.....	37
2.3.1.3 Regressão Logística Multinomial.....	38
3. ESTATÍSTICA ESPACIAL.....	39
3.1 CONCEITOS DE ESTATÍSTICA ESPACIAL.....	39
3.2 GEOESTATÍSTICA.....	42
3.2.1 Análise das variáveis regionalizadas.....	43
3.2.2 Modelagem de semivariogramas experimentais.....	44
3.2.2.1 Parâmetros para definição do semivariograma.....	44
3.2.2.2 Processo de obtenção do semivariograma.....	45

3.2.2.3 Interpretação gráfica do semivariograma.....	47
3.2.2.4 Ajuste do semivariograma experimental.....	49
3.2.3 Validação.....	51
3.2.4 Krigagem.....	54
3.2.4.1 Krigagem Indicativa (KI).....	55
3.2.5 Geoestatística e Transportes.....	57
4. MATERIAIS E MÉTODO.....	59
4.1 MATERIAIS.....	59
4.1.1 Banco de dados e área de estudo.....	59
4.1.2 <i>Software</i>	62
4.2 MÉTODO.....	62
4.1.1 Tratamento e visualização de dados.....	63
4.1.2 Escolha da variável objeto de estudo.....	66
4.1.3 Modelagem de demanda por transportes.....	66
4.1.3.1 Modelagem geoestatística.....	67
4.1.3.2 Modelagem tradicional.....	71
4.1.4 Comparação dos resultados.....	72
5. RESULTADOS E DISCUSSÕES.....	73
5.1 TRATAMENTO E VISUALIZAÇÃO DE DADOS.....	73
5.2 ESCOLHA DA VARIÁVEL OBJETO DE ESTUDO.....	74
5.3 MODELAGEM GEOESTATÍSTICA.....	75
5.3.1 Análise da estrutura espacial.....	76
5.3.2 Krigagem.....	80
5.3.3 Validação.....	81

5.3.3.4 Validação por município	83
5.4 MODELAGEM TRADICIONAL	84
5.4.1 Regressão Logística Múltipla	84
5.4.2 Validação	86
5.4.2.1 Validação por município	87
5.5 COMPARAÇÃO DOS RESULTADOS.....	88
6. CONCLUSÕES E RECOMENDAÇÕES.....	91
5.1 CONCLUSÕES.....	91
5.2 RECOMENDAÇÕES.....	95
REFERÊNCIAS BIBLIOGRÁFICAS	97
LISTA DE ANEXOS E APÊNDICE	105

1. INTRODUÇÃO

No capítulo 1 dessa dissertação, uma introdução é feita apresentando o problema da pesquisa, a justificativa do estudo, os objetivos e a estrutura do texto.

Estimativas de demanda por transportes auxiliam a tomada de decisão relativa ao planejamento urbano e de transportes. Tradicionalmente, são utilizados modelos diretos ou sequenciais para prever a demanda por transportes. O modelo mais utilizado ainda é o modelo quatro etapas, as quais são: (1) geração de viagens; (2) distribuição de viagens; (3) divisão modal e (4) alocação de tráfego.

Os modelos clássicos, usualmente, não consideram a dependência espacial de dados na análise. Diante deste fato, a inclusão de técnicas de análise espacial preenche essa lacuna e abre espaço para uma ferramenta adicional para incremento de modelagem e estimação da demanda por transportes, a partir da incorporação de localizações espaciais.

Técnicas de análise espacial diferenciam-se das técnicas tradicionais. Ao invés de serem promovidas análises estatísticas globais, ou seja, dos dados como um todo sem a preocupação da posição espacial, são realizadas análises estatísticas locais com o enfoque de expor os resultados obtidos em uma representação gráfica.

No contexto da análise espacial, a geoestatística se diferencia, pois, após a realização das estatísticas locais por meio de semivariogramas, os resultados obtidos são utilizados como instrumento para o processo de estimação denominado Krigagem. Mostra-se, então, como uma ferramenta completa, porque, além de promover a análise dos dados pré-existentes, é capaz de inferir sobre dados desconhecidos.

A aplicação da Geoestatística passou a ser mais usual na década de 1980, sendo utilizada pela lavra e prospecção mineira, agricultura de precisão, análise

espacial de crimes, cartografia, climatologia, oceanologia, geotecnia, hidrogeologia (YAMAMOTO e LANDIM, 2013). A utilização das técnicas de Geoestatística na Engenharia de Transportes é ainda incipiente. Autores como Braxmeier *et al.* (2005) e Du e Aultman-Hall (2006) iniciaram pesquisa com aplicação de técnicas de interpolação espacial de dados (Krigagem) para estimação da velocidade e do tempo de viagem.

1.1 PROBLEMA DE PESQUISA

O problema da presente pesquisa, que motivou a realização do trabalho, baseia-se no fato de que a modelagem tradicional de demanda por transportes desconsidera a localização espacial das variáveis. No entanto, sabe-se que a demanda por transportes é um problema locacional, além de socioeconômico e demográfico.

1.2 JUSTIFICATIVA

Embora os métodos tradicionais não abordem a localização espacial das variáveis na análise, a Geoestatística considera essa condição e se mostra potencialmente vantajosa para modelagem e estimação de demanda por transportes. Contudo, são poucos os trabalhos na área de estatística espacial para o planejamento de transportes. O trabalho aqui exposto procura contribuir com informações e propor método que viabilize a aplicação da Geoestatística no planejamento de transportes para que a técnica aprimore estudos já existentes. Da mesma forma, pretende-se contribuir com as adequações de dados relativos a viagens na abordagem da análise geoestatística.

Este trabalho de mestrado utilizou dados desagregados (por domicílio), os quais puderam se adequar melhor às suposições de continuidade espacial da

Geoestatística, quando comparados a dados agregados por zonas de tráfego, por exemplo.

1.3 OBJETIVO

O trabalho teve como objetivo principal estimar dados desagregados de demanda por transportes através de modelagem geoestatística e tradicional.

1.3.1 Objetivos específicos

- Obter estimativas da variável de estudo pela abordagem tradicional de análise de demanda por transportes;
- Obter a superfície espacial de estimação da variável objeto de estudo;
- Testar a adequabilidade da técnica de análise espacial por meio de comparação entre as abordagens;
- Analisar os resultados da modelagem por município da RMSP, discutindo as similaridades e diferenças.

1.4 ESTRUTURA DO TRABALHO

A dissertação de mestrado, aqui apresentada, é formada por seis capítulos, além desta introdução.

No segundo e terceiro capítulos, são apresentados embasamentos teóricos para o pleno entendimento deste trabalho. O segundo capítulo aborda alguns conceitos básicos de modelagem tradicional utilizados na análise de demanda por transportes. No terceiro capítulo, são apresentadas considerações sobre as técnicas de análise espacial de dados e conceitos relacionados à análise espacial de superfícies, conhecida como Geoestatística.

No quarto capítulo são explanados os materiais, *software* e o método utilizados para a obtenção da estimação da variável objeto de estudo a ser definida por meio de modelagem geoestatística e tradicional (regressão).

O quinto capítulo apresenta os resultados obtidos seguindo-se as etapas do método proposto, bem como as principais discussões. O sexto capítulo expõe as conclusões e recomendações.

2. MODELAGEM DE DEMANDA POR TRANSPORTES

O capítulo 2 tem por intuito oferecer ao leitor um referencial teórico acerca dos métodos tradicionais de modelagem de demanda. O texto enfatiza o modelo de quatro etapas, pois este ainda é o mais utilizado na prática. Além disso, o capítulo reserva um enfoque relacionado à escolha modal, especificamente descrevendo modelos logísticos.

2.1 CONSIDERAÇÕES INICIAIS

Um modelo é criado com a finalidade de representar um sistema. Tal representação pode ser realizada através de vários tipos de linguagem: matemática, lógica, gráfica, física (modelo reduzido).

A elaboração de um modelo objetiva a simplificação da realidade e a definição de elementos e relações relevantes.

Modelos podem ser classificados com referência às técnicas utilizadas (WILSON¹, 1967 *apud* NOVAES, 1982): i) análises estatísticas; ii) sistemas de equação; iii) simulação; iv) algoritmos ou processos heurísticos.

Tradicionalmente, o embasamento para modelagem de fenômenos urbanos e de demanda por transportes segue quatro tipos básicos de posturas científicas (NOVAES, 1982):

- a) Abordagem empírica: ajuste de relações matemáticas, comumente amparado por análise de regressão ou outro processo estatístico.
- b) Enfoque microanalítico ou comportamental: conduta geral determinada por análise de comportamento de indivíduos. Um entrave encontrado pela adoção dessa técnica está no fato de que indivíduos submetidos a mesmas condições atuam de formas diferentes.

¹ Wilson, A. (1967) Mathematical models in planning. Architectural Association Journal, 82, p. 260-65.

- c) Analogias: os elementos objetos de estudo são analisados de maneira alusória a técnicas e modelos utilizados em outros campos da ciência. Um exemplo clássico no planejamento de transportes é o modelo gravitacional.
- d) Enfoque macroanalítico ou da física social: consiste no estabelecimento de relações baseadas no comportamento individual de cada categoria e os níveis de demanda agregados, que são observados na realidade através de formulações probabilísticas.

Embora desenvolvidas no passado, essas técnicas ainda são amplamente utilizadas. Por meio do advento de recursos computacionais e matemáticos modernos e mais precisos, tais técnicas têm sido aprimoradas por meio de ferramentas, que antes eram inviáveis. Dentre essas ferramentas, pode-se citar a Análise Espacial (Regressão Espacial), Redes Neurais, Lógica Fuzzy, Algoritmos Genéticos e Autômatos Celulares e Sistemas de Informação Geográfica (SIG).

Lopes (2005) demonstrou, em seu trabalho, que a introdução de variáveis espaciais ao estudo de modelos de demanda por transporte apresentou resultados positivos, quando comparados a modelos tradicionais. Isso acontece, pois os SIGs permitem a identificação de padrões espaciais das variáveis de demanda por transportes e das variáveis socioeconômicas, enquanto as demais ferramentas restringem-se à caracterização das relações intervenientes nas variáveis modeladas. (PAVA, 2010).

Contudo, apesar de conveniente a introdução de novos conceitos e técnicas, é necessário o entendimento das abordagens tradicionais.

Estimativas de demanda por transportes auxiliam a tomada de decisão relativa ao planejamento urbano. A previsão da demanda por modelos tradicionais considera a utilização do sistema de transportes em função da dinâmica do ambiente. O modelo mais utilizado ainda é o modelo quatro etapas, descrito na próxima seção.

2.2 MODELO QUATRO ETAPAS

Um dos modelos largamente empregado no planejamento de transportes é o modelo sequencial de quatro etapas (SOUZA e D'AGOSTO, 2013; PALHARES, 2008; PEREIRA e OLIVEIRA, 2014). Apesar de ser conhecido como um modelo, se configura mais como um método ou um procedimento e sua finalidade é, principalmente, a de previsão de viagens entre as diversas zonas de tráfego (matriz O/D). O método se baseia no contexto de que a realização de uma viagem tem estreita relação com um processo desenvolvido em quatro fases: i) Geração de viagens, ii) Distribuição de viagens, iii) Divisão Modal e iv) Alocação de Tráfego (LOPES, 2005; TEDESCO, 2008). As quatro etapas são descritas, de maneira sintetizada, em seguida.

i) Geração de viagens

Origina o processo com a finalidade de prever o número de viagens de e para uma determinada região em estudo (viagens produzidas e atraídas por unidade de área considerada). Tal previsão geralmente é realizada por métodos de estimação de demanda. Os resultados dessa etapa servem de ponto de partida para as demais etapas do processo. Ortúzar e Willumsen (2011) apresentam os seguintes fatores que influenciam a produção de viagens: renda, propriedade de automóvel, estrutura do domicílio, tamanho da família, valor do solo, densidade residencial, acessibilidade. Como fatores que influenciam a atração de viagens numa dada zona pode-se citar: o número de empregos ofertados, o nível de atividade comercial e o número de matrículas escolares. Os seguintes modelos de geração de viagens podem ser citados: Fator de Crescimento, Taxas de Viagens, Análise de Categorias e Regressão.

ii) Distribuição de viagens

Este é o procedimento de distribuição de demanda de uma área para as demais, originando uma matriz de origens e destinos de viagens (pares O/D).

Usualmente utilizam-se modelos baseados em analogia, sendo o gravitacional o mais pertinente para esse fim (MEYER e STRASZHEIM, 1974). Outros modelos comumente aceitos são os Modelos de Fator de Crescimento tais como o Método do Crescimento Uniforme, o Método do Crescimento Médio, Método de Fratar e o Método de Furness.

iii) Divisão modal

Esta etapa objetiva a alocação de pares O/D entre os meios de transporte existentes (ou a serem criados), sem a preocupação com a definição das rotas. A divisão modal é feita com base nas variáveis que o usuário utiliza para fazer a sua escolha. Os fatores que influenciam a escolha modal incluem usualmente características socioeconômicas e características dos serviços dos modos de transporte. Os modelos de divisão modal classificam-se em determinísticos e probabilísticos. Os modelos determinísticos utilizam métodos quantitativos para determinar a proporção de viagens em cada modo, tais como Regressão Linear, Classificação Cruzada, Curvas de Desvio. Os modelos probabilísticos utilizam a probabilidade de escolha de cada modo para determinar a proporção de viagens. Os modelos probabilísticos mais utilizados são os econométricos (logístico e probit).

iv) Alocação de Tráfego

A última etapa do processo consiste na alocação dos fluxos de cada modo para cada par O/D da rede de transportes (OLIVEIRA, 2013). Os modelos de alocação são utilizados para fornecer estimativas de fluxos de tráfego em vias estratégicas. Um conceito que deve ser considerado em quase todos os modelos de atribuição de tráfego é o custo de deslocamento. Diante das diversas alternativas que o usuário tenha para efetuar uma viagem, naturalmente é ponderada uma série de fatores para decidir qual percurso seguir.

O método de quatro etapas foi um dos componentes do *Urban Transportation Planning System* (UTPS), desenvolvido nos EUA, com o objetivo de

promover uma padronização dos métodos de modelagem para transportes (HANSON e GIULIANO, 2004). Um fato importante a ser considerado é o de que, apesar de induzir o pensamento de que as etapas devam ter sua sequência respeitada, esta não é uma obrigação. Inicialmente, o modelo foi concebido em três etapas, sendo que a etapa de divisão modal foi adicionada a partir dos anos 60, passando a considerar os modos privados e públicos. Desde então, este modelo passou a considerar as viagens como um fator associado às pessoas e não aos automóveis (TEDESCO, 2008).

Apesar de amplamente utilizado, o modelo de quatro etapa, devido à sua natureza sequencial, pode propagar eventuais erros sem que se saibam quais serão os efeitos finais (VASCONCELOS, 2000). Além disso, o modelo não considera a dependência espacial das variáveis. Neste contexto, torna-se conveniente a utilização de novas abordagens relativas aos modelos, como a inclusão de variáveis com dependência espacial (variáveis regionalizadas) por meio da Geoestatística, por exemplo.

2.3 ESCOLHA MODAL

Variáveis como o número de viagens produzidas e atraídas, fluxo de tráfego e escolha modal são essenciais para o planejamento e previsão de demanda por transportes. Dentre essas variáveis, a previsão da variável de escolha modal, que se caracteriza como altamente dependente de fatores humanos, se baseia em fatores comportamentais. Os fatores comportamentais podem ser divididos em:

- Descritivos: explicam como os indivíduos se comportam;
- Abstratos: são baseados em termos não específicos a uma situação particular;
- Operacionais: os parâmetros e variáveis necessitam ser estimados e mensurados (RODRIGUES, 2012).

Conforme Rodrigues (2012), modelos probabilísticos são capazes de representar o comportamento dos indivíduos de maneira satisfatória, considerando os três fatores comportamentais.

Existem muitos casos em que o fenômeno que se deseja modelar é discreto, ao invés de contínuo. Esses são casos em que as decisões estudadas são resultados de uma seleção dentre distintas alternativas discretas. Nesses casos, é conveniente o uso de modelos de escolha discreta. Os modelos de escolha discreta podem ser de três tipos:

- Categóricos: a variável dependente não tem significado por si só, representa apenas uma codificação, ex: “1”, se o indivíduo trabalha e “0”, caso contrário.
- Com resposta ordenada: a variável dependente reflete um conceito por trás de sua ordem, ex: “1” muito satisfeito, “2” satisfeito, “3” insatisfeito.
- Contáveis: a variável dependente tem significado por si só, ex: número de patentes registradas de um país.

Além dos modelos econométricos e de pressupostos de escolha discreta, relevantes a estudos de demanda por transportes, é importante destacar a Função Utilidade. Uma Função Utilidade é uma expressão matemática que determina o grau de satisfação que o usuário do transporte obtém com a escolha do modo. De forma geral, a Função Utilizada é definida pela soma de variáveis e seus pesos relativos, tal como indicada pela Equação 2.1.

$$U = a_0 + a_1x_1 + a_2x_2 + \dots a_nx_n + \varepsilon \quad (2.1)$$

Onde, U é a utilidade derivada da escolha medida pelos atributos, x_i , x_i são as variáveis de transporte, a_i são os pesos relativos às variáveis e ε é o erro aleatório que segue distribuição de Weibull. Exemplos dos atributos relativos a transportes são: custo, tempo de viagem, tempo de espera, e algumas variações destes. A escolha pela utilidade prevê que o tomador de decisões realize compensações entre os atributos de cada modo para escolher um no qual seu benefício seja o maior possível (NOVAES, 1982), tal como:

$$U_{\text{MODO1}} > U_{\text{MODO2}} > U_{\text{MODO3}} \rightarrow \text{Escolhe-se o modo 1.}$$

2.3.1 Regressão Logística

Para prever se uma alternativa será escolhida, o valor de um determinado modo pode ser contrastado com uma opção alternativa, de maneira a ser transformado em uma probabilidade entre “0” e “1”. Entretanto, escolhas discretas não podem ser calibradas usando ajustes de curva como mínimos quadrados, pois sendo a variável dependente uma probabilidade, a modelagem por uma curva em forma de S se faz necessária. Um dos modelos de escolha discreta mais utilizado na Engenharia de Transporte é o logístico (ORTÚZAR e WILLUMSEN, 2011).

Se a discretização se restringe apenas a duas opções, a regressão logística é dita binomial, caso haja mais opções, a regressão é generalizada e nomeada multinomial. Vale salientar que essas regressões também são denominadas por Modelos Logit Binomial e Logit Multinomial.

A diferença da Regressão Logística para a Regressão Linear é de que a variável dependente não é contínua, mas sim, qualitativa. A Regressão Logística Binomial é normalmente expressa por duas categorias (binárias), conforme representado pela Figura 2.1, que representa a relação entre variável dependente binária (y) e variável independente contínua (x).

Figura 2.1: Gráfico de dispersão para variável dependente com apenas duas categorias.

Portanto, o método dos mínimos quadrados que era a base da Regressão Linear, já não oferece estimadores plausíveis. A solução para isso é o uso da Regressão Logística, pois essa permite o uso de um modelo (curva em S) para prever a probabilidade π de um evento específico, ilustrado na Figura 2.2.

Figura 2.2: Curva S de probabilidade para a variável dependente binária.

2.3.1.1 Regressão Logística Simples

Na Regressão Logística Simples, a variável dependente possui apenas duas categorias (“0” e “1”) e apenas uma variável independente envolvida. O modelo de Regressão para a curva S normalmente implica em dificuldades relativas ao cálculo. Para isto é possível considerar uma transformação logística de π que é calculada mais facilmente. A função logística $g(x)$ é dada por:

$$g(x) = \ln\left(\frac{\pi}{1-\pi}\right) \quad (2.2)$$

$$g(x) = \beta_0 + \beta_1 x \quad (2.3)$$

Sendo β_0 e β_1 os coeficientes da equação de calibração.

Em posse da Regressão Logística Simples (Equação 2.3), é simples calcular o valor da probabilidade π .

$$\pi = \frac{1}{1 + e^{-g(x)}} \quad (2.4)$$

Análises qualitativas podem ser feitas considerando as equações 2.3 e 2.4:

$$\left\{ \begin{array}{l} \text{Se } \beta_1 < 0, \pi \text{ é crescente} \\ \text{Se } \beta_1 > 0, \pi \text{ é decrescente} \\ \text{Se } x \rightarrow \infty \text{ e } \beta_1 < 0, \pi \rightarrow 0 \\ \text{Se } x \rightarrow \infty \text{ e } \beta_1 > 0, \pi \rightarrow 1 \\ \text{Se } \beta_1 = 0, g(x) \text{ é independente de } x \end{array} \right.$$

Em níveis muito baixos, a probabilidade tende a “0”. Entretanto, quando a variável aumenta, a probabilidade aumenta para cima da curva, mas em seguida, a inclinação começa a diminuir, de forma que, em qualquer nível, a probabilidade tenderá a “1”, não extrapolando esse valor.

A qualidade do ajuste pode ser mensurada através de medidas estatísticas tradicionais como erro padrão e significância; e também por medidas estatísticas apropriadas apenas para a análise de Regressão Logística, tais como: valor Wald, valor da verossimilhança (-2LL), testes de regressão de Cox & Snell e Nagelkerke, chi-quadrado avaliado pelo teste de Hosner & Lemeshow (HAIR et al., 2010).

2.3.1.2 Regressão Logística Múltipla

A Regressão Logística Múltipla (RLM) consiste da generalização do modelo de Regressão Logística Simples. Novamente, a variável dependente é binária, porém, existe mais de uma variável independente, ou seja, caso múltiplo.

A generalização consiste em uma alteração na função logística, que agora considera as múltiplas variáveis independentes:

$$g(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \quad (2.5)$$

$$g(x) = \sum_{i=0}^n \beta_i x_i \quad (2.6)$$

Sendo $x_0 = 1$, $\beta_0, \beta_1, \dots, \beta_n$ os coeficientes da equação de calibração e n o número de variáveis independentes.

Por meio da Regressão Logística Múltipla (Equação 2.6), calcula-se a probabilidade π (Equação 2.4), com $g(x)$ adequado.

As mesmas medidas de avaliação da Regressão Logística Simples podem ser usadas para RLM.

2.3.1.3 Regressão Logística Multinomial

Pode-se modificar a Regressão Logística Múltipla de maneira que a variável dependente apresente mais de dois níveis de codificação. Ao invés de apenas ter-se “0” e “1”, agora pode-se obter “0”, “1”, “2”, por exemplo.

Faz-se necessário k funções logísticas para cada uma das k categorias que a variável dependente pode assumir:

$$g_1(x) = \beta_{1,0} + \beta_{1,1}x_1 + \beta_{1,2}x_2 + \cdots + \beta_{1,n}x_n$$

$$g_2(x) = \beta_{2,0} + \beta_{2,1}x_1 + \beta_{2,2}x_2 + \cdots + \beta_{2,n}x_n$$

$$g_k(x) = \beta_{k,0} + \beta_{k,1}x_1 + \beta_{k,2}x_2 + \cdots + \beta_{k,n}x_n$$

$$g_k(x) = \sum_{i=0}^n \beta_{k,i}x_i \quad (2.7)$$

Sendo $x_0 = 1$, $\beta_{k,0}$, $\beta_{k,1}, \dots, \beta_{k,n}$ os coeficientes da equação de calibração e n o número de variáveis independentes.

Isso irá gerar as probabilidades π_k para cada uma das k classes.

$$\pi_k = \frac{1}{1 + e^{-g_k(x)}} \quad (2.8)$$

As medidas de avaliação usadas na Regressão Logística Simples e Regressão Logística Múltipla também podem ser usadas para Regressão Logística Multinomial (FIGUEIRA, 2006).

3. ESTATÍSTICA ESPACIAL

O capítulo 3 versa sobre o referencial teórico acerca da estatística espacial, com ênfase na Geoestatística. A estatística espacial é de grande relevância, uma vez que apresenta um método alternativo de análise quando comparado aos modelos clássicos de demanda por transportes. O capítulo também descreve e conceitualiza o processo de modelagem geoestatística: (1) análise de variáveis regionalizadas, (2) modelagem de semivariogramas experimentais, (3) validação e (4) Krigagem.

3.1 CONCEITOS DE ESTATÍSTICA ESPACIAL

Compreender a distribuição espacial de dados oriundos de fenômenos ocorridos no espaço constitui um grande desafio para a elucidação de questões centrais em diversas áreas do conhecimento, seja em saúde (LAPA *et al.*, 2006), em economia (PEROBELLI, 2004), em meteorologia (SIQUEIRA *et al.*, 2007), em transportes (PITOMBO *et al.*, 2010a). Com o advento de recursos computacionais poderosos, tornam-se cada vez mais usuais estudos onde a distribuição espacial dos dados é avaliada. Este processo é denominado Análise Espacial (DRUCK *et al.*, 2004).

Análises estatísticas tradicionais, baseadas na independência das observações, têm sido substituídas por análises espaciais, as quais consideram as correlações entre observações vizinhas. Estas análises são baseadas na teoria das variáveis regionalizadas (MATHERON, 1971), a qual considera as características estruturais e aleatórias de uma variável espacialmente distribuída, de forma a descrevê-la adequadamente.

Um dos conceitos básicos de análise espacial é o estudo do comportamento das variáveis. Caso haja dispersão espacial não uniforme dos valores amostrais, pode-se dizer que a variável tem comportamento anisotrópico. Caso contrário, o comportamento espacial é isotrópico.

Uma vantagem da análise espacial é a de que esta é mais propícia para lidar com a anisotropia, pois representa bem a variabilidade espacial da propriedade em estudo. Procedimentos determinísticos para este fim são considerados limitados, pois não consideram a estrutura de autocorrelação espacial bem como a anisotropia presente (CAMARGO, 2001).

Diversos tipos de análises podem ser promovidos de acordo com o tipo do dado o qual se trabalha. A Tabela 3.1, adaptada de Druck *et al.* (2004), ilustra a questão.

Tabela 3.1: Tipos de dados e problemas em análise espacial

Problemas em análise espacial	Tipos de dados	Exemplo	Aplicações típicas
Análise de padrões pontuais	Eventos localizados	Ocorrência de doenças	Determinação de padrões e agrupamentos
Análise de superfícies	Amostras de campo e matrizes	Depósitos minerais	Interpolação e medidas de incerteza
Análise de áreas	Polígonos e atributos	Dados censitários	Regressão e distribuições conjuntas

Fonte: Druck *et al.*, 2004 (adaptado).

A análise espacial, em seu caráter exploratório, pode ocorrer de diversas maneiras. Em seguida, são citadas representações de análise espacial exploratória.

- Visualização de Dados: Trata-se da forma mais simples e intuitiva de realizar uma análise espacial. Consiste em considerar valores de corte de maneira a identificar valores extremos no espaço amostral. Isto é, considerando um valor de corte propício a análise, apenas os valores acima do valor de corte são exibidos em uma superfície espacial (DRUCK *et al.*, 2004).

- Gráficos de Médias e Medianas: Estes gráficos são construídos levando em conta as coordenadas geográficas dos centroides de áreas com valores comuns de maneira a montar uma matriz de médias e medianas entre cada par de áreas. Dessa forma, estes gráficos (matrizes) permitem explorar simultaneamente a presença de tendências espaciais. Quando se encontra uma tendência espacial ao longo de alguma direção, os valores dessa matriz variam suavemente, caso

contrário, para valores muito discrepantes, considera-se a não presença de tendência espacial (DRUCK *et al.*, 2004).

- Análise de Autocorrelação Espacial: Esta análise é muito parecida com a anterior, alterando apenas a abordagem considerada. Ela visa identificar estrutura de correlação espacial pela estimação da magnitude da autocorrelação espacial entre as áreas. Quando se dispõe de um número grande de áreas, é muito provável que existam diferentes regimes de correlação (DRUCK *et al.*, 2004).

- Semivariograma: O semivariograma é uma representação gráfica indicadora da dependência espacial através do cálculo de variância entre pontos da amostra. Nele é considerada a associação de um valor único de atributo a cada ponto geográfico. Através dos valores calculados, pode-se descrever dados como sendo não-estacionários quando o semivariograma não se estabiliza, e estacionários quando o semivariograma é estabilizado. Os semivariogramas devem ser realizados para todas as direções de interesse (DRUCK *et al.*, 2004).

- Análises baseadas em estimadores: Algumas vezes é conveniente realizar a reestimação dos valores do espaço amostral, de maneira a suprimir a ausência de dados em determinadas áreas ou até mesmo realizar uma suavização de flutuações aleatórias muito intensas. Para isto, algumas técnicas são utilizadas. Um exemplo é a estimativa bayesiana local, que consiste em estimar valores convergindo os dados para médias locais. Após a estimação, os dados são visualizados como um todo em um mapa. Esta técnica pode ser combinada com as quatro citadas anteriormente (DRUCK *et al.*, 2004).

Enquanto as técnicas de análise exploratória descrevem e visualizam dados espaciais para a verificação da existência de padrões de associação espacial, as técnicas de análise confirmatória envolvem o conjunto de modelos de estimação e procedimentos para sua validação, geralmente realizados por regressões (KREMPL, 2004). No entanto, a utilização dessa ferramenta depende de uma averiguação preliminar de existência de homogeneidade espacial. Caso seja identificada

homogeneidade espacial, podem-se usar modelos globais, caso contrário, são utilizados modelos locais.

Para o caso de modelos com efeitos espaciais globais, supõe-se que é possível capturar a estrutura de correlações espaciais por meio de um único parâmetro (variável), que deve ser adicionado ao modelo de regressão tradicional.

Já para o caso de modelos com efeitos espaciais locais, os coeficientes de regressão precisam refletir a heterogeneidade espacial. Para isso é conveniente a divisão do espaço de estudo em subregiões, cada uma com seu padrão espacial próprio (DRUCK *et al.*, 2004). Cada regime das subregiões colabora com os seus próprios valores para os coeficientes que irão compor, por exemplo, uma regressão geograficamente ponderada, regressão esta que considera que pontos mais próximos ao ponto da região de estudo têm maior influência nos parâmetros estimados desta regressão (CARVALHO *et al.*, 2006).

Dentre estas análises, baseadas na teoria das variáveis regionalizadas (MATHERON, 1971), destaca-se a Geoestatística, também conhecida como análise espacial de superfícies, cuja base conceitual é análoga. A Geoestatística desempenha papel semelhante às análises de estatística espacial descritas anteriormente, porém mais rebuscada, pois combina a utilização de semivariogramas experimentais com um poderoso estimador, a Krigagem.

3.2 GEOESTATÍSTICA

O processo de modelagem geoestatística é melhor definido quando segregado pelas etapas de: (1) Análise das variáveis regionalizadas; (2) Modelagem de semivariogramas experimentais; (3) Validação e (4) Krigagem. Tais etapas são descritas nas seções subsequentes.

3.2.1 Análise das variáveis regionalizadas

Quando um fenômeno distribui-se espacialmente e exibe certa estrutura espacial, deve-se dizer que este é um fenômeno regionalizado. Haverá então pontos do espaço com certas características referentes a este fenômeno. Diz-se então que essas variáveis, que descrevem as características do fenômeno, são denominadas variáveis regionalizadas (MATHERON, 1971).

A teoria das variáveis regionalizadas, desenvolvida por Matheron, diz que, sendo a variável regionalizada distribuída no espaço (ou tempo), esta pode ser vista como resultado de uma função aleatória (processo aleatório, campo aleatório ou processo estocástico). Esta teoria consolida o alicerce da Geoestatística (CAMARGO *et al.*, 2001).

A Geoestatística estuda o comportamento das variáveis regionalizadas, que, diferentemente das aleatórias, possuem uma distribuição e variação espacial com continuidade aparente. A variação espacial de uma variável regionalizada é composta de três componentes (PEREIRA, 2011):

- Componente estrutural não aleatória: associada a um valor médio constante ou uma tendência constante;
- Componente aleatória: componente espacialmente correlacionada;
- Componente de ruído aleatório: trata-se do erro residual da variável regionalizada.

A estrutura espacial de uma variável regionalizada pode apresentar comportamento diferente dependendo da direção de estudo (anisotropia). Dessa forma, a análise e modelagem da estrutura espacial de uma variável regionalizada tem como ferramenta primária o semivariograma (GONÇALVES *et al.*, 2001; SOUZA *et al.*, 2004).

A natureza das variáveis regionalizadas pode ser de dois tipos: quantitativa e qualitativa. As variáveis quantitativas podem ser contínuas ou discretas. As variáveis qualitativas (categóricas) podem ser subclassificadas em nominal e ordinal.

3.2.2 Modelagem de semivariogramas experimentais

O semivariograma condensa as informações obtidas pelas variáveis regionalizadas por meio de uma representação gráfica em que a ordenada é a variância entre pares de observações (γ) e a abscissa é a distância entre estes pares (h).

A construção do semivariograma é uma etapa da modelagem geoestatística que trata da análise do comportamento estrutural das variáveis regionalizadas. O fenômeno espacial é avaliado pela mensuração das variâncias entre pares de pontos. A variância entre pontos (γ) é definida como a diferença quadrática média entre observações de uma variável aleatória $Z(x)$, separadas por uma distância h (Equação 3.1).

$$\gamma(h) = [Z(x) - Z(x + h)]^2 \quad (3.1)$$

Dessa forma, valores baixos indicam menor variabilidade, ou seja, maior similaridade entre as observações (MELLO, 2004). A função do semivariograma é determinada pela metade da média das variâncias entre os pontos:

$$\gamma(h) = \frac{1}{2N} \sum_{i=1}^{n(h)} [Z(x) - Z(x + h)]^2 \quad (3.2)$$

Em que, N é número total de observações da amostra em cada lag (h).

Alguns autores utilizam a terminologia variograma para a modelagem geoestatística. O variograma é definido como metade do semivariograma.

3.2.2.1 Parâmetros para definição do semivariograma

A elaboração de semivariogramas demanda a definição dos seguintes parâmetros:

- Distância de corte (d_{corte}): limite superior para cálculo do semivariograma. O valor usual é metade da maior distância entre pontos da amostra (LANDIM, 2004).
- Lag: segmentos delimitados a fim de auxiliar o cálculo das variâncias entre pares de pontos. São definidos os parâmetros para o lag ($\vec{h}$):
 - Tamanho do lag (h): distância média entre amostras vizinhas;
 - Número de lags (n_{lags}): valor escolhido de maneira que:

$$h \times n_{\text{lags}} < d_{\text{corte}} \text{ para que haja representação completa das amostras;}$$
 - Tolerância de lag (Δh): valores dentro dos limites ($h + \Delta h$) e ($h - \Delta h$) são considerados como pertencentes a h .
- Direção angular (θ): direção adotada para elaboração do semivariograma que melhor representa a distribuição espacial dos dados. Deste parâmetro, definem-se:
 - Tolerância angular ($\Delta\theta$): valores dentro dos limites ($\theta + \Delta\theta$) e ($\theta - \Delta\theta$) são considerados como pertencentes a θ .
 - Direção ortogonal ($\theta + 90^\circ$): direção ortogonal à direção adotada para elaboração do semivariograma.

3.2.2.2 Processo de obtenção do semivariograma

Considera-se o lag de tamanho h com tolerância Δh e a direção angular θ com tolerância angular $\Delta\theta$; n_{lags} o número de lags, $N=[n_1, n_2, n_3, \dots, n_N]$ o número total de amostras; e d_{corte} a distância de corte.

1) O processo tem início em um primeiro ponto n_1 da amostra. Para toda observação distante $h \pm \Delta h$ e com ângulo entre $\theta \pm \Delta\theta$ deste ponto, calcula-se a diferença entre os valores dos pares (Figura 3.1);

Figura 3.1: Etapa 1) para processo de obtenção do semivariograma experimental.

2) Refaz-se o passo 1) para todas as N amostras (Figura 3.2);

Figura 3.2: Etapa 2) para processo de obtenção do semivariograma experimental.

3) Obtém-se a variância γ para o lag de tamanho h . Este valor estará representado por um dos pontos γ no semivariograma experimental, onde o eixo y é representado pela variância γ e o eixo x pela distância h (Figura 3.3);

Figura 3.3: Etapa 3) para processo de obtenção do semivariograma experimental.

4) Repete-se os passos de 1) até 3) para os lags $2h, 3h, \dots, n_{lags} * h$, afim de se obter $\gamma(2h), \gamma(3h), \dots, \gamma(n_{lags} * h)$, desde que os valores de distância não excedam a distância de corte (d_{corte}), conforme Figura 3.4.

Figura 3.4: Etapa 4) para processo de obtenção do semivariograma experimental.

3.2.2.3 Interpretação gráfica do semivariograma

Os parâmetros gráficos do semivariograma podem ser observados na Figura 3.5. O alcance ou amplitude (a) é a distância em que as amostras se apresentam correlacionadas espacialmente. Desta distância em diante, considera-se que não existe mais dependência espacial entre as amostras, porque a variância γ entre pares de amostras não se altera com a distância h (CAMARGO, 1998). O patamar (C) é o valor do eixo de variância γ correspondente ao alcance (a).

À medida que h tende a zero, a variância γ se aproxima de um valor que revela a descontinuidade do semivariograma para distâncias menores do que a distância média entre as amostras. Este valor é definido como efeito pepita (C_0). O efeito pepita representa a incerteza à pequena escala, como também a alta variabilidade da variável em distâncias pequenas (SOARES, 2006). Podem ocorrer também casos onde o efeito pepita seja nulo, ou seja, são identificadas variâncias nulas entre pontos a curtas distâncias ($h \rightarrow 0$). Quando existente, esta descontinuidade junto à origem pode ocorrer devido a erros de medição ou à

inconstância de variáveis (SOUSA e MUGE, 1990). O efeito pepita pode ser devido à: i) possível existência de erros de amostragem e ii) microrregionalizações não perceptíveis à escala de análise.

O patamar (*sill*), observado na Figura 3.5 mede a variância do espaço amostral. O *sill* pode ser definido como a variância teórica dos dados ou, conforme Leung (2002) pode ser substituído pela média entre as variâncias experimentais (média dos pontos de variância γ).

A contribuição (C_1) é a diferença entre o patamar (C) e o efeito pepita (C_0).

Figura 3.5: Parâmetros para interpretação gráfica do semivariograma.

O comportamento espacial isotrópico da variável (Figura 3.6 (a)) justifica a escolha de um semivariograma omnidirecional (CORREIA, 2010). Caso contrário, para anisotropia (efeito de direção), deve-se examinar semivariogramas para diversas direções de variabilidade espacial. Na anisotropia (Figura 3.6 (b)), a função do semivariograma varia de acordo com a direção (YAMAMOTO e LANDIM, 2013). De modo contínuo, a variabilidade dos dados acontece em uma dada direção particular (SOARES, 2006). A direção principal refere-se ao ângulo θ de máxima variabilidade espacial dos dados, no exemplo de anisotropia da Figura 3.6 (b), a direção principal é visualizada como $\theta = -45^\circ$.

Figura 3.6: Representação de uma variável espacial isotrópica (a) e anisotrópica (b).

A anisotropia é representada por dois tipos principais: geométrica e zonal. Segundo Soares (2006), a anisotropia geométrica é um modelo no qual a continuidade espacial, denunciada pelas amplitudes dos semivariogramas, varia da direção de maior continuidade (amplitude) à direção de menor amplitude, perpendicular àquela, de um modo gradual segundo a equação da elipse no plano daquelas duas direções. Dessa forma, os semivariogramas de dados com estrutura anisotrópica geométrica se caracterizam pela existência de um único patamar e amplitudes diferentes. No entanto, dados espaciais com anisotropia zonal apresentam patamares diferentes, conforme direção analisada, mas todos sob um mesmo alcance (YAMAMOTO e LANDIM, 2013).

3.2.2.4 Ajuste do semivariograma experimental

O ajuste de semivariogramas experimentais é realizado de acordo com modelos teóricos, conforme Figura 3.7.

Figura 3.7: Exemplos de melhores ajustes para semivariogramas experimentais.
Fonte: OLEA (2006).

Os modelos teóricos podem ser para um semivariograma cúbico, esférico, exponencial, gaussiano, pentaesférico, lei de potência e efeito senoidal. Embora as sete curvas sejam convenientes a modelos de ajuste, as três curvas mais utilizadas na literatura são: exponencial, esférica e gaussiana, cujas expressões estão representadas nas Equações 3.3, 3.4 e 3.5, respectivamente (OLEA, 2006).

- Exponencial:

$$Ex(h) = C \left(1 - e^{-3h/a} \right) \quad (3.3)$$

- Gaussiano:

$$G(h) = C \left(1 - e^{-3(h/a)^2} \right) \quad (3.4)$$

- Esférico:

$$Sp(h) = \begin{cases} C \left(\frac{3h}{2a} - \frac{1}{2} \left(\frac{h}{a} \right)^3 \right), & 0 \leq |h| < |a| \\ C, & |a| \leq |h| \end{cases} \quad (3.5)$$

A correta elaboração de semivariogramas experimentais e a qualidade do ajuste são fundamentais para a etapa de Krigagem.

3.2.3 Validação

O processo de validação é uma das etapas de obtenção de um modelo que represente de forma adequada a realidade. Kramer (1998) desenvolve em seu estudo um esquema básico que deve ser seguido durante o processo de escolha de um modelo representativo mais adequado. A Figura 3.8 ilustra o esquema.

Figura 3.8: Fluxograma para abordagem básica de calibração e validação de modelos.
Fonte: Kramer, 1998 (traduzido e adaptado).

O uso de dados adequados à pesquisa pode parecer algo intrínseco, porém não é um passo trivial. Muitos fatores influenciam a adequabilidade dos dados, tais como número de amostras destinadas ao conjunto de calibração e validação, precisão dos dados utilizados na calibração. Entretanto, apenas após realizar os passos seguintes é que muitas vezes problemas são evidenciados e é por isso que, de acordo com o esquema, sempre é possível o retorno ao passo inicial após aprimoramento de acordo com as necessidades.

Dos passos do esquema, a construção dos modelos é o mais simples. Através de adventos computacionais modernos, vários pacotes computacionais permitem a manipulação de uma grande quantidade de dados a fim de estabelecer procedimentos que possibilitem a extração de informações relativas aos objetos de estudo. Esta etapa é nomeada calibração e é realizada utilizando uma parte dos dados originais destinada apenas para esse propósito.

O passo seguinte consiste em testar os modelos implementados no passo anterior. A melhor forma de averiguar se uma calibração foi feita de maneira errônea é por meio de um conjunto de testes. Esse conjunto de testes, denominado validação, é realizado com a parcela de dados complementar à utilizada na calibração. Para tal, considera-se que essas amostras de validação sejam desconhecidas e, utilizando a calibração prévia, tenta-se estimar seus valores. A seguir, os valores previstos são comparados com os valores conhecidos para essas amostras. Dessa forma, com o objetivo de avaliar a qualidade do método estabelecido, medidas de erro entre as previsões e os valores observados são mensuradas.

Posterior à calibração e validação dos modelos criados, pode-se escolher o modelo que melhor atendeu às expectativas. Entretanto, este modelo deve ser usado com cautela, pois contempla apenas fatores envolvidos nas condições propostas. Isso quer dizer que, para o conjunto de dados em questão, o melhor modelo escolhido pode ser uma boa representação da realidade nas condições de tempo e espaço estabelecidas. Dessa forma, o modelo deve ser usado com certo

nível de ceticismo uma vez que o evento modelado possa ser submetido a diferentes fatores externos dos estabelecidos inicialmente.

De maneira a suprimir eventuais inadequações provenientes de novas condições, a etapa seguinte estabelece que o modelo possa sofrer alterações e aprimoramentos a fim de se adaptar a possíveis necessidades futuras.

Para a Geoestatística, as etapas do esquema apresentado são igualmente válidas. Um dos fatores que influencia a adequabilidade dos dados, além do número de elementos da amostra, é a correta escolha da variável de estudo. A escolha dos dados deve contemplar o uso de variáveis regionalizadas, pois estas são fundamentais no processo de construção de semivariogramas.

O processo de ajustes dos semivariogramas identifica a direção principal de tendência dos dados utilizados. Através do semivariograma relativo à direção principal, o modelo de calibração é estabelecido por meio do processo de Krigagem.

A verificação da qualidade da modelagem geoestatística é realizada por validação. Nesta etapa, é quantificado o erro e a variância dos resíduos a partir dos valores observados e estimados nas coordenadas de valores conhecidos das variáveis. Para o tamanho do conjunto de validação, em geral utiliza-se um número igual a 30% do total de elementos da amostra (LUZ, 2003).

Um procedimento de validação utilizado com frequência na Geoestatística é o teste do ponto fictício (validação cruzada). Este teste consiste em subdividir o conjunto de dados destinados à validação em grupos de mesmo tamanho. Cada grupo é usado individualmente como teste, enquanto os outros são usados para calibração. Quando um grupo é usado como teste, apesar de se conhecer os valores *a priori*, assume-se esses valores como sendo desconhecidos. Então, tenta-se estimar os valores deste grupo através da calibração feita pelos grupos complementares. O processo se repete até que todos os grupos tenham passado pelo processo de validação. Esta técnica é interessante, pois, diferente da validação convencional, evita a superposição dos conjuntos de teste. A validação cruzada adotada por este trabalho define grupos unitários.

3.2.4 Krigagem

A Krigagem é um dos métodos de estimativa utilizada na Geoestatística. A teoria foi desenvolvida a partir do trabalho do seu inventor Daniel G. Krige. Trata-se de um processo de predição linear, pois suas estimativas são combinações lineares ponderadas por dados existentes. A Krigagem fornece, em geral, estimativas não tendenciosas e com variância mínima (LANDIM, 2003).

A Krigagem difere-se de outros interpoladores, pois parte do pressuposto que pontos próximos no espaço tendem a ter valores mais parecidos do que pontos afastados. Esse pressuposto é garantido por uma ponderação intrínseca ao método.

A partir das curvas ajustadas de semivariogramas do espaço amostral, valores de efeito pepita (C_0), contribuição (C_1) e amplitude (a) são utilizados para a definição de pesos W (ponderadores). A função dos ponderadores é a de ditar a influência dos dados amostrais na estimação de novos valores. O estimador de Krigagem é dado pela Equação 3.6.

$$Z(x_0) = \sum_{i=1}^n \lambda_i \times Z(x_i) \quad (3.6)$$

O sistema de Krigagem é constituído por ponderadores com o objetivo de conduzir estimativas a erros nulos (não enviesamento universal) e minimização da variância (otimização):

$$\lambda = (G)^{-1} \times M \quad (3.7)$$

Os ponderadores do estimador de Krigagem resumem dois fatores: 1) o fator de distância estrutural entre amostras e o ponto a estimar; 2) o fator desagregação originado pela matriz de covariâncias entre as amostras, ou seja, quanto mais correlacionadas estiverem, maior a redundância e menor será o peso individual na construção do estimador (SOARES, 2006). A solução para os ponderadores da Krigagem Ordinária é dada pelas representações matriciais:

$$\lambda = \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ \dots \\ \lambda_n \\ \mu \end{pmatrix}, \quad G = \begin{pmatrix} \gamma(x_1, x_1) & \gamma(x_1, x_2) & \dots & \gamma(x_1, x_n) & 1 \\ \gamma(x_2, x_1) & \gamma(x_2, x_2) & \dots & \gamma(x_2, x_n) & 1 \\ \dots & \dots & \dots & \dots & 1 \\ \dots & \dots & \dots & \dots & 1 \\ \gamma(x_n, x_1) & \gamma(x_n, x_2) & \dots & \gamma(x_n, x_n) & 1 \\ 1 & 1 & 1 & 1 & 0 \end{pmatrix}, \quad M = \begin{pmatrix} \gamma(x_1, x_0) \\ \gamma(x_2, x_0) \\ \dots \\ \dots \\ \gamma(x_n, x_0) \\ 1 \end{pmatrix}.$$

Os ponderadores de Krigagem são considerados dentro de uma área de atuação estabelecida por um elipsóide de raios determinados pelas amplitudes da direção principal e ortogonal.

As formas mais usuais de Krigagem são a Krigagem Simples (KS) e a Ordinária (KO). A KS é utilizada quando a média é assumida como estatisticamente constante para toda área de abrangência das amostras. Já a KO, por sua vez, considera a média flutuante ou móvel por toda área.

As variâncias de Krigagem não são normalmente medidas de acurácia de estimativa local. Para satisfazer esta necessidade, uma das soluções é a Krigagem Indicativa (KI). Na KI, o enfoque passa a ser o de definir áreas com maior ou menor probabilidade de ocorrência de um determinado evento. Diferentemente da KO, onde o enfoque é a estimação de um determinado valor. Neste trabalho, o interpolador utilizado é a KI, descrita na subseção seguinte.

3.2.4.1 Krigagem Indicativa (KI)

A KI consiste basicamente na aplicação da KO para variáveis transformadas, ou seja, a variável resultante da aplicação da função não linear $f(x)=0$ ou $f(x)=1$. O conceito inicial foi apresentado como proposta para construção de uma função de distribuição de probabilidades acumuladas para estimativa de distribuições espaciais (LANDIM e STURARO, 2002).

O primeiro passo da KI consiste em transformar os dados iniciais em indicadores, isto é, transformar valores que estão acima de um determinado valor de corte em “0” e os que estão abaixo em “1”.

$$i_j(v_c) = \begin{cases} 1, & \text{se } v_j \leq v_c \\ 0, & \text{se } v_j > v_c \end{cases}$$

Onde v_c é o nível de corte e v_j é o valor observado. Os passos seguintes são análogos à KO, porém aplicados aos dados transformados.

A frequência acumulada dos valores observados abaixo do nível de corte é representada por:

$$F(v_c) = \frac{1}{n} \sum_{j=1}^n i_j v_c \quad (3.6)$$

De maneira análoga, a proporção de valores abaixo do nível de corte pode também ser considerada como a média ponderada dos indicadores situados na vizinhança do local avaliado segundo:

$$\hat{F}(v_c) = \sum_{j=1}^n w_j i_j v_c \quad (3.7)$$

Sendo w_j os pesos, i_j os indicadores e v_c o nível de corte.

Em posse das frequências acumuladas, da proporção de valores e de semivariogramas experimentais, são calculadas as probabilidades para determinados locais da área amostral (LIMA *et al.*, 2008).

No presente trabalho, a variável objeto de estudo escolhida não é uma variável contínua transformada de probabilidade, mas sim, uma variável discreta relativa à escolha de modo motorizado, conforme descrito no capítulo posterior.

3.2.5 Geoestatística e Transportes

A proposta de utilização da Geoestatística na Engenharia de Transportes ainda é incipiente. Alguns trabalhos, como os de Braxmeier et al. (2005) e Du e Aultman-Hall (2006) deram início aos estudos na área de Engenharia de Tráfego. O conhecimento da análise espacial de Geoestatística, no campo de modelagem de demanda por transportes, ainda requer muitos conhecimentos básicos, principalmente acerca das variáveis, que encontram obstáculos por envolverem questões de comportamento humano. Alguns trabalhos provenientes do estudo da Geoestatística na área de demanda por transportes são mencionados a seguir.

Teixeira (2003) relacionou aspectos socioeconômicos à geração de viagens através da Geoestatística. Neste trabalho, foi aplicada a Krigagem Ordinária para a apresentação de uma metodologia para definir o zoneamento por áreas homogêneas relacionadas ao planejamento de transportes por meio de aspectos econômicos e de viagens da população através dos setores censitários.

Pitombo e Sousa (2009) aplicaram conceitos geoestatísticos para analisar geração de viagens urbanas por modo de transporte e motivo de viagem. No trabalho, em que foram utilizadas as técnicas de Krigagem Ordinária e um banco de dados agregado por Zona de Tráfego, estimou-se as variáveis para 40.000 pontos. No trabalho de Pitombo *et al.* (2010b) e Pitombo *et al.* (2014) também foram previstos dados de geração de viagens de dados agregados pela Geoestatística, contudo, foi utilizada, além da Krigagem Ordinária, a Krigagem com Deriva Externa.

Miura (2010) descreveu, em seu estudo, uma abordagem para estimação de tempo de viagem, através de Krigagem Ordinária. Os resultados demonstraram que as propriedades geoestatísticas de dados de tempo de viagem são tão favoráveis quanto os dados de malha viária.

Costa *et al.* (2013) estimaram a probabilidade da escolha modal através da Geoestatística por meio de dados desagregados, utilizando técnicas de Árvore de Decisão e Krigagem Ordinária.

Pitombo *et al.* (2015) propuseram um método de duas etapas para a estimação da escolha modal, através de um banco de dados domiciliar na cidade de São Carlos. Foram selecionados os atributos que mais influenciam na escolha modal por meio de Árvore de Decisão e após a comparação da eficácia da Árvore de Decisão com um Modelo Multinomial Logístico, realizou-se uma Krigagem Ordinária para prever a escolha modal, a partir da localização espacial.

Rocha (2014) analisou a geração de viagens urbanas por transporte coletivo através da Geoestatística, por meio de Krigagens Ordinária e com Deriva Externa em um banco de dados agregado referente à Região Metropolitana de Salvador.

Rocha *et al.* (2014) propuseram uma tentativa inicial de melhorar a modelagem Geoestatística a partir de dados muito agregados, sabendo que, muitas vezes, apenas estes estão disponibilizados e que, no entanto, a Geoestatística manipula melhor dados contínuos no espaço.

4. MATERIAIS E MÉTODO

Uma vez que nos capítulos 2 e 3 foi discutido o referencial teórico necessário para o entendimento das atividades deste trabalho, o capítulo 4 descreve os materiais e o método. Foram utilizados como banco de dados a Pesquisa Origem/Destino da área de estudo explanada a seguir e como pacotes computacionais: o IBM SPSS 22, o ESRI ArcGIS Platform 10.1, e o GeoMS 1.0. As etapas metodológicas também são descritas, posteriormente aos materiais.

4.1 MATERIAIS

4.1.1 Banco de dados e área de estudo

Os dados utilizados neste trabalho são referentes à Pesquisa Origem/Destino (O/D) realizada na Região Metropolitana de São Paulo (RMSP) em 2007 pela Companhia do Metropolitano de São Paulo. Na RMSP estão contidas 460 Zonas de Tráfego em 39 municípios (Figura 4.1). O Anexo 01 apresenta o mapa com a denominação de todos os municípios da RMSP.

Figura 4.1: Região Metropolitana de São Paulo e Zonas de Tráfego.
Fonte: Metrô (2007)

A Pesquisa O/D fornece dados para o conhecimento da natureza dos deslocamentos da população, identificando as principais viagens diárias, conforme motivo e modo de transporte. Além disso, a Pesquisa O/D permite estabelecer relações quantitativas entre as viagens realizadas e outras variáveis, como características socioeconômicas da população e aspectos de ocupação urbana, de forma a estabelecer projeções futuras das necessidades de viagens das pessoas.

Na Pesquisa O/D da RMSP de 2007 foram levantadas informações de 30 mil domicílios (Figura 4.2), escolhidos aleatoriamente e distribuídos nas 460 Zonas de Tráfego na RMSP (Figura 4.1). A distância entre o par de domicílios mais afastados entre si é de aproximadamente 142 quilômetros, enquanto a média entre os pares de domicílios mais próximos é de 146 metros.

Figura 4.2: Representação dos domicílios apurados na Pesquisa O/D da RMSP (2007).

A RMSP possui uma população de 19,8 milhões de habitantes (IBGE, 2010) na área de 7.946 km². A Tabela 4.1 apresenta as informações relacionadas aos municípios da RMSP e aos domicílios da Pesquisa O/D. Observa-se que a maior quantidade de domicílios entrevistados (Pesquisa O/D) está representada pelos municípios de São Paulo, Osasco, Santo André, Mogi das Cruzes e Guarulhos. Estes municípios são os municípios mais populosos, no entanto, estes não condizem com as respectivas maiores áreas ou densidades.

Tabela 4.1: Informações dos municípios da RMSP

Cód	Município	Domicílios*	Habitantes**	Área** km²	Densidade hab/km²
36	São Paulo	77,25%	11.253.503	1521,11	7.398
24	Osasco	2,52%	666.740	64,954	10.265
32	Santo André	2,08%	676.407	175,782	3.848
23	Mogi das Cruzes	2,04%	387.779	712,541	544
15	Guarulhos	2,03%	1.221.979	318,675	3.835
33	São Bernardo do Campo	1,54%	765.463	409,508	1.869
34	São Caetano do Sul	1,01%	149.263	15,331	9.736
2	Barueri	0,97%	240.749	65,701	3.664
22	Mauá	0,86%	417.064	61,909	6.737
6	Carapicuíba	0,79%	369.584	34,546	10.698
27	Ribeirão Pires	0,71%	113.068	99,075	1.141
7	Cotia	0,66%	201.150	323,994	621
37	Suzano	0,65%	262.480	206,236	1.273
8	Diadema	0,61%	386.089	30,756	12.553
9	Embu das Artes	0,55%	240.230	70,398	3.412
18	Itaquaquecetuba	0,48%	321.770	82,622	3.894
19	Jandira	0,42%	108.344	17,449	6.209
11	Ferraz de Vasconcelos	0,39%	168.306	29,564	5.693
26	Poá	0,39%	106.013	17,264	6.141
28	Rio Grande da Serra	0,38%	43.974	36,341	1.210
38	Taboão da Serra	0,38%	244.528	20,388	11.994
10	Embu-Guaçu	0,33%	62.769	155,641	403
5	Cajamar	0,30%	64.114	131,386	488
17	Itapevi	0,29%	200.769	82,658	2.429
31	Santana de Parnaíba	0,26%	108.813	179,949	605
12	Francisco Morato	0,23%	154.472	49,001	3.152
13	Franco da Rocha	0,23%	131.604	132,775	991
1	Arujá	0,21%	74.905	96,167	779
39	Vargem Grande Paulista	0,21%	42.997	42,489	1.012
16	Itapeçerica da Serra	0,17%	152.614	150,742	1.012
21	Mairiporã	0,16%	80.956	320,697	252
35	São Lourenço da Serra	0,16%	13.973	186,456	75
30	Santa Isabel	0,15%	50.453	363,332	139
3	Biritiba-Mirim	0,14%	28.575	317,406	90
25	Pirapora do Bom Jesus	0,14%	15.733	108,489	145
20	Juquitiba	0,13%	28.737	522,169	55
29	Salesópolis	0,09%	15.635	424,997	37
14	Guararema	0,07%	25.844	270,816	95
4	Caieiras	0,02%	86.529	97,642	886

*Informação obtida pela proporção de domicílios entrevistados pela Pesquisa O/D (2007).

**Informação obtida pelo censo 2010 (IBGE).

Obs: Em verde estão representados os cinco menores valores para cada atributo e em vermelho os cinco maiores valores.

4.1.2 Software

Os resultados obtidos neste trabalho de mestrado valeram-se de um conjunto de ferramentas computacionais, que auxiliaram em diversos campos, tais como: análises estatísticas tradicionais, análise espacial e modelagem geoestatística aplicadas ao campo da Engenharia de Transportes.

No âmbito das estatísticas tradicionais, cálculos como o de regressão e testes de qualidade (chi-quadrado, taxa de acertos) foram realizados através do *software* IBM SPSS Statistics 22.

Com relação à análise espacial, os sistemas de informação geográfica (SIGs) provêm ferramental necessário para o estabelecimento da visualização de dados por meio de mapas. A geração de superfície espacial foi realizada pelo *software* ESRI ArcGIS Platform, versão 10.1.

No campo da Geoestatística, as etapas de visualização estatística dos dados, geração de semivariogramas experimentais, ajuste por curvas teóricas, interpolação por Krigagem e validação cruzada (teste do ponto fictício) foram obtidas através do *software* GeoMS 1.0, desenvolvido pelo Centro de Modelização de Reservatórios Petrolíferos (CMRP-Portugal).

4.2 MÉTODO

O processo metodológico adotado se baseia nas etapas ilustradas na Figura 4.3 e descritas na sequência.

Figura 4.3: Esquema simplificado do método proposto.

4.1.1 Tratamento e visualização de dados

Os dados desagregados originalmente apresentavam 196.698 registros, em que cada registro representava uma viagem da Pesquisa Origem/Destino. Em se tratando de um estudo espacial, torna-se necessário que as informações estejam vinculadas a coordenadas geográficas. Como o estudo das viagens isoladas não representa um espaço, o método visou à transformação do banco de dados relacionados a viagens para informações atreladas aos domicílios.

Por se tratar de um banco de dados referente à Pesquisa O/D, observou-se que havia casos em que coordenadas geográficas estavam associadas a mais de um domicílio. Como uma das suposições da Geoestatística sugere que cada elemento do espaço tenha representação única, houve necessidade da adoção de um valor singular para cada elemento de espaço. Dessa forma, para cada caso de coordenada geográfica repetida, sorteou-se (através de pseudoaleatoriedade baseada em semente temporal) um domicílio para sua representação espacial.

As informações de interesse para este projeto estão descritas na Tabela 4.2, juntamente com as suas naturezas. Originalmente, a Pesquisa O/D indica os modos realizados para cada viagem. Neste trabalho, como cada linha de viagem tem suas informações compiladas em informações/domicílios, a variável de *modo de viagem* de cada linha do banco de dados foi transformada em variáveis de *proporção (%) de viagens por modo*.

Os modos de transporte foram classificados como: Transporte Individual Motorizado, Transporte não Motorizado e Transporte Público. O Transporte Individual Motorizado inclui viagens em que o indivíduo dirige automóvel ou motocicleta. O Transporte não Motorizado refere-se ao uso de bicicleta e modo a pé. O Transporte Público inclui viagens por ônibus do município de São Paulo, ônibus de outros municípios, ônibus metropolitano, microônibus do município de São Paulo, microônibus de outros municípios, microônibus metropolitano, metrô e trem.

Tabela 4.2: Informações utilizadas da Pesquisa O/D RMSP 2007.

Informações por domicílio	Natureza	
Coordenada geográfica x (*)	UTM	
Coordenada geográfica y (*)	UTM	
Zona	Qualitativa	
Município	Qualitativa	
Número de pessoas (**)	Quantitativa	Discreta
Critério de renda (**)	Qualitativa	
Quantidade de automóveis (**)	Quantitativa	Discreta
Quantidade de motocicletas (**)	Quantitativa	Discreta
Quantidade de bicicletas (**)	Quantitativa	Discreta
Número total de viagens (*) (**)	Quantitativa	Discreta
% de viagens por transporte individual motorizado (*)	Quantitativa	Contínua
% de viagens por transporte não motorizado (*)	Quantitativa	Contínua
% de viagens por transporte público (*)	Quantitativa	Contínua
Distância total de viagens (*) (**)	Quantitativa	Contínua
Distância por transporte individual motorizado (*) (**)	Quantitativa	Contínua
Distância por transporte não motorizado (*) (**)	Quantitativa	Contínua
Distância por transporte público (*) (**)	Quantitativa	Contínua
Preferência por modo motorizado (*) (**)	Categórica Dicotômica	

(*) Variável avaliada para a determinação da variável objeto de estudo na análise de estrutura espacial.

(**) Variável utilizada para a modelagem tradicional.

A variável *preferência por modo motorizado* originalmente não consta nos dados da Pesquisa O/D. Esta variável foi inserida com o intuito de prover informações complementares ao estudo e suavizar possíveis efeitos de microrregionalizações inerentes às preferências modais/domicílio, principalmente no caso de viagens não motorizadas. Dessa forma, para esta variável, foi necessário um tratamento adicional. Este consistiu da transformação da variável de *proporção de viagens por transporte público e por transporte individual motorizado* em uma variável categórica, em que:

Proporção de viagens por Transporte Público > Transporte Individual Motorizado → 1

Proporção de viagens por Transporte Individual Motorizado > Transporte Público → 0

Esta variável caracteriza a preferência entre o modo de transporte público e modo individual motorizado nos domicílios da RMSP. Portanto, foram eliminados os

casos em que a *proporção de viagens por transporte público* foi igual à *proporção por transporte individual motorizado*. Devido a esta etapa de eliminação, esta variável tem um número de registros menores do que as demais variáveis.

A Tabela 4.3 apresenta a caracterização das variáveis quantitativas através das medidas descritivas de médias, desvios padrão, modas, mínimos e máximos.

Tabela 4.3: Medidas descritivas para as variáveis quantitativas.

Informações por domicílio	média	desvio padrão	moda	mínimo	máximo
Número de pessoas	3,3	1,6	3	1	24
Quantidade de automóveis	0,9	0,9	1	0	8
Quantidade de motocicletas	0,1	0,3	0	0	9
Quantidade de bicicletas	0,5	0,9	0	0	12
Número total de viagens	6,8	4,6	2	1	68
% de viagens por transporte individual motorizado	27%	32%	0%	0%	100%
% de viagens por transporte não motorizado	17%	24%	0%	0%	100%
% de viagens por transporte público	33%	32%	0%	0%	100%
Distância total de viagens (km)	38,4	34,5	7,2	0,0016	427,1
Distância por transporte individual motorizado (km)	12,1	19,9	0,0	0,0000	273,0
Distância por transporte não motorizado (km)	2,0	7,5	0,0	0,0000	124,5
Distância por transporte público (km)	19,8	27,5	0,0	0,0000	423,9

A Tabela 4.4 apresenta o percentual de observações para a caracterização da variável qualitativa *critério de renda*.

Tabela 4.4: Percentual de observações da variável *critério de renda*.

Critério de Renda (Classes)							
1 (A1)	2 (A2)	3 (B1)	4 (B2)	5 (C1)	6 (C2)	7 (D)	8 (E)
3,0%	11,7%	17,3%	24,3%	23,2%	14,4%	5,7%	0,4%

A Tabela 4.5 apresenta o percentual de observações para a variável categórica dicotômica *preferência por modo motorizado*.

Tabela 4.5: Percentual de observações da variável *preferência por modo motorizado*.

Preferência por modo motorizado	
0	1
Transporte individual motorizado	Transporte público
44,8%	55,3%

Devido ao tratamento diferenciado entre a variável de *preferência por modo motorizado e as demais*, este trabalho possui uma amostra considerando a variável dicotômica (*preferência por modo motorizado*) e uma amostra considerando as demais variáveis, cujo número de registros é superior ao da primeira.

4.1.2 Escolha da variável objeto de estudo

A escolha da variável objeto de estudo se deu pela análise univariada dos dados. Atentou-se para a escolha de uma variável que se adequasse melhor à estatística espacial.

Com o objetivo de analisar o comportamento da variável e as tendências espaciais, a elaboração de mapas temáticos pontuais para as nove variáveis não se mostra como alternativa viável, pois a grande quantidade de pontos (domicílios) faz com que as informações se sobreponham. Dessa forma, a análise univariada para cada caso específico de variável foi considerada.

As nove variáveis relativas à demanda de transporte analisadas foram: *proporção de viagens por modo de transporte* (3), *distância total de viagens* (1), *total de viagens por domicílio* (1), *distância de viagens por modo de transporte* (3) e *preferência por modo motorizado* (1).

4.1.3 Modelagem de demanda por transportes

Posterior à definição da variável de demanda mais adequada às suposições de análise espacial, os dados passaram pelas etapas de modelagem geoestatística e de modelagem tradicional, com o intuito de obter as estimações das variáveis para coordenadas de valores conhecidos. Assim, tornou-se possível a realização de análises comparativas. Os processos de modelagem da variável de demanda escolhida (objeto de estudo) podem ser observados na Figura 4.4.

Figura 4.4: Processo de calibração/validação pela abordagem geoestatística e tradicional.

A amostra foi separada em duas partes: a primeira parte (70%), reservada para calibração dos dados e a segunda (30%), para validação. Tais proporções de subamostras foram utilizadas para ambas as abordagens. A escolha da razão 70/30 foi empírica. No entanto, não há uma proporção padrão para estipular os tamanhos dos conjuntos de calibração e validação. Valores comuns para a razão são: 60/40, 70/30, 80/20. É importante salientar que a divisão deve ser feita de maneira randômica e a maior proporção deve ser para o conjunto de calibração.

4.1.3.1 Modelagem geoestatística

O processo de modelagem geoestatística segue os passos de: i) Análise da estrutura espacial, ii) Krigagem, iii) Validação.

i) Análise da estrutura espacial

A análise da estrutura espacial é realizada pela elaboração de semivariogramas experimentais e posterior ajuste por curvas teóricas (esférica, exponencial, gaussiana), o qual permite a identificação da direção principal.

Com o objetivo de analisar as tendências espaciais, a elaboração de mapas temáticos pontuais não se mostra como alternativa viável, pois a grande quantidade

de pontos (domicílios) faz com que as informações se sobreponham. Assim sendo, semivariogramas experimentais foram calculados. Os semivariogramas demonstram graficamente a variabilidade espacial, além disso, esta informação torna-se mensurável.

Calculou-se o semivariograma para o caso omnidirecional, considerando, dessa forma, uma tolerância de 180° . Entretanto, uma abordagem exclusivamente omnidirecional seria válida apenas para casos de isotropia. Os casos de anisotropia demandam a definição da direção preferencial dos dados. Considerando a possibilidade de anisotropia da variável, os semivariogramas foram obtidos também para ângulos de teste variando de 15° e adotando uma tolerância de 1° .

O cálculo dos semivariogramas teve início com a escolha do tamanho de lag para cada uma das direções realizadas no estudo. Utilizou-se para o caso omnidirecional um tamanho de lag obtido pela média da distância euclidiana entre vizinhos próximos considerando todas as direções angulares. Para os demais semivariogramas, utilizou-se a média da distância euclidiana para cada uma das direções angulares estudadas.

Posteriormente à elaboração, foram investigados os parâmetros dos semivariogramas experimentais. A investigação intencionou a identificação de características para semivariogramas com indícios de um bom ajuste:

- Menor efeito pepita;
- Adequabilidade a curvas teóricas (esférica, exponencial, gaussiana);
- Maiores valores para a razão de amplitude entre os pares de direção.

Quando satisfeitos, estes parâmetros indicam semivariogramas com maiores tendências espaciais. Por meio desta análise geoestatística, os resultados apontaram a direção principal de estudo.

A partir de todos os semivariogramas, avaliou-se a estrutura espacial para realização dos passos de ajuste e definição dos parâmetros gráficos. O patamar foi considerado pela média entre as variâncias dos dados. Caso houvesse isotropia,

denunciada por semivariogramas direcionais similares, seria utilizado o caso omnidirecional e escolher-se-ia a curva com melhor ajuste. Contudo, o trabalho verificou anisotropia geométrica. Dessa forma, a direção principal é dada pelo par de semivariogramas (ortogonais) com maior razão (r):

$$r = \frac{\text{amplitude maior}}{\text{amplitude menor}} \quad (4.1)$$

ii) Krigagem

A etapa de Krigagem se dá por um método sofisticado de estimativa de valores. Métodos simples são realizados por interpolações não estocásticas e não utilizam semivariogramas. No entanto, a Krigagem utiliza os semivariogramas para a estimação dos valores das variáveis. Isso faz com que a Krigagem seja um método que considere, além da distância entre os pontos, a existência ou não de efeito pepita, a amplitude e a presença de anisotropia. (YAMAMOTO e LANDIM; 2013).

O resultado das estimativas é um mapa krigado com a superfície espacial, que consiste de um novo conjunto de dados binários, obtido a partir de dados conhecidos (dados de calibração – 70% da amostra). Este mapa é representado por pontos espaçados a uma distância d e dotados de três informações: latitude x , longitude y e valor da variável.

iii) Validação

A validação ideal seria através da leitura direta dos valores estimados, ou seja, a partir dos domicílios do conjunto de validação, relacioná-los a valores estimados para cada coordenada geográfica. Dessa forma, a Krigagem seria realizada para uma distância unitária ($d=1$), o que torna o arquivo gerado extenso e de difícil manipulação. Este trabalho buscou a utilização de um valor maior para a distância (d) das células. Em vista disso, foram adotados dois critérios de validação: a) pela média dos nós adjacentes e b) por validação cruzada.

a) Validação pela média dos nós adjacentes

Utilizando dados de validação (30% da amostra), a estimativa é calculada através da média entre os nós adjacentes. A Figura 4.5 representa um exemplo deste processo para um ponto P. Vale ressaltar que o processo de Krigagem adotado considera células com duas dimensões e são estimados valores nas extremidades destas células. Valores internos às células são calculados aproximadamente pela média das extremidades.

Figura 4.5: Representação da interpolação dos valores de coordenadas conhecidas.

Ao final, estes valores foram arredondados para transformar esta variável em binária, considerando que não se trata de probabilidade, mas sim de uma variável categórica.

b) Validação cruzada

Nos passos anteriores, puderam-se obter os parâmetros de ajuste do semivariograma através do conjunto de calibração. Estes parâmetros foram utilizados para a realização de uma nova Krigagem (realizada apenas para a validação cruzada), que considerou a direção principal, a ortogonal e parâmetros de contribuição (C_1), efeito pepita (C_0), amplitude (a) e o conjunto de 30% da amostra. Vale ressaltar que, apesar de este método utilizar os parâmetros gráficos do semivariogramas da amostra de calibração, a validação cruzada se trata de uma nova calibração com menor número de informações (30% da amostra).

A etapa de validação cruzada considera um ponto do conjunto de validação como desconhecido e calcula seu valor através dos outros pontos de valores conhecidos por meio de uma função. No caso deste trabalho, a função é advinda da Krigagem, que considera os ponderadores e a proximidade espacial. O processo de considerar um ponto como desconhecido se repete para os demais, até que todos os pontos do conjunto de validação sejam estimados. Ao final desta etapa, os valores estimados foram também binarizados.

Uma vez que, para ambas as abordagens a) e b), foram obtidos valores em coordenadas de valores conhecidos, pode-se avaliar a eficácia das duas vertentes através do uso de medidas estatísticas de comparação entre valores previstos e observados. As medidas adotadas foram: teste chi-quadrado e taxa de acertos (Tabela 4.6), considerando que foram comparados dois vetores de valores qualitativos.

Tabela 4.6: Medidas estatísticas avaliadas nas modelagens.

Chi-quadrado (*)	
$\chi^2 = \sum_{i=1}^N \frac{O - E}{E} \quad (4.3)$	
Taxa de acertos	
$\frac{1}{N} \sum_{i=1}^N k_i, \quad \begin{cases} \text{se } x_i = y_i, k_i = 1 \\ \text{se } x_i \neq y_i, k_i = 0 \end{cases} \quad (4.4)$	

(*) O: frequência observada na célula;
E: frequência esperada na célula;
N: número total de componentes da amostra.

Realizou-se, também para ambas as propostas geoestatísticas, a validação por município, cujo conjunto de validação é uma parte dos 30%, referente à proporção de registros de domicílios de cada município. Nesta validação, a medida estatística adotada foi a porcentagem de acerto.

4.1.3.2 Modelagem tradicional

O processo de modelagem tradicional segue os passos de: i) Elaboração de uma equação de calibração e ii) Validação.

i) Elaboração de uma equação de calibração

A equação de calibração foi obtida por regressão considerando como variável dependente a variável de demanda escolhida pela etapa de análise geoestatística. As demais variáveis domiciliares representadas na Tabela 4.2 são utilizadas como variáveis independentes para calibração da equação.

ii) Validação

A estimação da variável dependente foi realizada em coordenadas de valores conhecidos para o conjunto de validação e posteriormente foi binarizada, para estar referente à preferência por transporte individual motorizado ("0") e à preferência por transporte público ("1").

Novamente, o processo de validação teve sua eficácia avaliada através do uso de medidas estatísticas de comparação entre os valores previstos e observados. As medidas adotadas foram as mesmas, representadas na Tabela 4.6.

Também para esta abordagem, foi realizada a validação por município, considerando a medida estatística de porcentagem de acertos.

4.1.4 Comparação dos resultados

Finalmente, os resultados das medidas estatísticas da modelagem tradicional foram comparados aos resultados da modelagem geoestatística nas coordenadas geográficas de valores conhecidos referentes a 30% da amostra.

Atenta-se ao fato de que o trabalho não objetiva a comparação entre a modelagem tradicional e a modelagem geoestatística, considerando que a primeira pondera fatores socioeconômicos e de viagens, enquanto que a segunda analisa a influência das localizações e os valores da variável de interesse. O real objetivo é o de testar a adequabilidade do uso da Geoestatística na análise de demanda por transportes.

5. RESULTADOS E DISCUSSÕES

O capítulo 5 descreve a análise e discussão dos resultados obtidos através do método descrito no capítulo 4. São apresentados resultados principais provenientes da modelagem geoestatística e tradicional, bem como comparações das duas técnicas considerando amostra de validação.

5.1 TRATAMENTO E VISUALIZAÇÃO DE DADOS

O tratamento e a visualização de dados visaram à triagem das informações. Dessa forma, no estudo de variáveis relacionadas ao comportamento das viagens, os domicílios em que o número de viagens era nulo foram eliminados. Inicialmente, foram utilizadas duas amostras, considerando: i) a variável dicotômica de *preferência por modo motorizado* e ii) as demais variáveis.

i) Amostra considerando a variável dicotômica

Por se tratar de uma variável binomial, a *preferência por modo motorizado* teve tratamento diferenciado para o banco de dados quando comparado às demais variáveis. O tratamento de dados categorizou as variáveis de proporção de viagens por transporte público e transporte individual motorizado entre “1” e “0”, respectivamente. Foram selecionados 18.733 domicílios.

ii) Amostra considerando as demais variáveis

Para prosseguir com o estudo das variáveis de: *proporção de viagens por modo de transporte, total de viagens por domicílio, distância de viagem total* e por *modo de transporte*, foram eliminados domicílios em que havia viagens, porém o valor da distância das viagens era nula. No total, 23.636 domicílios foram selecionados.

5.2 ESCOLHA DA VARIÁVEL OBJETO DE ESTUDO

Uma forma de análise univariada pode ser obtida por meio da representação de frequências. Os histogramas das variáveis: *proporção de viagens por modo de transporte* (3), *distância de viagens por modo de transporte* (3), *distância total de viagens* (1) e *total de viagens por domicílio* (1) foram representados na Figura 5.1. No entanto, não há necessidade da representação para a variável *preferência por modo motorizado*, pois o percentual de observações da variável indica que a frequência de domicílios com preferência para o transporte público (55,3%) é maior do que a frequência de domicílios com preferência para o transporte individual motorizado (44,8%).

Figura 5.1: Histogramas das variáveis de demanda por transporte.

Nota-se que os histogramas apresentam grande quantidade de registros com valores nulos para as variáveis *proporção de viagens por modo de transporte* e *distância de viagens por modo de transporte* Figura 5.1 (a, b, c, d, e, f). Isto acontece, pois nos domicílios pode não existir pessoas que realizam viagens por um determinado modo. Por exemplo, caso uma família não possua automóvel ou motocicleta, o domicílio apresentará valor nulo para as variáveis de *proporção de viagens por transporte individual motorizado* e *distância de viagens por transporte individual motorizado*. Devido ao fato de estas seis variáveis apresentarem tal característica, a previsão por meio da Geoestatística se torna comprometida. Vale ressaltar, ainda, a pouca variabilidade dessas variáveis, pois apresentam as principais medidas descritivas (média, moda, mediana, desvio padrão – Tabela 4.3) nulas.

Embora a *distância total de viagens* e *total de viagens por domicílio* apresentassem estatísticas univariadas mais adequadas, realizou-se uma análise exploratória de variografia para tais variáveis. Esta análise demonstrou semivariogramas sem tendências espaciais claras.

Com isso, a variável *preferência por modo motorizado* se torna mais adequada ao prosseguimento deste estudo de estatística espacial, sendo esta abordada como variável objeto de estudo.

5.3 MODELAGEM GEOESTATÍSTICA

Através da identificação da variável objeto de estudo, o processo de modelagem geoestatística consiste das etapas de: análise da estrutura espacial, Krigagem e validação.

5.3.1 Análise da estrutura espacial

Nesta etapa, seguiram-se os passos de: elaboração de semivariogramas experimentais, ajuste de semivariogramas e verificação da direção principal e da respectiva direção ortogonal. O produto desta etapa é a definição dos parâmetros gráficos da direção principal e ortogonal, que são utilizados na Krigagem.

A elaboração dos semivariogramas experimentais direcionais e das respectivas direções ortogonais procedeu-se a partir do ângulo 0° (Direção Norte-Sul) até 90° (Direção Leste-Oeste), conforme padronização de eixos do *software* GeoMS, variando de 15° em 15° e tolerância angular de 1°. O número de lags utilizado foi o máximo permitido pelo *software*, 100 lags. Os tamanhos do lag (h) foram obtidos através da distância euclidiana (distância métrica) da amostra de pontos na RMSP para todos os ângulos de teste. A distância de corte foi de 15.000 metros, considerado coerente, pois se trata de um valor menor que a distância entre o par de domicílios mais afastados entre si da amostra (142 km). A Tabela 5.1 apresenta os parâmetros de cálculo do semivariograma experimental para todos os ângulos de teste.

Tabela 5.1: Parâmetros para cálculo dos semivariogramas experimentais.

Tabela 07.1: Parâmetros para cálculo dos semivariogramas experimentais.				
Ângulo de teste	Lag (m)	Tolerância	Número de lags	Distância de corte (m)
1°	146	180°	100	15.000
0°	1.794	1°		
-90°	975			
15°	2.077			
-75°	207			
30°	1.260			
-60°	299			
45°	1.071			
-45°	272			
60°	1.352			
-30°	596			
75°	787			
-15°*	924			

Os semivariogramas resultantes da análise da estrutura espacial da variável de *preferência por modo motorizado* são apresentados na Figura 5.2 e 5.3.

Figura 5.2: Semivariogramas experimentais nas direções: omnidirecional, (0°, -90°), (15°, -75°), (30°, -60°) e (45°, -45°) para *preferência por modo motorizado*.

Figura 5.3: Semivariogramas experimentais nas direções (60°, -30°) e (75°, -15°) para *preferência por modo motorizado*.

A direção principal foi verificada, para este trabalho, de forma conjunta ao ajuste do semivariograma experimental. *A priori*, constatou-se que a curva exponencial possui melhor ajuste, dentre as três opções aqui sugeridas (ajuste exponencial, esférico, gaussiano).

Os semivariogramas experimentais apresentaram tendências distintas, porém seus pontos de variâncias (em azul, na Figura 5.2 e Figura 5.3) demonstraram estacionariedade com patamar de 0,247, referente à média de variâncias de todos os pontos da amostra (linha vermelha). Esta observação denuncia a presença da estrutura espacial de anisotropia geométrica, em que existe um único patamar, contudo as amplitudes variam em função da direção.

O ajuste foi realizado para todas as direções de teste, considerando o efeito pepita (C_0) igual ao menor valor observado (0,1869). Consequentemente, a contribuição (C_1) foi de 0,0601. O efeito pepita teve valor alto (76% do valor do patamar), porém aceitável, considerando a natureza dos dados.

Definidas as medidas de C_0 , C_1 e patamar, ajustaram-se curvas exponenciais a todos os semivariogramas experimentais da Figura 5.2 e 5.3. Como resultado, foram definidas as amplitudes e as razões entre as amplitudes ortogonais,

apresentadas na Figura 5.4, juntamente com a representação do elipsóide de anisotropia geométrica.

Figura 5.4: Amplitudes e razões das direções estudadas e elipsóide de anisotropia geométrica.

A partir da Figura 5.4, nota-se que o maior valor de razão foi para as direções de 75° e -15°. A maior amplitude dentre as duas direções é representada pelo ângulo de -15°, definido como ângulo de direção principal da variável de *preferência por modo motorizado* deste trabalho.

A Figura 5.5 apresenta os resultados de ajuste da curva (em verde) e os parâmetros gráficos.

Figura 5.5: Semivariogramas ajustados para a direção principal (-15°) e direção ortogonal (75°).

A partir dos resultados de direção principal e ortogonal e dos parâmetros gráficos, realizou-se a etapa de estimação por Krigagem.

5.3.2 Krigagem

A definição da direção principal e ortogonal permitiu a extração de parâmetros necessários para a composição do mapa krigado. O mapa com a superfície espacial representa um novo conjunto de pontos estabelecidos a partir de pontos conhecidos (70% da amostra). Foi escolhido, como espaçamento d , o valor de 100 metros. Para as dimensões dos pontos estudados e da área da RMSP, o espaçamento de 100 metros gerou 1594 células no eixo x e 1059 células no eixo y .

Como a variável objeto de estudo não se trata de uma probabilidade entre a preferência por transporte individual motorizado e transporte público, mas sim, a escolha, de forma discreta, entre os modos motorizados, utilizou-se a Krigagem Indicativa como forma de estimação discreta e binária. A Figura 5.6 representa o mapa krigado da variável *preferência por modo motorizado*.

Figura 5.6: Mapa krigado para a variável *preferência por modo motorizado*.

Na Figura 5.6, a cor verde (valor “1”) representa a preferência por transporte público, enquanto que a cor vermelha (valor “0”) indica a preferência por transporte individual motorizado. A cor cinza indica posições espaciais onde não foi realizada a estimação (Krigagem). Isto se deve ao fato de o elipsóide não ter alcance para tais áreas e de não haver informações próximas o suficiente para que o processo de Krigagem ocorresse.

5.3.3 Validação

Utilizando os 30% da amostra reservados para validação, foram adotadas: a) a validação pela média dos nós adjacentes e b) a validação cruzada.

a) Validação pela média dos nós adjacentes

Nesta etapa, deparou-se com o problema de que os pontos reservados para a validação não coincidiam com as adjacências das células espaçadas a 100 metros da superfície espacial. Dessa forma, fez-se necessário encontrar as quatro adjacências que envolviam cada um dos pontos do conjunto de validação e o valor destes foram interpolados por meio da Equação 4.2.

Sendo os valores destes pontos conhecidos de antemão, puderam-se realizar medidas estatísticas de comparação entre valores observados e estimados a fim de avaliar a qualidade da interpolação realizada. A Tabela 5.2 apresenta os resultados do teste chi-quadrado para avaliar a associação entre as variáveis binárias estimadas e observadas e contém a taxa de acertos para a validação pela média dos nós adjacentes (a).

Tabela 5.2: Resultados de validação pela média dos nós adjacentes (modelagem geoestatística).

Chi-quadrado	556
Significância	0,00
Grau de liberdade	1
Taxa de acertos	0,67

O teste chi-quadrado objetiva a aceitação ou não da Hipótese Nula (BARBETTA, 2012). Neste caso:

- $\left\{ \begin{array}{l} \text{Hipótese Nula} \rightarrow \text{as variáveis estimadas e observadas não estão associadas.} \\ \text{Hipótese Alternativa} \rightarrow \text{as variáveis estimadas e observadas estão associadas.} \end{array} \right.$

Como chi-quadrado resultou em um valor alto, sugere-se que exista associação entre os valores estimados e os observados, rejeitando-se a Hipótese Nula. Essa indicação é confirmada pelo valor de significância obtido, que é menor que o valor crítico (α) padrão do *software* (0,05).

b) Validação cruzada

De maneira análoga à etapa a), a validação cruzada (teste do ponto fictício) valeu-se dos 30% da amostra para validação. Neste teste, assumiu-se um ponto como desconhecido e interpolou-se seu valor por meio de Krigagem, utilizando o restante dos pontos reservados para validação mais os parâmetros de Krigagem advindos do processo de calibração (70% da amostra). A Tabela 5.3 apresenta resultados do teste chi-quadrado e contém a taxa de acertos para a validação cruzada (b).

Para a validação cruzada, chi-quadrado também resultou em um valor alto, sugerindo a dependência entre os valores estimados e observados. O valor de significância obtido foi menor do que o valor crítico (α), confirmando a rejeição da Hipótese Nula.

Tabela 5.3: Resultados da validação cruzada (modelagem geoestatística).

Chi-quadrado	524
Significância	0,00
Grau de liberdade	1
Taxa de acertos	0,66

5.3.3.4 Validação por município

A validação por município se deu pela utilização do conjunto referente a cada um dos 39 municípios. Os resultados da validação por município podem ser visualizados no Apêndice 1. A porcentagem de acertos para a validação pela média dos nós adjacentes está representada, por classes, na Figura 5.7.

Figura 5.7: Mapa da taxa de acertos para a abordagem geoestatística (validação pela média dos nós adjacentes).

Com exceção de Salesópolis, que teve 100% de acertos, em geral, os resultados de porcentagem de acertos se apresentam maiores aonde existe maior quantidade de registros. Contudo, São Paulo apresentou menor taxa em relação a alguns municípios da periferia, demonstrando que o modelo de calibração se adequou bem a tais municípios, apesar de 73% dos registros estarem no município de São Paulo.

A representação, por classes, para a taxa de acertos da validação cruzada é referente à Figura 5.8. Quando ambas os mapas das abordagens geoestatísticas são comparadas, é notável que, para a validação cruzada, houve maior número de municípios com taxa de acertos nula. Da mesma forma, os resultados para a

validação cruzada indicaram menores taxas de acertos, com exceção de Embu-Guaçu e Suzano, que tiveram taxas de acertos muito próximas, mas maiores.

Figura 5.85.1: Mapa da taxa de acertos para a abordagem geoestatística (validação cruzada).

5.4 MODELAGEM TRADICIONAL

A modelagem tradicional se deu pela análise da variável de *preferência por modo motorizado*. Sendo esta variável do tipo dicotômica, modelos de regressão que utilizam ajustes por métodos de mínimos quadrados (regressão linear) não são adequados a este tipo de variável. O modelo mais correto para esta proposta é o modelo de Regressão Logística Múltipla (RLM).

5.4.1 Regressão Logística Múltipla

A Regressão Logística Múltipla foi escolhida para a variável dependente de *preferência por modo motorizado*, por incluir mais de uma variável independente. As variáveis independentes domiciliares utilizadas para a regressão foram: *número total*

de viagens, quantidade de automóveis, quantidade de motocicletas, quantidade de bicicletas, número de pessoas, critério de renda (separado pelas classes A1, A2, B1, B2, C1, C2, D e E e representado pelos números de 1 a 8, respectivamente), distância total de viagens, distância de viagens por transporte público, distância de viagens por transporte individual motorizado e distância de viagens por transporte não motorizado.

A regressão logística resultou em valores significativos para a maioria dos coeficientes correspondentes a cada variável independente. As exceções foram as variáveis *quantidade de bicicletas* e *distância total de viagens por domicílio*, as quais não foram consideradas significativas. Por esta razão, estas variáveis foram suprimidas por não influenciarem os resultados do ajuste do modelo. A equação de calibração para a RLM é demonstrada na Equação 5.1.

$$\pi = \frac{1}{1 + e^{-(g(x))}}$$

$$g(x) = -0,547 + 0,271\beta_1 + 0,318\beta_2 - 0,878\beta_3 - 1,180\beta_4 - 0,172\beta_5 + 0,000193\beta_6 - 0,000177\beta_7 + 0,00003\beta_8 \quad (5.1)$$

Onde:

- 1 - número de pessoas,
- 2 - critério de renda,
- 3 - quantidade de motocicletas,
- 4 - quantidade de automóveis,
- 5 - total de viagens,
- 6 - distância de viagens transporte público,
- 7 - distância de viagens por transporte individual motorizado,
- 8 - distância de viagens por transporte não motorizado.

A qualidade do ajuste pode ser mensurada através de medidas estatísticas, como erro padrão e significância; e também por medidas estatísticas apropriadas apenas para a análise de RLM, tais como: valor Wald, valor da verossimilhança

(-2LL), testes de regressão de Cox & Snell e Nagelkerke, chi-quadrado avaliado pelo teste de Hosner e Lemeshow. Tais medidas encontram-se na Tabela 5.4.

Tabela 5.4: Resultados para qualidade do ajuste da RLM.

		Valor observado		% de acertos	Wald	130,9
					-2LL	4.214
		0	1		Cox & Snell	0,652
Valor Previsto	0	5662	264	95,5	Nagelkerke	0,873
	1	294	6947	95,9	Hosner & Lemeshow	7.250
					Erro padrão	0,018
					Significância	0,000

A análise da Tabela 5.4 indica que houve um ótimo ajuste pela RLM. Os valores de erro padrão e significância foram extremamente baixos. Tais medidas estatísticas têm caráter genérico. Entretanto, para as medidas de caráter específico, resultados igualmente satisfatórios foram obtidos.

Similar ao coeficiente de determinação R^2 utilizado para casos de Regressão Linear, por exemplo, as medidas Cox & Snell e Nagelkerke têm valor ótimo quando tendem a “1”. No caso da RLM atual, obtiveram-se valores próximos ao valor ótimo.

As estatísticas Wald, -2LL e Hosner & Lemeshow também atenderam às expectativas. Contudo, a estatística que deixa mais evidente a qualidade do ajuste é a matriz de classificação, onde a contagem de valores previstos e observados é comparada e a taxa de acertos é mensurada. A taxa de acertos foi superior a 95%.

5.4.2 Validação

De maneira análoga à validação realizada por dois procedimentos na modelagem geoestatística (a e b), realizou-se a validação na modelagem tradicional. A Tabela 5.5 apresenta resultados do teste chi-quadrado e contém os resultados de validação para RLM.

O valor alto para o teste chi-quadrado indicou dependência entre os valores estimados e observados. Como o valor de significância obtido foi menor do que o valor crítico (α), confirmou-se a rejeição da Hipótese Nula.

Tabela 5.5: Resultados de validação para RLM.

Chi-quadrado	4.529
Significância	0,00
Grau de liberdade	1
Taxa de acertos	0,96

5.4.2.1 Validação por município

Os resultados de taxa de acertos para a validação por município para o caso da RLM podem ser visualizados no Apêndice 1. A Figura 5.9 apresenta no mapa os resultados, por classes, para a taxa de acertos para a validação pela abordagem tradicional.

A modelagem tradicional apresentou resultados satisfatórios quando avaliada a taxa de acertos por município. Percebe-se que o modelo atendeu com qualidade as cidades periféricas, que em geral, tiveram 100% de acertos, com exceção de Guararema, que teve 67% de acertos.

Figura 5.9: Mapa da taxa de acertos por município para a RLM.

5.5 COMPARAÇÃO DOS RESULTADOS

Através da extração dos resultados para os casos de modelagem geoestatística e para o caso de RLM, foram promovidas comparações a fim de a porcentagem de acertos.

O teste chi-quadrado observado nas três abordagens indicou um valor maior para a RLM. Dessa forma, é possível afirmar que existe uma maior associação entre os valores estimados e observados na modelagem tradicional em comparação à abordagem geoestatística.

Quanto às taxas de acertos (Figura 5.10), nota-se que a RLM teve uma maior eficácia, quando comparada às técnicas geoestatísticas. Isso se deve ao fato de que na RLM mais dados foram envolvidos no processo de calibração, sobretudo variáveis importantes para estimação da escolha modal como posse de automóveis, por exemplo. Enquanto a RLM utilizou oito variáveis independentes na elaboração do modelo de calibração, as técnicas geoestatísticas valeram-se apenas da localização da variável dicotômica. Ou seja, o método RLM utilizou quatro vezes mais informação do que as técnicas geoestatísticas.

No âmbito desta dissertação de mestrado, esse resultado é de grande valia, pois as técnicas geoestatísticas não têm por objetivo substituir as modelagens tradicionais de previsão de demanda. Embora, tenha-se obtido taxas de acertos de aproximadamente 1,4 vezes menor do que as taxas para RLM, este pode ser considerado um resultado apreciável, levando-se em conta que sua obtenção foi feita utilizando quatro vezes menos informação.

Figura 5.10: Gráfico de comparação de taxas de acertos entre a modelagem Geoestatística (média dos nós adjacentes e validação cruzada) e a modelagem por RLM.

Os valores de taxa de acertos obtidos pelos métodos geoestatísticos, apesar de não serem tão altos quanto a RLM, ainda assim credibilizam seus usos para estimação de valores em coordenadas de valores previamente desconhecidos.

Os resultados da modelagem tradicional foram muito satisfatórios. Apesar de 77% dos registros estarem referentes ao município de São Paulo, a equação de validação pela RLM obteve taxas de acertos superiores a 80%, com única exceção do município de Guararema (67%). Para este município, a validação pela média dos nós adjacentes teve 33% de acertos, enquanto para a validação cruzada, não houve acertos.

Dentre as abordagens, a única que teve porcentagens de acertos nulas foi a validação cruzada, que apresentou três municípios com taxas de 0%.

Comparando os resultados por município, a validação pela média dos nós adjacentes obteve taxa de acertos maiores para a maioria dos municípios do que a validação cruzada, com exceção de dois municípios, cujas taxas foram muito similares. Dessa forma, torna-se certo dizer que, para este estudo de caso (número alto de observações e tamanhos de lag relativamente baixos), dentre as modelagens geoestatísticas, é mais conveniente realizar a validação pela média dos nós adjacentes.

O município de São Paulo apresentou taxas de acertos de 66% e 67% para as validações geoestatísticas, enquanto que, para a RLM, a proporção foi de 96%.

A validação pela média nos nós adjacentes e a RLM demonstraram que os municípios periféricos obtiveram maiores taxas de acertos, tendo estes municípios menor número de registros (Figura 5.11).

Figura 5.11: Proporção de registros por município da RMSP.

6. CONCLUSÕES E RECOMENDAÇÕES

5.1 CONCLUSÕES

Este trabalho de mestrado teve como propósito o estudo e a aplicação da Geoestatística na área de demanda por transportes. Objetivou-se a análise de dados desagregados de demanda por transportes através de modelagem geoestatística e tradicional. Assim, a utilização da Geoestatística propicia a consideração da dependência espacial na análise, enquanto que a abordagem tradicional considera aspectos socioeconômicos e características de viagens.

Dispondo de um banco de dados desagregado da Pesquisa O/D da Região Metropolitana de São Paulo, inicialmente realizou-se a primeira etapa do método, tratamento e visualização dos dados, onde foram eliminados possíveis empecilhos como viagens nulas e redundâncias de localizações domiciliares. Neste momento, parte dos dados já estava apta ao trabalho, bastando apenas uma última adequação dos dados para a criação de uma variável dicotômica denominada *preferência por modo motorizado*.

Na sequência, realizou-se a análise univariada das variáveis. Percebeu-se que as variáveis de *proporção de viagens por modo de transporte* e *distância de viagens por modo de transporte* apresentaram mais de 50% dos registros nulos. As variáveis de *total de viagens por domicílio* e *distância total de viagens* tiveram uma análise prévia de variografia não satisfatória. Dessa forma, a variável objeto de estudo foi a *preferência por modo motorizado*.

Escolhida a variável objeto de estudo, aplicaram-se a modelagem geoestatística e a modelagem tradicional para previsão da variável. Como a variável de interesse escolhida foi a variável dicotômica, tornou-se necessária a escolha da melhor forma de modelagem em ambas as abordagens. Isso quer dizer que para o caso tradicional, recorreu-se à Regressão Logística Múltipla, e para o caso de

modelagem geoestatística, recorreu-se à Krigagem Indicativa. Ambos os procedimentos são aplicáveis quando se utiliza uma variável dependente binária. Contudo, a utilização da Krigagem Indicativa não resultou estimativas de probabilidade, pois se trata de uma variável categórica.

A modelagem geoestatística teve como primeiro passo o ajuste das curvas e a definição da direção principal. A curva com melhor ajuste para os dados deste trabalho foi a exponencial. Diante de todos os semivariogramas para as direções de teste de 15°, verificou-se que a estacionariedade era atingida para um patamar similar a todas as direções, contudo as amplitudes variavam. Dessa forma, para o ajuste, considerou-se a existência de anisotropia geométrica, em que a direção principal foi relativa ao par de direções com maior razão de amplitudes (-15°, 75°).

Os semivariogramas apresentaram efeitos pepitas de 76% do valor do patamar. Este fenômeno pode ser explicado pela diferença alta entre os valores da variável entre pares de pontos amostrais (domicílios) próximos entre si. Por exemplo, uma família faz somente viagens de carro, enquanto a vizinhança não. No caso prático do planejamento de transportes, ocorre muito esse episódio, pois envolve o comportamento humano. Isso faz com que o efeito pepita seja alto.

O processo posterior, de estimativa (Krigagem), foi realizado para células de espaçamento de 100 metros. O resultado demonstrou áreas com preferência por transporte individual motorizado, áreas com preferência por transporte público e houve áreas onde não foi realizada a estimativa. Isto ocorreu, pois em algumas áreas não havia informações próximas o suficiente para que o processo de Krigagem ocorresse.

A validação para o método geoestatístico se deu pela média dos nós adjacentes e pelo teste do ponto fictício (validação cruzada). Em geral, os resultados de porcentagem de acertos para toda a RMSP foram similares, ambos próximos a 67%. Contudo, a validação por município, indicou que a validação pela média dos nós adjacentes, em comparação com a validação cruzada, teve maiores taxas de acertos para todos os municípios, com exceção de Embu-Guaçu e Suzano. Dessa

forma, a validação pela média dos nós adjacentes foi mais eficiente, dentre as validações do método geoestatístico. Isto aconteceu, pois o processo de validação cruzada utilizou apenas 30% da amostra para Krigagem, juntamente com os parâmetros gráficos do semivariograma (calibração).

Para este trabalho, cuja área é extensa e o banco de dados possui informações de 30.000 domicílios, as abordagens de validação foram válidas. Quando a área de estudo é menor e as estimações são propostas para centenas, ou até dezenas, de pontos, a validação pela geoestatística pode ser realizada diretamente pela leitura no mapa krigado. Neste contexto de áreas de estudo menores, verifica-se que a elaboração de mapas krigados com células com menores distâncias entre si, por exemplo, distâncias unitárias, tornam o processo de validação mais acurado.

A modelagem tradicional utilizou informações de oito variáveis para a calibração: *número de pessoas, critério de renda, quantidade de motocicletas, quantidade de automóveis, total de viagens, distância de viagens transporte público, distância de viagens por transporte individual motorizado, distância de viagens por transporte não motorizado*. Nesta abordagem, foram obtidas taxas de acertos superiores a 80% para os municípios, com exceção de Guararema. Considerando a RMSP, a taxa de acertos foi de 96%.

Pôde-se perceber que a maior parte dos domicílios pesquisados se concentra nas áreas mais populosas, principalmente no município de São Paulo, em que se concentram 77% dos registros. Isto ocorre, pois a Pesquisa O/D é realizada a partir de uma amostragem proporcional aleatória baseada na população. Dessa forma, as estimações geoestatísticas tornam os municípios mais afastados mais suscetíveis a erro, exatamente pela generalização do comportamento da variável. Contudo, os resultados apresentaram taxas de acertos eventualmente até maiores para as áreas periféricas.

O trabalho aqui exposto apresentou um método que se adéqua melhor às características da amostra e da natureza das informações (grande número de

informações, inconstância dos dados e possíveis erros no processo da Pesquisa O/D). É importante salientar que esse método é particular para esse caso. Se houvesse um tipo de informação menos inconstante e mais preciso (como por exemplo, no caso da hidrologia, em que se observa a contaminação de bacias hidrográficas por meio de processos de análises laboratoriais), os dados seriam mais escassos e acurados, conseqüentemente teriam outro tipo tratamento na geoestatística.

De modo geral, os resultados para os dois enfoques, geoestatístico e tradicional, foram extremamente satisfatórios. O valor chi-quadrado se mostrou alto para os três métodos de validação, sendo o valor de chi-quadrado para RLM maior, indicando maior associação entre os valores estimados e observados na modelagem tradicional em comparação à geoestatística. Os valores para taxa de acertos enfatizam tal afirmação. A abordagem tradicional obteve uma taxa de acertos de 96% enquanto a abordagem geoestatística apresentou uma taxa de acertos de aproximadamente 67%. Levando-se em conta a abordagem tradicional, a altíssima taxa de acertos se deve ao fato de que a regressão logística utilizada foi a múltipla, isso quer dizer que a variável dependente dicotômica foi modelada por uma equação envolvendo oito outras variáveis independentes (por essa razão chamado caso múltiplo). Entretanto, a dependência espacial da variável dicotômica, considerada pela coordenada geográfica, não foi utilizada.

A coordenada geográfica foi utilizada na análise geoestatística, que se valeu somente da localização espacial e de valores conhecidos da variável de interesse para a estimação. Uma vez que menos dados foram utilizados, é esperado que a taxa de acertos seja menor, pois um montante menor de informação foi utilizado. Entretanto, o resultado de 67% é aceitável, pois a análise geoestatística não objetiva substituir a análise tradicional de regressão, mas sim tentar colaborar para elevar as taxas de acertos desse tipo de análise. Os 67% de acertos foram gerados utilizando dados complementares aos dados que geraram a taxa de acertos de 96% da regressão. Isso gera uma perspectiva enorme para excelentes resultados combinando os dois métodos, que além de estimar valores em coordenadas de

domicílios já pesquisados, torna possível a previsão de variáveis em outras localizações. Isto faz com que a Geoestatística certamente seja de grande valia e contribuição para a área de planejamento e previsão de demanda por transportes.

5.2 RECOMENDAÇÕES

Os estudos associados à Geoestatística ainda têm uma longa trajetória para tornarem a técnica uma ferramenta incremental a métodos tradicionais de previsão de demanda por transportes. A fim de dar prosseguimento ao aprendizado a partir dos conhecimentos já existentes na área e, obtidos por meio deste trabalho, são feitas algumas sugestões a seguir.

Conforme este trabalho, que comparou métodos com diferentes informações de entrada, recomenda-se que sejam ainda acrescentadas as coordenadas geográficas a modelos de regressão para testar de forma justa as abordagens.

Pode-se agregar outras características locais a modelos tradicionais, além da localização espacial, como: variâncias em diferentes raios de distância, autocorrelação espacial (conceitos de semi-variogramas) e índice de Moran.

Uma recomendação de estudos futuros segue pela combinação da Geoestatística a métodos clássicos. Essa combinação pode ocorrer pela inserção de resultados das técnicas tradicionais como informação secundária para Krigagem com Deriva Externa.

Sugere-se, também, uma complementação da proposta deste trabalho para a previsão modal individual, além da preferência domiciliar.

REFERÊNCIAS BIBLIOGRÁFICAS

BARBETTA, P. A. (2012) **Estatística aplicada às Ciências Sociais**. 8ª ed. rev. Editora da UFSC, Florianópolis.

BRAXMEIER H.; SCHMIDT V.; SPODAREV E. (2005) Kriged road-traffic maps. Em: Pilz J (ed) Interfacing geostatistics, GIS and spatial data bases. **Proceedings of the international conference**, StatGIS 2003. Advances in spatial science. Springer, Berlim, p. 39–52

CAMARGO, E. C. G. (1998) **Geoestatística: Fundamentos e aplicações. Geoprocessamento para projetos ambientais**. São José dos Campos: INPE.

CAMARGO, E. C. G.; FELGUEIRAS, C. A.; MONTEIRO, A. M. V. (2001) A importância da modelagem da anisotropia na distribuição espacial de variáveis ambientais utilizando procedimentos geoestatísticos. Anais do X **Simpósio Brasileiro de Sensoriamento Remoto**, Foz do Iguaçu, 21-26 Abril de 2001, INPE, p. 395-402.

CARVALHO, L. E.; SILVA, H. N.; LOUREIRO, C. F. G.; MENESES, H. B. (2006) Regressão linear geograficamente ponderada em ambiente SIG. **Transportes**. v. XIV, nº2, p. 18-26, Dezembro de 2006.

CORREIA, P. (2010) **Modelação e estimação**: uma introdução à Geoestatística. Apostila do Núcleo de Minas do Instituto Superior Técnico. Universidade de Lisboa.

COSTA, A. S. G.; PITOMBO, C. S.; SALGUEIRO, A. R. (2013) Estimação de escolha modal através da geoestatística. Anais do XXVII **Congresso de Pesquisa e Ensino em Transportes** (ANPET). Belém, Pará. 04 a 08 de novembro de 2013.

DRUCK, S.; CARVALHO, M.S.; CÂMARA, G.; MONTEIRO, A.V.M. (eds) (2004) **Análise Espacial de Dados Geográficos**. Brasília, EMBRAPA, 2004 (ISBN: 85-7383-260-6) 209 p.

DU J.; AULTMAN-HALL L. (2006) Using spatial analysis to estimate link travel times on local roads. **Transportation Research Board** (TRB).

FIGUEIRA, C. V. (2006) **Modelos de regressão logística**. Dissertação (Mestrado em Matemática) – Instituto de Matemática da Universidade Federal do Rio Grande do Sul, Porto Alegre.

GONÇALVES, A. C. A.; FOLEGATTI, M. V.; MATA, J. D. V. (2001) Análise exploratória e geoestatística da variabilidade de propriedades físicas de um Argissolo Vermelho. **Acta Scientiarum**. Maringá, v. 23, nº 5, p. 1149-1157.

HAIR JR, J. F.; BLACK, W. C.; BABIN, B. J.; ANDERSON, R. E. (2010) **Multivariate Data Analysis**. Prentice Hall. 7ª ed. 785 p.

HANSON, S.; GIULIANO, G. (2004) **The Geography of Urban Transportation**. Guildford Press 3ª ed. 419 p.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATISTICA – IBGE. (2010) Censo Demográfico Brasileiro 2010. Disponível em: < <http://cidades.ibge.gov.br/> >. Acesso em 02 de março de 2015.

KRAMER, R. (1998) **Chemometric techniques for quantitative analysis**. Nova Iorque: Marcel Dekker, Inc. 202 p.

KREMPI (2004) A. P. **Explorando recursos de estatística espacial para análise da acessibilidade da cidade de Bauru**. Dissertação (Mestrado em Engenharia Civil) – Escola de Engenharia de São Carlos da Universidade de São Paulo, São Carlos.

LANDIM, P. M. B. (2003) **Análise Estatística de Dados Geológicos**. 2ª ed. São Paulo: Universidade Estadual Paulista, 254p.

LANDIM, P. M. B. (2004) **Introdução à análise variográfica com o Variowin**– UNESP – texto didático 14.

LANDIM, P. M. B.; STURARO, J. R. (2002) **Krigagem Indicativa** aplicada à elaboração de Mapas Probabilísticos de Riscos. Universidade Estadual Paulista. Rio Claro.

LAPA, T. M.; ALBUQUERQUE, M. F. P. M.; CARVALHO, M. S.; SILVEIRA JR, J. C. (2006) Análise da demanda de casos de hanseníase aos serviços de saúde através do uso de técnicas de análise espacial. **Cadernos de Saúde Pública**. Rio de Janeiro, vol. 22, nº 12, pp. 2575-2583, Dezembro de 2006.

LEUNG, J. K. Y. (2002) Kriging analysis on Hong Kong rainfall data. **The Hong Kong Institution of Engineers**. v. 9. Número 1. Abril de 2002.

LIMA, M. L. A.; MELO JR, H. R.; CHAVES, C. A. P.; SIMÃO, F. B. (2008) Krigagem Indicativa aplicada à pesquisa de nitrato nas águas subterrâneas da área urbana de Porto Velho – RO. **XV Congresso Brasileiro de Águas Subterrâneas**. Natal. 11 a 14 de novembro de 2008.

LOPES, S. B. (2005) **Efeitos da dependência espacial em modelos de previsão de demanda por transporte**. Dissertação (Mestrado em Engenharia Civil) – Departamento de Transportes da Escola de Engenharia de São Carlos da Universidade de São Paulo, São Carlos.

LUZ, E. R. (2003) **Predição de propriedades de gasolinas usando espectroscopia FTIR e regressão por mínimos quadrados parciais**. Dissertação (Mestrado em Química) – Departamento de Química da Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro.

MATHERON, G. (1971) **The theory of regionalized variables and its applications**. Les Cahiers du Centre de Morphologie Mathématique de Fontainebleau. École Nationale Supérieure des Mines de Paris nº 5. 211 p.

MELLO, J. M. (2004) **Geoestatística aplicada ao inventário florestal**. Tese (Doutorado em Recursos Florestais) - Universidade de São Paulo, Piracicaba.

METRÔ (2007) Companhia do Metropolitano de São Paulo. **Pesquisa Origem e Destino**. Disponível em < <http://www.metro.sp.gov.br/> > Acesso em: 8 de Janeiro de 2014.

MEYER, J. R.; STRASZHEIM, M. R. (1974) **Techniques of Transport Planning**. v.1 Washington: The Brookings Institution. 336p.

MIURA, H. (2010) A study of travel time prediction using universal kriging. **Sociedad de Estadística e Investigación Operativa**, p. 257-270.

NOVAES, A. G. (1982) **Modelos em Planejamento Urbano, Regional e de Transportes**. São Paulo: Edgard Blücher. 290p.

OLEA, R. A. (2006) A six step practical approach to semivariogram modeling **Stochastic Environmental Research and Risk Assessment**, v. 20. p.307–318.

OLIVEIRA, A. M. L. (2013) **Escolha entre o transporte interestadual aéreo e rodoviário de passageiros**: avaliação da distância como fator determinante. Dissertação (Mestrado em Regulação e Gestão de Negócios) – Universidade de Brasília, Brasília.

ORTÚZAR, J. D.; WILLUMSEN, L. G. (2011) **Modelling Transport**. Londres: Wiley. 4ª ed. 586p.

PANA, C. E. (2010) **Modelos de Transporte e Tráfego** Novas Tecnologias Transporte e Tráfego. Disponível em:

<http://www.sinaldetransito.com.br/artigos/novos_modelos_3.pdf>. Acesso em 02 de março de 2015.

PALHARES, C. R. C. (2008) **Desenvolvimento de subcentros como forma de planejamento urbano e de transportes em cidades de porte médio**. Dissertação (Mestrado em Engenharia Civil) - Faculdade de Engenharia Civil da Universidade Federal de Uberlândia, Uberlândia.

PEREIRA, L. S. F.; OLIVEIRA, L. K. (2014) Metodologia para estimativa de fluxos de carga a partir de dados secundários: uma aplicação em Belo Horizonte. **Journal of Transporte Literature**, vol. 8, n. 4, pp. 279-315.

PEREIRA, W. C. (2011) **Simulação geoestatística via métodos convolutivos**. Dissertação (Mestrado em Estatística) do Instituto de Ciências Exatas e Naturais da Universidade Federal do Pará, Belém.

PEROBELLI, F. S. (2004) **Análise espacial das interações econômicas entre os estados brasileiros**. Tese (Doutorado em Economia) – Universidade de São Paulo, São Paulo.

PITOMBO, C. S.; FERNANDES, V. O.; ALIXANDRINI JR, M. J.; SANTOS, D. V. C.; TEIXEIRA, T. C. S. (2010a) Aplicação conjunta de técnicas de análise multivariada de dados e análise espacial exploratória para avaliação do desempenho de transporte público por ônibus. Anais do XXIV **Congresso de Pesquisa e Ensino em Transportes** (ANPET). Salvador, Bahia. 29 de Novembro a 3 de Dezembro de 2010.

PITOMBO, C. S.; ROCHA, S. S.; SALGUEIRO, A. R.; BITTENCOURT, D. C. (2014) Uso de krigagem para estimação de geração de viagens. Em: 6° **Congresso Luso-Brasileiro para Planejamento Urbano, Regional, Integrado e Sustentável**, 2014, Lisboa. PLURIS 2014.

PITOMBO, C. S.; SALGUEIRO, A. R.; COSTA, A. S. G.; ISLER, C. A. (2015) A two-step method for mode choice estimation with socioeconomic and spatial information. **Spatial Statistics**. v. 11 p. 45-64. DOI: 10.1016/j.spasta.2014.12.002.

PITOMBO, C. S.; SOUSA, A. J. (2009) Aplicação de conceitos geoestatísticos para análise de geração de viagens urbanas. Anais do XXIII **Congresso de Pesquisa e Ensino em Transportes** (ANPET). Vitória, Espírito Santo. 9 a 13 de novembro de 2009.

PITOMBO, C. S.; SOUSA, A. J.; QUINTANILHA, J. A.; BIRKIN, M. (2010b) Comparing different spatial data analysis to forecast trip generation. 12° **World Conference on Transport Research**, 11 a 15 de Julho de 2010. Lisboa, Portugal.

ROCHA, S. S. (2014) **Análise de geração de viagens urbanas por transporte coletivo através de técnicas de geoestatística**. Dissertação (Mestrado em Engenharia Ambiental Urbana) – Escola Politécnica da Universidade Federal da Bahia, Salvador.

ROCHA, S. S.; LINDNER, A.; PITOMBO, C. S. (2014) Modelagem geoestatística aplicada à amostra sistemática desagregada: estimação de geração de viagens. Anais do XXVIII **Congresso de Pesquisa e Ensino em Transportes** (ANPET). Curitiba, Paraná. 24 a 28 de novembro de 2014.

RODRIGUES, F. S. P. (2012) **Modelos comportamentais desagregados: uma análise conceitual**. Trabalho final de curso (Graduação em Engenharia Civil). Escola de Engenharia da Universidade Federal do Rio Grande do Sul, Porto Alegre.

SIQUEIRA, H. R.; ALVES, G. F.; GUIMARÃES, E. C. (2007) Comportamento da precipitação pluviométrica mensal do estado de Minas Gerais: análise espacial e temporal. **Horizonte Científico**. vol. 1. n°1 (2007).

SOARES, A. (2006) **Geoestatística para as Ciências da Terra e do Ambiente**. 2ª ed. Lisboa: IST Press, 214p.

SOUSA, A. J.; MUGE, F. (1990) **Elementos da Geoestatística**. Lisboa: Laboratório de Mineralogia e Planejamento Mineiro. IST. 63p.

SOUZA, C. D. R.; D'AGOSTO, M. A. (2013) Modelo de quatro etapas aplicado ao planejamento de transporte de carga. **Journal of Transport Literature**, vol. 7, n. 2, pp. 207-234.

SOUZA, Z. M.; MARQUES JR, J.; PEREIRA, G. T. (2004) Variabilidade espacial da estabilidade de agregados e matéria orgânica em solos de relevos diferentes. **Pesquisa Agropecuária Brasileira**, Brasília, v. 39, nº5, p.491-499, Maio de 2004.

TEDESCO, G. M. I. (2008) **Metodologia para elaboração do diagnóstico de um sistema de transporte**. Dissertação (Mestrado em Transportes) – Departamento de Engenharia Civil e Ambiental da Faculdade de Tecnologia da Universidade de Brasília, Brasília.

TEIXEIRA, G. L. (2003) **Uso de dados censitários para identificação de zonas homogêneas para planejamento de transportes utilizando estatística espacial**. Dissertação (Mestrado em Transportes). Faculdade de Tecnologia da Universidade de Brasília, Brasília.

VASCONCELOS, E. A. (2000) **Transporte urbano nos países em desenvolvimento**: reflexões e propostas. Annablume. 282 p.

YAMAMOTO, J. K.; LANDIM, P. M, B. (2013) **Geoestatística: Conceitos e Aplicações**. São Paulo: Oficina de Textos. 215p.

LISTA DE ANEXOS E APÊNDICE

ANEXO 1: Mapa da Região Metropolitana de São Paulo

Legenda:

1	Arujá	14	Guararema	27	Ribeirão Pires
2	Barueri	15	Guarulhos	28	Rio Grande da Serra
3	Biritiba-Mirim	16	Itapecerica da Serra	29	Salesópolis
4	Caieiras	17	Itapevi	30	Santa Isabel
5	Cajamar	18	Itaquaquecetuba	31	Santana de Parnaíba
6	Carapicuíba	19	Jandira	32	Santo André
7	Cotia	20	Juquitiba	33	São Bernardo do Campo
8	Diadema	21	Mairiporã	34	São Caetano do Sul
9	Embu das Artes	22	Mauá	35	São Lourenço da Serra
10	Embu-Guaçu	23	Mogi das Cruzes	36	São Paulo
11	Ferraz de Vasconcelos	24	Osasco	37	Suzano
12	Francisco Morato	25	Pirapora do Bom Jesus	38	Taboão da Serra
13	Franco da Rocha	26	Poá	39	Vargem Grande Paulista

APÊNDICE 1: Resultados de taxa de acertos para a validação por município

Município	Média dos nós adjacentes	Validação cruzada	Regressão Logística Múltipla
Arujá	92%	92%	92%
Barueri	73%	70%	98%
Biritiba-Mirim	55%	55%	100%
Caieiras	73%	73%	100%
Cajamar	81%	69%	100%
Carapicuíba	74%	70%	98%
Cotia	71%	65%	100%
Diadema	75%	75%	94%
Embu das Artes	52%	48%	94%
Embu-Guaçu	61%	78%	100%
Ferraz de Vasconcelos	88%	88%	100%
Francisco Morato	93%	93%	100%
Franco da Rocha	79%	79%	93%
Guararema	33%	0%	67%
Guarulhos	63%	61%	94%
Itapecerica da Serra	88%	75%	100%
Itapevi	83%	78%	89%
Itaquaquecetuba	63%	50%	96%
Jandira	50%	23%	91%
Juquitiba	50%	0%	100%
Mairiporã	83%	0%	83%
Mauá	75%	75%	96%
Mogi das Cruzes	67%	62%	95%
Osasco	63%	52%	94%
Pirapora do Bom Jesus	85%	77%	100%
Poá	77%	77%	100%
Ribeirão Pires	71%	61%	84%
Rio Grande da Serra	71%	71%	93%
Salesópolis	100%	100%	100%
Santa Isabel	89%	67%	89%
Santana de Parnaíba	81%	75%	100%
Santo André	68%	43%	97%
São Bernardo do Campo	70%	55%	98%
São Caetano do Sul	78%	60%	91%
São Lourenço da Serra	71%	43%	100%
São Paulo	71%	67%	96%
Suzano	76%	78%	93%
Taboão da Serra	53%	37%	84%
Vargem Grande Paulista	67%	67%	100%