UNIVERSIDADE DE SÃO PAULO
ESCOLA DE ENGENHARIA DE SÃO CARLOS
DEPARTAMENTO DE ENGENHARIA DE TRANSPORTES

SAMILLE SANTOS ROCHA

Uso de Algoritmos Genéticos para otimização de modelagem
geoestatística aplicada à demanda por transportes

São Carlos

2019

SAMILLE SANTOS ROCHA

Uso de Algoritmos Genéticos para otimização de modelagem
geoestatística aplicada à demanda por transportes

Programa de Pós-graduação em
Engenharia de Transportes da EESC-USP

**Exemplar definitivo (corrigido). O exemplar
original está disponível na CPG da EESC-USP**

São Carlos, 02/05/2019

RESOLUÇÃO CoPGr N º 6018, DE 13 DE OUTUBRO DE 2011, artigo 5º

Tese apresentada à Escola de Engenharia de São Carlos/USP, como requisito para obtenção do título de doutora em Ciências - Programa de Pós-Graduação em Engenharia de Transportes.

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

Orientadora: Prof. Dr^a. Cira Souza Pitombo

São Carlos

2019

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

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

R672u	<p>Rocha, Samille</p> <p>Uso de Algoritmos Genéticos para otimização de modelagem geoestatística aplicada à demanda por transportes / Samille Rocha; orientadora Cira Souza Pitombo. São Carlos, 2019.</p> <p>Tese (Doutorado) - Programa de Pós-Graduação em Engenharia de Transportes e Área de Concentração em Planejamento e Operação de Sistemas de Transporte -- Escola de Engenharia de São Carlos da Universidade de São Paulo, 2019.</p> <p>1. Algoritmos Genéticos. 2. Otimização. 3. Semivariogramas. 4. Análise espacial. 5. Demanda por transportes. I. Título.</p>
-------	--

Eduardo Graziosi Silva - CRB - 8/8907

FOLHA DE JULGAMENTO

Candidato: Bacharel **SAMILE SANTOS ROCHA.**

Título da tese: "Uso de algoritmos genéticos para otimização de modelagem geoestatística aplica à demanda por transportes".

Data da defesa: 20/02/2019.

Comissão Julgadora:

Resultado:

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

Aprovada

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

APROVADA

Prof. Dr. **Luís Henrique Magalhães Costa**
(Universidade Veiga de Almeida/UVA)

Aprovada

Prof. Dr. **Antônio Jorge Gonçalves de Sousa**
(Universidade de Técnica de Lisboa/UTL)

Aprovada

Prof. Dr. **Jose Alberto Quintanilha**
(Escola Politécnica/EP-USP)

Aprovada

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

Profa. Associada **Ana Paula Camargo Larocca**

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

El. 50.15

AGRADECIMENTOS

A Deus, sempre presente em todos os momentos, me dando força para seguir em frente.

À minha família, em especial meus pais, Arlete e Roque, e à minha irmã Larissa, pelo apoio e incentivo ao longo da minha formação pessoal e profissional.

À minha orientadora, Cira Pitombo, por estes sete anos de parceria, ensinamentos e troca de conhecimento.

Ao Profº Luís Henrique (Universidade Estadual Vale do Acaraú – CE), pela parceria e contribuição na etapa da otimização.

Aos professores e funcionários do STT pela sabedoria transmitida, disponibilidade e colaboração nos momentos em que precisei de ajuda.

Aos professores Jaime Gómez-Hernández, Universidade Politécnica de Valência (Espanha) e José Alberto Quintanilha (POLI/USP) pela colaboração e sugestões em diversos momentos deste trabalho.

Aos professores Antônio Nélon (EESC/USP) e Jorge Sousa (Instituto Superior Técnico de Lisboa) pelas contribuições no exame de qualificação deste trabalho.

Aos amigos, amigas de São Carlos e do STT, que colaboraram direta e indiretamente com este trabalho, como Cíntia de Campos, Luan Carvalho, Marcos Vinícius Nicolino, Samuel Marques, Wana Dorigo e, especialmente, a Anabele Lindner pela amizade e colaborações nos projetos acadêmicos.

A todos e todas que estiveram presentes, de alguma forma, durante este período de doutorado, especialmente a Andressa Passos, Murilo Teixeira e Thaíse Sá pelo incentivo, paciência, companheirismo, amizade e torcida. A Oscar, minha melhor companhia de trabalho e escrita, desde a dissertação de mestrado.

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPQ), pelo auxílio financeiro.

RESUMO

ROCHA, S. S. **Uso de Algoritmos Genéticos para otimização de modelagem geoestatística aplicada à demanda por transportes**. 2019. 139p. Tese (Doutorado) – Departamento de Engenharia de Transportes, Escola de Engenharia de São Carlos, Universidade de São Paulo, São Carlos, 2019.

Geralmente, dados relacionados à demanda por transportes não são independentes no espaço. Por esta razão, o uso de técnicas de estatística espacial torna-se relevante para aprimoramento de estimativas. A geoestatística está entre as técnicas que incorporam a dependência espacial em suas análises e o semivariograma é uma ferramenta indispensável para descrever e ilustrar a estrutura espacial de uma Variável Regionalizada. Muitas vezes, o cálculo e ajuste do semivariograma experimental são realizados visualmente, de acordo a familiaridade com os dados ou experiência do pesquisador, o que exige, sobretudo, tempo de dedicação às análises. A partir destas considerações, técnicas de otimização podem ser uma alternativa viável para cálculo e ajuste de semivariograma experimental. Diante disso, este trabalho objetiva avaliar a adequabilidade do uso de Algoritmos Genéticos (AG) para otimização da modelagem geoestatística aplicada à demanda por transportes. Outro ponto importante é a forma de representação de dados de transportes: quando disponíveis, dados desagregados, por domicílios, comprometem a qualidade dos modelos variográficos, devido à sua alta aleatoriedade espacial. Diante disso, outra importante contribuição deste estudo foi a implementação de um código em *software* livre para definir a dimensão de uma grade para agregação de dados pontuais. A implementação do AG permitiu a obtenção de inúmeros modelos variográficos de duas variáveis relacionadas à demanda por transportes, para dois diferentes suportes geográficos. Além disso, foi possível obter os intervalos mais frequentes dos parâmetros dos semivariograma com melhor *fitness*. Finalmente, uma proposta primária de análise da semivariância foi apresentada, a fim de validar os resultados obtidos pelo AG. A análise de mapas de semivariância permitiu verificar o comportamento estrutural das variáveis estudadas. Apesar da abordagem tradicional (mapas de semivariância e ajuste manual) apresentar algumas dissimilaridades quando comparada aos melhores semivariogramas provenientes do AG, as medidas de desempenho, obtidas através da validação cruzada, foram bem similares. Conclui-se que a otimização da modelagem geoestatística, através de AG, pode trazer contribuições importantes, relativas a maior facilidade de cálculo e ajuste, além de distribuição de parâmetros variográficos associados a soluções quase ótimas. Vale ressaltar que o código desenvolvido ao

longo desta tese, disponível ao público, pode ser utilizado em qualquer área do conhecimento onde se verifique a existência de dependência espacial entre as observações.

Palavras-chave: Algoritmos Genéticos, Otimização, Semivariogramas, Análise Espacial, Demanda por Transportes.

ABSTRACT

ROCHA, S. S. **Genetic Algorithms for optimization of geostatistical modeling applied to transport demand**. 2019. 139p. Doctoral Theses – Department of Transport Engineering, São Carlos Engineering School, University of São Paulo, São Carlos, 2019.

Data related to travel demand are generally not independent in space. Due to this, using spatial statistics techniques is important for improving estimates. Geostatistics is among the techniques that incorporate spatial dependence in its analyses and the semivariogram is an indispensable tool to describe and illustrate the spatial structure of a Regionalized Variable. The calculation and fitting of the experimental semivariogram are often performed visually, according to the familiarity with the data or the researcher's experience, which requires, above all, time for the analyses. Based on these considerations, optimization techniques can be a viable alternative to calculate and fitting the experimental semivariogram. Therefore, this study aims to evaluate the adequacy of using Genetic Algorithms (GAs) to optimize geostatistical modeling applied to travel demand. Another important point is the way of representing travel data: when it is available, disaggregated data by households affect the quality of variographic models, due to their high spatial randomness. Therefore, another important contribution of this study was the implementation of a free software code to define the size of a grid for aggregation of point data. Implementing the GA enabled us to obtain numerous variographic models of two variables related to travel demand for two different geographical supports. In addition, the most frequent intervals of the semivariogram parameters could be obtained with better fitness. Finally, a primary proposal for semivariance analysis was presented in order to validate the results obtained by the GA. The semivariance analysis maps verified the structural behavior of the studied variables. In spite of traditional approach (semivariance maps and manual fit) to present some dissimilarities when compared to the best semivariograms from GA, the performance measures obtained through cross validation were very similar. It can be concluded that the geostatistical modeling optimization, through GA, can bring important contributions, related to making calculations and fits easier, as well as distribution of variographic parameters associated with almost optimal solutions. It is worth mentioning that the code developed throughout this thesis, available to the public, can be used in any area of knowledge where there is a spatial dependence between observations.

Keywords: Genetic Algorithms, Optimization, Semivariograms, Spatial Analysis, Travel Demand.

LISTA DE FIGURAS

Figura 1.1: Estrutura da tese.	28
Figura 2.1: Técnicas aplicadas na modelagem da demanda por transportes.	34
Figura 2.2: Representação hipotética de diferentes níveis de agregação de um município.	38
Figura 2.3: Representação hipotética do MAUP.	39
Figura 3.1: Esquema básico de funcionamento do AG.	49
Figura 3.2: Exemplo de representação binária no AG.	50
Figura 3.3: Representação do <i>crossover</i>	51
Figura 3.4: Representação do processo de mutação.	51
Figura 4.1: Comportamento de uma VR.	57
Figura 4.2: Representação anisotropia geométrica. Fonte: Soares, 2000.	58
Figura 4.3: Estrutura do semivariograma	59
Figura 5.1: Mapa de localização do município de São Paulo.	67
Figura 5.2: Uso e ocupação do solo do município de São Paulo	68
Figura 5.3: Uso e ocupação do solo da área de estudo	69
Figura 5.4: Dados da Pesquisa Origem/Destino 2007	70
Figura 5.5: Divisão modal da RMSP	71
Figura 5.6: Localização dos domicílios e distribuição espacial das variáveis estudadas.	72
Figura 5.7: Divisão modal da área de estudo.	73
Figura 5.8: Percentual de viagens por automóvel (sem valores ‘zero’).	74
Figura 5.9: Distribuição de frequência da variável Percentual de viagens por automóvel (domicílio).	74
Figura 5.10: Distribuição de frequência da variável <i>Renda familiar</i> (SM) por domicílio.	75
Figura 5.11: Distribuição de probabilidade acumulada para os suportes pontual x grade.	77
Figura 5.12: Distribuição de frequência das variáveis para as grades selecionadas.	78
Figura 5.13: Fluxograma das etapas do método.	80
Figura 6.1: Ilustração do método proposto.	82
Figura 6.2: Representação hipotética de um cromossomo.	83
Figura 6.3: Esquema do cálculo da largura de banda.	84
Figura 6.4: Esquema ilustrativo para identificar se um ponto pertence à janela de busca.	85
Figura 6.5: Ilustração da janela de busca para amostras irregulares	86
Figura 6.7: Gráfico de evolução do AG – Fitness x Geração – Percentual de viagens por automóvel.	89

Figura 6.8: Gráfico de evolução do AG – <i>Fitness</i> x Geração – <i>Renda (SM)</i>	89
Figura 7.1: Distribuição de frequência dos parâmetros do AG para a variável <i>Percentual de viagens por automóvel</i> – domicílio.	92
Figura 7.2: Distribuição de frequência dos parâmetros do AG para a variável <i>Renda familiar</i> – domicílio.	93
Figura 7.3: Distribuição de frequência dos parâmetros do AG para a variável <i>Percentual de viagens por automóvel</i> – grade regular.	94
Figura 7.4: Distribuição de frequências dos parâmetros do AG para a variável <i>Renda familiar</i> – grade regular.	95
Figura 7.5: Semivariogramas obtidos por AG. a) <i>Percentual de viagens por automóvel</i> no domicílio; b) <i>Percentual de viagens por automóvel</i> – grade regular; c) <i>Renda familiar</i> ; d) <i>Renda familiar</i> – grade regular.	97
Figura 7.6: Renda familiar média por zona de tráfego.	98
Figura 7.7: Mapas estimados através dos modelos otimizados.	99
Figura 7.8: Esquema de um mapa de semivariância	100
Figura 7.9: Mapas de semivariância da variável <i>Percentual de viagens por automóvel</i>	101
Figura 7.10: Mapas de semivariância da variável <i>Renda familiar (SM)</i>	102
Figura 7.11: Semivariogramas obtidos por ajuste manual. a) <i>Percentual de viagens por automóvel no domicílio</i> ; b) <i>Percentual de viagens por automóvel por grade regular</i> ; c) <i>Renda familiar</i> ; d) <i>Renda familiar por grade regular</i>	103
Figura 7.12: Mapas estimados através ajuste variográfico manual.	104
Figura 7.13: Semivariogramas referentes a anisotropia do <i>varmap</i> obtidos com a otimização	105
Figura 8.1: Síntese da tese.	108

LISTA DE TABELAS

Tabela 5.1: Estatísticas descritivas das variáveis por domicílio.....	73
Tabela 5.2: Estatísticas descritivas das variáveis por grades.....	77
Tabela 7.1: Intervalos dos parâmetros para cálculo e ajuste de semivariogramas (soluções quase ótimas e ótimas).....	96
Tabela 7.2: Parâmetros dos indivíduos – soluções ótimas do AG	96
Tabela 7.3: Medidas de desempenho dos modelos geoestatísticos.	106

LISTA DE QUADROS

Quadro 2.1: Problemas detectados e soluções sugeridas.....	45
Quadro 3.1: Particularidades do AG e SA	53
Quadro 4.1: Modelos teóricos de semivariogramas	60
Quadro 5.1: Resultados do teste não paramétrico de Mann – Whitney	76
Quadro 8.1: Resumo dos entraves relacionados à tese	110

SUMÁRIO

1. INTRODUÇÃO	23
1.1. CONSIDERAÇÕES INICIAIS	23
1.2. OBJETIVOS E CONTRIBUIÇÕES	26
1.2.1. Objetivo geral	26
1.2.2. Objetivos específicos	26
1.2.3. Contribuições da pesquisa	26
1.3. HIPÓTESES DE PESQUISA	27
1.4. ESTRUTURA DO TRABALHO	27
 2. ANÁLISE ESPACIAL E DEMANDA POR TRANSPORTES.....	29
2.1. ESTUDO DA DEMANDA POR TRANSPORTES: CONSIDERAÇÕES INICIAIS	29
2.1.1. Modelos de demanda por transportes: análise e aplicações	30
2.2. O PROBLEMA DA ESCALA EM ANÁLISE ESPACIAL	36
2.2.1. O Problema da Unidade de Área Modificável – <i>MAUP</i>	37
2.3. ESTUDO DA DEMANDA POR TRANSPORTES E A GEOESTATÍSTICA	41
2.3.1. Otimização para escolha de parâmetros da geoestatística	43
 3. MÉTODOS DE OTIMIZAÇÃO	47
3.1. ALGUMAS CONSIDERAÇÕES SOBRE TÉCNICAS DE OTIMIZAÇÃO	47
3.2. ALGORITMOS EVOLUTIVOS	48
3.2.1. Algoritmos Genéticos (AG)	48
3.3. <i>SIMULATED ANNEALING (SA)</i>	52
3.4. PARTICULARIDADES DOS DOIS MÉTODOS	53
 4. GEOESTATÍSTICA	55
4.1. CONCEITOS BÁSICOS DA GEOESTATÍSTICA	55
4.1.1. Variáveis Regionalizadas (VR)	56
4.1.2. Comportamento de uma VR	57
4.1.3. Semivariograma	58
4.1.4. Validação Cruzada	61
4.2. ESTIMATIVAS POR GEOESTATÍSTICA	62
4.2.1. Krigagem Ordinária (KO)	62
4.3. MUDANÇA DE SUPORTE E GEOESTATÍSTICA	64

5. MATERIAIS E MÉTODO: DADOS E ÁREA DE ESTUDO	67
5.1. APRESENTAÇÃO DA ÁREA DE ESTUDO	67
5.2. PESQUISA ORIGEM-DESTINO 2007	69
5.2.1. Seleção da área de estudo e apresentação dos dados	71
5.2.2. Análise das variáveis estudadas	73
5.3. PROCEDIMENTO PARA ESCOLHA DO TAMANHO DA GRADE	75
5.4. FERRAMENTAS UTILIZADAS	78
5.5. CONSIDERAÇÕES METODOLÓGICAS	79
6. APLICAÇÃO DO AG EM GEOESTATÍSTICA	81
6.1. CONSIDERAÇÕES INICIAIS	81
6.2. DETERMINAÇÃO DOS PARAMETROS DO SEMIVARIOGRAMA	82
6.2.1. Estrutura dos indivíduos	83
6.2.2. Etapa de otimização do ajuste dos semivariogramas	86
6.3. ETAPA DE OTIMIZAÇÃO POR AG	87
7. RESULTADOS DA ETAPA DE OTIMIZAÇÃO	91
7.1. OBTENÇÃO DE MODELOS VARIOGRÁFICOS ATRAVÉS DE AG	91
7.1.1. Distribuição de parâmetros de semivariogramas	91
7.1.2. Geração de semivariogramas experimental e teórico	97
7.2. ESTIMAÇÃO DE DADOS DE VIAGENS POR KRIGAGEM ORDINÁRIA	98
7.3. VALIDAÇÃO DA OTIMIZAÇÃO ATRAVÉS DE PROCEDIMENTOS TRADICIONAIS	100
7.4. MEDIDAS DE DESEMPENHO DOS MODELOS – VALIDAÇÃO CRUZADA	105
8. CONSIDERAÇÕES FINAIS	107
8.1. SÍNTESE DO ESTUDO	107
8.2. CONCLUSÕES E RECOMENDAÇÕES	109
8.2.1. Conclusões e contribuições	109
8.2.2. Recomendações para trabalhos futuros	111
REFERÊNCIAS	113
APÊNDICE I	121
APÊNDICE II	122

APÊNDICE III.....	134
APÊNDICE IV	137
APÊNDICE V	138
APÊNDICE VI.....	139

Capítulo

INTRODUÇÃO

1

Nesta introdução serão apresentados os objetivos, hipótese de pesquisa e a estrutura deste texto, além de uma sucinta discussão teórica sobre as ideias que nortearam este trabalho.

1.1. CONSIDERAÇÕES INICIAIS

A análise de demanda por transportes, comumente, é realizada através de métodos estatísticos tradicionais. Muitas das técnicas, tais como, a regressão linear tradicional, modelos lineares generalizados ou modelos sintéticos, buscam modelar e estimar variáveis de demanda por transportes (MELLO, 1975; BRUTON, 1979). Usualmente, estes modelos representam médias de atributos ou total de viagens geradas, representadas por um setor censitário, uma zona de tráfego, ou um domicílio, por exemplo. No entanto, entende-se que a análise de demanda por transportes relaciona-se também com a localização espacial dos indivíduos e/ou domicílios e dos destinos das viagens. Assim, a utilização de variáveis que consideram a localização geográfica passa a ser uma importante ferramenta para o aprimoramento das estimativas de dados relacionados à demanda por transportes (PITOMBO, COSTA e SALGUEIRO, 2015).

Com a popularização dos Sistemas de Informações Geográficas (SIG), houve o aumento da utilização de técnicas estatísticas que consideram a dependência espacial dos dados. Análises urbanas de uso do solo e transportes vêm utilizando, ao longo dos anos, ferramentas que consideram a localização das informações, como indicadores de associação espacial, modelos espaciais locais e globais e/ou técnicas de interpolação espacial (PÁEZ e SCOTT, 2005). Bhat e Zhao (2002) demonstraram que a análise de demanda por transportes é afetada por questões espaciais e estas precisam ser incorporadas no contexto de pesquisas de análise de geração de viagens. Adjemian, Cynthia Lin e Williams (2010) discutiram a existência de efeitos espaciais na variável posse de automóveis. Os resultados deste estudo apontaram para a importância de considerar as localizações espaciais em pesquisas relacionadas à engenharia ou planejamento de sistemas de transportes.

Dentre as técnicas que consideram a dependência espacial de variáveis, destaca-se a geoestatística. A geoestatística é uma técnica de análise que possibilita o desenvolvimento de

estudos onde os valores das variáveis não são independentes no espaço. A aplicação da geoestatística permite, principalmente, prever valores de uma variável em pontos não amostrados, além de ser possível avaliar a acurácia das estimativas. A capacidade de prever valores de uma variável em locais sem informações, através de pontos conhecidos, é uma vantagem da geoestatística em relação a outros métodos de análise espacial (ISAACS e SRIVASTAVA, 1989).

A utilização da geoestatística na área de planejamento de transportes é relativamente recente. Eom, Park, Heo e Huntsinger (2006) utilizaram técnicas de análise de dependência espacial (dentre elas, a krigagem) para prever o tráfego médio diário anual. Estes autores concluíram que o volume de tráfego em uma estação de monitoramento está correlacionado com os volumes nas estações vizinhas. Além disso, os modelos que levaram em conta a autocorrelação espacial apresentaram melhor capacidade preditiva em relação aos modelos tradicionais não espaciais. Miura (2010) e Yong, Zhuang e Haod (2012) utilizaram a geoestatística para prever tempo de viagem e a variação espacial na velocidade veicular, respectivamente. Selby e Kockelman (2013) compararam os resultados de estimativas por krigagem para contagem de tráfego considerando a distância euclidiana e de rede. No entanto, esta última não mostrou melhor desempenho em relação à primeira.

Na área de demanda por transportes, Zhang e Wang (2014) estimaram o número de passageiros em uma linha de metrô de Nova York. As distâncias euclidianas foram substituídas pelas distâncias em rede, considerando que atividades de transportes, geralmente, ocorrem em uma rede estruturada. A krigagem em rede apresentou melhores ajustes em relação aos resultados da regressão linear e de krigagem realizada através da distância euclidiana convencional.

O grupo de pesquisa do Departamento de Engenharia de Transportes da Escola de Engenharia de São Carlos/Universidade de São Paulo, onde se insere este trabalho, vem buscando incorporar análise espacial a estudos de demanda por transportes, utilizando a geoestatística (GOMES, PITOMBO, ROCHA e SALGUEIRO, 2016; LINDNER, PITOMBO, ROCHA e QUINTANILHA, 2016; LINDNER e PITOMBO, 2018; PITOMBO, SALGUEIRO, COSTA e ISLER, 2015; ROCHA, LINDNER e PITOMBO, 2017).

No entanto, é fundamental considerar que dados relacionados à demanda por transportes possuem algumas particularidades quando analisados através da geoestatística. As unidades de área que, geralmente, representam dados relacionados à demanda por transportes, como zonas de tráfego e setores censitários, por exemplo, possuem formas e dimensões irregulares

(GOOVAERTS, 2008). Este é mais um obstáculo da aplicação da geoestatística em estudos de transportes, visto que esta técnica foi desenvolvida para ser aplicada a suportes (unidades de análise) regulares (com mesma forma e tamanho).

Além disso, o ideal em análises de demanda por transportes é conhecer o comportamento individual. Por outro lado, dados desagregados relativos a indivíduos, quando disponibilizados, muitas vezes, apresentam alta variabilidade a curtas distâncias (LINDNER, 2015), o que dificulta a caracterização do comportamento espacial deste tipo de variável. Portanto, a escala de amostragem é um dos fatores que influenciam na qualidade dos modelos. A escolha da escala de representação espacial dos dados de uma variável é imprescindível para a modelagem geoestatística de variáveis relacionadas à demanda por transportes, pois há duas importantes questões a serem consideradas: 1) representar adequadamente o comportamento individual e, 2) prezar pela qualidade dos modelos espaciais.

Na geoestatística, além do processo de escolha da escala de obtenção das amostras, há uma etapa não menos importante: a caracterização da estrutura espacial de uma Variável Regionalizada (VR). Para demonstrar o comportamento espacial de uma variável, utiliza-se o semivariograma, um gráfico que explica e ilustra a variabilidade de um atributo em função da distância. A obtenção do semivariograma experimental demanda tempo e amplo conhecimento sobre o comportamento estrutural de uma VR, pois esta etapa, muitas vezes, é realizada visualmente, de acordo com a familiaridade e experiência do pesquisador. Além disso, não há um método totalmente adequado para ajuste de semivariogramas experimentais. Os critérios automáticos mais populares são baseados na função de mínimos quadrados ponderados, assumindo suposição de distribuição gaussiana (CRESSIE, 1985). O objetivo principal deste método de ajuste é diminuir a distância entre a curva teórica e o semivariograma experimental, tendo como ponderador as semivariâncias dos *lags* menores, como também, o número de pares de pontos considerados para o cálculo. Desta maneira, técnicas de otimização podem ser uma alternativa viável para cálculo e ajuste de um semivariograma experimental, pois elas geram milhares de respostas para os parâmetros a serem estimados.

Um dos métodos que utilizam a busca exaustiva de soluções para problemas são os Algoritmos Genéticos (AGs). Os AGs são um campo de estudo dos Algoritmos Evolucionários e, como tal, são baseados na Teoria de Seleção Natural, proposta no século XIX por Charles Darwin (BANZHAF, NORDIN, KELLER e FRANCONI, 1998).

Diante dos problemas de pesquisa apresentados, o presente texto discute os desafios para cálculo e ajuste de semivariogramas experimentais para dados relacionados à demanda por transportes. A representação adequada destes dados também se apresenta como um problema metodológico em análise espacial, visto que a escala utilizada pode interferir nos resultados dos modelos.

1.2. OBJETIVOS E CONTRIBUIÇÕES

1.2.1. Objetivo geral

Verificar a adequabilidade do uso de Algoritmos Genéticos para a otimização de modelagem geoestatística aplicada à demanda por transportes.

1.2.2. Objetivos específicos

- a. Propor um procedimento que auxilie na definição da escala de análise para modelagem de dados relativos a viagens;
- b. Otimizar cálculo e ajuste de semivariogramas experimentais;
- c. Obter intervalos e distribuição de parâmetros de semivariogramas experimentais e teóricos, correspondentes a soluções quase ótimas;
- d. Disponibilizar código aberto para uso na área de demanda por transportes, como também em demais campos de conhecimento que utilizam dados espacialmente dependentes.

1.2.3. Contribuições da pesquisa

Considerando os objetivos anteriormente apresentados, esta tese apresenta duas importantes contribuições em estudos de análise espacial de demanda por transportes: a) propor um método de otimização para definição de parâmetros do semivariograma, com base nos erros das estimativas da krigagem e, b) buscar por um suporte homogêneo, para representar dados desagregados, que diminua a aleatoriedade espacial associada aos domicílios e/ou irregularidade, associada às zonas de tráfego.

1.3. HIPÓTESES DE PESQUISA

Este trabalho foi norteado a partir de duas hipóteses principais de pesquisa, descritas em seguida.

- O uso de Algoritmos Genéticos otimiza a modelagem geoestatística aplicada a variáveis de demanda por transportes;
- O conhecimento de distribuição e intervalo de parâmetros do variograma teórico, a partir de soluções quase ótimas, oriundas do AG, pode colaborar para o entendimento da estrutura espacial da Variável Regionalizada.

1.4. ESTRUTURA DO TRABALHO

Este texto está estruturado em sete capítulos, além desta introdução. A Figura 1.1 resume as etapas deste trabalho, distribuídas através dos capítulos que compõem o presente texto.

O Capítulo 2 discorre a respeito dos principais aspectos relacionados aos modelos mais utilizados para a previsão de dados de demanda por transportes. Este capítulo busca sistematizar o avanço de técnicas estatísticas, fazendo uma revisão teórica sobre modelos não-espaciais e aqueles que consideram o fator espacial na área de demanda por transportes. Além disso, discute-se a relação entre a escala e a representação de dados em análise espacial.

O Capítulo 3 apresenta, sucintamente, alguns princípios fundamentais dos Algoritmos Evolutivos e introduz conceitos básicos sobre um dos seus ramos, os Algoritmos Genéticos, método de otimização utilizado neste trabalho.

O Capítulo 4 traz uma breve descrição e conceitos acerca da técnica de análise espacial utilizada neste estudo – a geoestatística.

No Capítulo 5 é feita uma caracterização da área de estudo e da pesquisa Origem-Destino de 2007 da Região Metropolitana de São Paulo. Além disso, as variáveis estudadas neste trabalho, bem como suas características estatísticas serão discutidas. Neste capítulo é apresentado um código em linguagem R para otimização da escolha das dimensões da grade regular (etapa relativa à escolha do suporte). Finalmente, neste capítulo, apresenta-se o ferramental utilizado.

Figura 1.1: Estrutura da tese.

O Capítulo 6 discutir a etapa de otimização, através de Algoritmos Genéticos, para busca de parâmetros que melhor representam fenômenos espaciais de demanda por transportes.

O sétimo capítulo apresenta os principais resultados e discussões. Neste capítulo é apresentada a análise geoestatística (comportamento estrutural e aplicação da krigagem) de dados relacionados a viagens. Os resultados da etapa de estimação geoestatística serão analisados, bem como avaliada a relação entre mudança de escala em análise espacial.

O Capítulo 8 busca discutir as principais conclusões e pontuar os problemas encontrados ao longo do desenvolvimento deste trabalho. Por fim, com base em questões que este trabalho não conseguiu responder, são apresentadas algumas recomendações para trabalhos futuros.

Capítulo

2

ANÁLISE ESPACIAL E DEMANDA POR TRANSPORTES

Este capítulo apresenta uma breve revisão de literatura acerca de técnicas estatísticas aplicadas a problemas de transportes, com enfoque especial em modelos de previsão de demanda por transportes. Os problemas de mudança de escala e da unidade de área modificável serão abordados. Além disso, procura-se discutir como este trabalho pode preencher algumas lacunas encontradas na literatura da área de estudo, através da busca de soluções para adequação do uso de técnicas de análise espacial à previsão de dados relacionados à demanda por transportes. Por fim, será apresentada uma breve discussão sobre técnicas de otimização aplicadas à análise geoestatística.

2.1. ESTUDO DA DEMANDA POR TRANSPORTES: CONSIDERAÇÕES INICIAIS

O crescimento econômico acelerado aumentou a necessidade de locomoção de pessoas e mercadorias diariamente pelas cidades, o que vem a ocasionar diversos problemas, como congestionamentos, poluição, riscos de acidentes, dentre outros. A Engenharia de Transportes busca atenuar estes conflitos através do estudo do planejamento de transportes (ORTUZAR e WILLUMSEN, 2011). Estudos relacionados à demanda por transportes englobam a aplicação de técnicas voltadas para a adequação da oferta à demanda por transportes, minimizando problemas na mobilidade, principalmente nas grandes cidades.

Em planejamento de transportes, comumente, utiliza-se métodos estatísticos de modelagem tradicionais, como a regressão linear, aplicada a modelos sequenciais, por exemplo (MELLO, 1975; BRUTON, 1979). Modelos tradicionais, utilizados na previsão de demanda por transportes, buscam representar valores médios de atributos agregados por uma unidade de área que, geralmente, é representada por um setor censitário ou uma zona de tráfego. Dentre os problemas relacionados aos modelos tradicionais estão os possíveis resultados negativos e a suposição de comportamento contínuo das variáveis de viagens, além da não incorporação de características individuais sobre comportamento dos viajantes (CHANG, JUNG, KIM e KANG, 2014). Na análise agregada, os dados sobre demanda por transportes (por exemplo, quantidade

de viagens por modo) são vistos como resultantes de um agrupamento de decisões individuais (PÁEZ e SCOTT, 2005). No entanto, características comportamentais representadas através de áreas podem sintetizar bastante o comportamento individual nos modelos de previsão, generalizando informações importantes para o planejamento de transportes, como, por exemplo, os fatores que influenciam as decisões dos viajantes. O comportamento habitual do indivíduo pode variar de acordo com diversos fatores que influenciam a escolha do modo ou destino, como o tempo e/ou o custo da viagem (ORTÚZAR e WILLUMSEN, 2011). Por isso, o comportamento individual não pode ser desconsiderado nas análises. Por outro lado, ele é bastante complexo, uma vez que envolve decisões a partir de percepções pessoais.

A definição da forma de representação de dados de transportes é uma etapa muito importante na modelagem da demanda, pois o nível de detalhe sob o qual o dado é representado pode interferir nas respostas dos modelos. Além disso, muitas vezes, as variáveis representadas por áreas são tratadas como contínuas, o que pode ocasionar problemas de representação espacial, como será visto mais adiante neste capítulo.

A subseção a seguir procura apresentar uma discussão sobre os diferentes tipos de modelos utilizados no estudo da demanda por transportes, além de procurar enfatizar possíveis lacunas encontradas neste campo de pesquisa. Busca-se, ainda, apresentar uma proposta de método baseado nos problemas identificados, que será discutida nos Capítulos 5 e 6 deste texto.

2.1.1. Modelos de demanda por transportes: análise e aplicações

Na análise de modelos de demanda por transportes, tradicionalmente, são utilizados modelos estatísticos não espaciais. A Regressão Linear Múltipla (RLM), embora amplamente utilizada na área de modelagem de demanda por transportes, apresenta diversas restrições, especialmente pela natureza dos dados de viagens: viagens são dados de contagem e não atendem à suposição de normalidade, por exemplo. Além disso, a relação existente entre a geração de viagens e variáveis explicativas, como população, por exemplo, muitas vezes não é linear.

Em geral, os resíduos provenientes de modelos tradicionais de demanda por transportes não são distribuídos aleatoriamente com média em torno de zero e não atendem à suposição de homoscedasticidade (variância constante dos erros). Como alternativa aos modelos de regressão linear simples e múltipla, Nelder e Wedderburn (1972) propuseram os Modelos Lineares Generalizados (MLGs), onde as observações, supostamente, veem de uma função de densidade de probabilidade da família exponencial. Os MLGs referem-se a uma classe maior

de modelos e podem ser relacionados a quatro tipos de distribuições de dados: Normal, Binomial, Poisson e Gama. Os MLGs fornecem base estatística para combinar componentes aleatórios a sistemáticos do modelo, através de uma função de ligação. O componente aleatório seria a variável dependente e sua função de distribuição de probabilidade, já o sistemático seria o preditor linear do modelo, que pode ser linearizado por uma transformação adequada. Modelos lineares generalizados são amplamente utilizados em diversos trabalhos na área de transportes (JOSHUA e GARBER, 1990; MIAOU e LUM, 1993; MIAOU, 1994; MAHER e SUMMERSGILL, 1996; THOMPSON, 1997; CHU, 2004; CHOI, LEE, KIM e SOHN, 2012)

Outra restrição quanto ao uso de modelos tradicionais em análise da demanda por transportes é o uso aplicado, tradicionalmente, ao estudo de zonas de tráfego, considerando-se fatores agregados socioeconômicos, de uso do solo ou acessibilidade. No entanto, no planejamento de transportes é importante obter informações relacionadas ao indivíduo, pois valores agregados médios, na maioria das vezes, podem estar vinculados a problemas como a *falácia ecológica*, apresentada na próxima seção deste capítulo.

Por outro lado, visto que a localização das atividades e uso do solo influenciam nas decisões individuais relacionadas à demanda por transportes, é importante pontuar que modelos tradicionais não consideram a dependência espacial dos dados, ou seja, a componente espacial não é considerada nos modelos de previsão. Uma vez que a demanda por transportes surge a partir da necessidade de realização de atividades cotidianas como ir ao trabalho, à escola ou fazer compras, é necessário entender como estas atividades estão distribuídas no espaço, ou seja, é importante conhecer o uso do solo urbano (MCFADDEN, 1974). Neste contexto, a estrutura espacial e a forma urbana das cidades estão diretamente relacionadas às distâncias e ao tempo das viagens. Usos mistos ou segregados do espaço urbano, como por exemplo, a densidade de atividades ou o espraiamento urbano, podem influenciar no quantitativo de viagens geradas em determinado local ou na escolha modal dos indivíduos (ORTÚZAR e WILLUMSEN, 2011; TSEKERIS e TSEKERIS, 2011). O atributo espacial, ou seja, a localização de um fenômeno, exerce importante influência, por exemplo, na origem, destinos e escolha do modo de transporte utilizado nas viagens. Com isso, neste capítulo são consideradas, sobretudo, a influência da dependência espacial na análise e previsão da demanda por transportes.

De acordo com a Primeira Lei da Geografia, observações no espaço geográfico são dependentes, ou seja, dados que estão espacialmente próximos tendem a ter valores mais parecidos que aqueles mais distantes (TOBLER, 1970). Esta é uma propriedade do espaço

geográfico que pode ser explorada para o entendimento de fenômenos levando em consideração a componente espacial. A aplicação de ferramentas e modelos que avaliam as relações dos valores de uma variável no espaço é conhecida, em vários campos de estudo, como **análise espacial de dados**. Esse conjunto de técnicas procura relacionar e quantificar características de um fenômeno (que varia significativamente no espaço) a partir de sua localização geográfica (FISCHER, 2001; DRUCK, CARVALHO, CÂMARA e MONTEIRO, 2004).

Além das ciências ambientais, muitas áreas de estudo utilizam ferramentas de análise espacial, como a criminologia, epidemiologia e planejamento urbano (GOOVAERTS, 2005; 2011; TSUTSUMI e SEYA, 2008; SOUSA, PITOMBO, ROCHA, SALGUEIRO e DELGADO, 2017). Principalmente após os avanços dos Sistemas de Informações Geográficas (SIG) e com a difusão de dados espaciais digitais, técnicas que incorporam a dependência espacial aos estudos relativos a problemas de transportes passaram a ser amplamente utilizadas.

A análise espacial possui duas diferentes abordagens (exploratória e confirmatória), que compreendem desde uma simples descrição estatística até a inferência de dados através de modelos. Mas, na prática, muitas técnicas de análise espacial utilizam os dois conceitos (ANSELIN e GETIS, 1992).

A análise exploratória, como o próprio nome já explicita, é utilizada quando se quer conhecer as características e padrões espaciais de uma variável. Muitos métodos são utilizados nesta etapa, desde a visualização de informações no espaço, até os indicadores de dependência ou autocorrelação espacial local e global, como o Indicador Local de Associação Espacial (*Local Indicators of Spatial Association – LISA*) e MORAN, respectivamente (ANSELIN, 1995; DRUCK, CARVALHO, CÂMARA e MONTEIRO, 2004).

O semivariograma, principal etapa da modelagem geoestatística, também mensura a dependência espacial e corresponde a uma ferramenta de análise exploratória de dados espaciais (DRUCK, CARVALHO, CÂMARA e MONTEIRO, 2004). Já a análise confirmatória engloba um conjunto de ferramentas para modelagem, previsão e validação de dados relativos a fenômenos espaciais. Dentre estas ferramentas de análise confirmatória, destaca-se a krigagem, método geoestatístico de estimação para superfícies, utilizado nesta tese.

Na Engenharia de Transportes, a aplicação de técnicas de análise espacial é relativamente recente. Pesquisas que envolvem esta área de estudo tiveram início por volta dos anos 2000. A partir deste período, muitos trabalhos discutiram como a incorporação da localização dos dados passou a exercer importante influência na qualidade dos modelos. Em análise exploratória de

dados espaciais, utilizada em estudos de Engenharia de Transportes, destacam-se trabalhos na área de segurança viária (NICHOLSON, 1998; SHAFABAKHSH, FAMILI e BAHADORI, 2017) e demanda por transportes (ALVES, DA SILVA e VAN DER WAERDEN, 2012; BHAT e ZHAO, 2002; DIMATULAC e MAOH, 2017; LOPES, BRONDINO e DA SILVA, 2014; PITOMBO, KAWAMOTO e SOUSA, 2011).

Técnicas de análise espacial confirmatória, como a Regressão Geograficamente Ponderada (RGP) foram utilizadas em diversos trabalhos (BLAINEY, 2010; BLAINEY e MULLEY, 2013; CLARK, 2007; HADAYEGHI, SHALABY e PERSAUD, 2003).

Zhao e Park (2004) utilizaram a RGP para analisar e estimar o volume de tráfego médio diário anual. Os modelos de RGP permitiram estimar parâmetros locais, ao invés de globais, como ocorre na análise de regressão linear tradicional. Os autores compararam os resultados destas duas técnicas e concluíram que os modelos de RGP, por considerarem o fator espacial, foram mais precisos do que os modelos de regressão tradicionais.

Em análise da demanda por transportes, Bhat e Zhao (2002) identificaram que questões espaciais precisam ser incorporadas nas pesquisas relacionadas à demanda por transportes. O modelo *logit* misto identificou a existência de fatores espaciais na análise de geração de viagens baseadas em atividades (compras) com múltiplas paradas. Tal modelo identificou que a acessibilidade é uma componente do espaço urbano que influencia a quantidade de viagens encadeadas, além de características familiares, como emprego e renda. Pitombo, Sousa e Birkin (2010) utilizaram a RGP para análise e previsão de variáveis relacionadas à demanda por transportes. Neste trabalho, verificou-se que há dependência espacial de variáveis explicativas (socioeconômicas e densidade de atividades). A aplicação da RGP apresentou bons resultados em relação à visualização dos padrões de deslocamentos espaciais, bem como estimativas de parâmetros locais, associados às variáveis explicativas.

Destaca-se que os trabalhos discutidos nesta subseção utilizaram análise espacial em diversas áreas de planejamento de transportes, através de técnicas que não preveem valores em pontos não amostrados.

A partir da breve discussão sobre modelos de demanda por transportes, sobretudo os modelos espaciais, a Figura 2.1 apresenta algumas características das técnicas mais frequentemente utilizadas no estudo de demanda por transportes. Além disso, a técnica de análise espacial utilizada nesta pesquisa, bem como algumas limitações encontradas a partir de seu uso, são apresentadas junto a uma sugestão de método que busca propor soluções para tais problemas.

Figura 2.1: Técnicas aplicadas na modelagem da demanda por transportes.

A principal vantagem da geoestatística, em relação a outros métodos de análise espacial, é estimar informações sobre a variável de interesse para uma superfície. Além disso, não é necessário utilizar informações de outras variáveis, além da amostra daquela que se quer conhecer. Esta característica da técnica é possível devido ao formalismo geoestatístico que permite, a partir das características estruturais de uma variável, a obtenção de um modelo matemático para prever dados em locais sem informações prévias.

Por outro lado, dados relacionados à demanda por transportes são, muitas vezes, disponibilizados de forma agregada e zonas de tráfego, por exemplo, possuem forma e tamanho irregulares. Este é um obstáculo encontrado para a aplicação da geoestatística a estudo de dados de transportes. Além disso, é fundamental conhecer o comportamento desagregado das viagens.

Outra importante consideração sobre a aplicação da geoestatística a dados de demanda por transportes é que tais dados precisam de algumas adaptações. Dados de viagens são derivados de características de comportamento de indivíduos, os quais não se apresentam de maneira contínua no espaço, como dados geológicos, onde a geoestatística se desenvolveu. Esta característica tem influência direta na qualidade dos modelos, pois a alta variabilidade a curtas distâncias, muitas vezes apresentada pelos dados, acarreta problemas de representação estrutural do semivariograma, como a alta taxa de aleatoriedade, representada no gráfico pelo efeito pepita. Por isso, neste trabalho propõe-se um método de agregação de dados, que minimiza dois importantes problemas encontrados na aplicação de geoestatística na modelagem de demanda por transportes: 1) quando agregados, os dados são obtidos por zoneamento irregular e, 2) a existência de alta variabilidade dos dados desagregados (quando disponibilizados).

O ajuste teórico do semivariograma experimental é um procedimento que, muitas vezes, é realizado visualmente. Este método, embora amplamente utilizado em diversas áreas de pesquisa que utilizam a geoestatística como ferramenta de análise espacial, muitas vezes pode ser trabalhoso e, sobretudo, demorado. Além disso, tal procedimento não é confiável por inúmeras razões, incluindo possíveis flutuações no semivariograma experimental e a precisão do cálculo das semivariâncias e anisotropias, ou seja, variação na função subjacente com mudança de direção (OLIVER e WEBSTER, 2014). Técnicas de otimização podem ser uma alternativa a estes obstáculos.

No entanto, o cálculo e ajuste automático do semivariograma experimental ainda é um desafio na geoestatística. Entre os poucos trabalhos encontrados sobre o tema, a maioria trata apenas

do ajuste do modelo teórico (CUI, YONEDA, SHIMADA e MATSUI, 2016; JIAO e LEI, 2011; LI e LU, 2010). Não foram encontradas pesquisas que abordem procedimentos para definição dos parâmetros (direção ou ângulo de anisotropia, tolerância angular e tamanho de *lag*) para o cálculo do semivariograma experimental.

Por outro lado, principalmente usuários da área de ciências ambientais e geologia, mesmo diante dos avanços em pesquisas que envolvem princípios de programação genética e otimização, defendem que a modelagem automática de semivariogramas pode não ser a melhor solução em geoestatística. Grande parte dos estudiosos da área concordam que, muitas vezes, o conhecimento prévio do pesquisador sobre o comportamento da variável pode ser uma etapa importante na modelagem da dependência espacial, o que não é possível incorporar aos métodos de otimização. No entanto, entende-se que métodos de inteligência computacional podem contribuir para a escolha dos parâmetros para o cálculo e ajuste do semivariograma e, ser uma alternativa viável às técnicas de modelagem tradicionais, já citadas.

Atualmente, técnicas de programação genética vêm sendo amplamente utilizadas em diversas áreas de pesquisa que buscam a otimização para solução de problemas. Além de estudos diretamente relacionados à computação, às engenharias (CUNHA, BESSA e SETTI, 2009; EGAMI, MON-MA, SETTI e RILETT, 2006), ciências ambientais, como em geoprocessamento e classificação de imagens (BANDYOPADHYAY e MAULIK, 2002; LI, WU, WAN e ZHU, 2011; WANG, 1991), além da própria geoestatística (ABEDINI, NASSERI e BURN, 2012; MASCHIO, VIDAL e SCHIOZER, 2008) passaram a utilizar ferramentas de inteligência computacional como técnicas de otimização para solução de problemas de pesquisa.

2.2. O PROBLEMA DA ESCALA EM ANÁLISE ESPACIAL

A escala possui diversos conceitos na geografia, desde o cartográfico até o epistemológico. Operacionalmente, o conceito de escala não é apenas utilizado na Geografia, mas em vários campos científicos, devido à sua influência no processo de experimentação, sendo uma medida que dá visibilidade ao fenômeno (DE CASTRO, CORRÊA, e GOMES, 1995). A escala, portanto, define o nível de análise, ou recorte do objeto de estudo.

Muitos problemas em análises ambientais e sociais exigem o conhecimento de processos que operam sob diferentes escalas. Portanto, há a necessidade de prever e controlar os efeitos da

mudança de escala e de diferentes níveis de agregação nos resultados estatístico dos estudos (MARCEAU, 1999).

Existem diversas formas de aquisição de dados utilizados em pesquisa científica. Muitas vezes, estes dados são obtidos pontualmente ou são associados a uma unidade de área (GELFAND, ZHU e CARLIN, 2001). Em estudo de dados agregados por área, um dos principais problemas são os critérios para definição dos polígonos, uma vez que diferentes agrupamentos podem gerar análises com diferentes resultados.

Por motivos como confidencialidade, na maioria das pesquisas sobre informações socioeconômicas da população, os dados coletados são vinculados a áreas. No entanto, o que se deseja conhecer, na maioria das pesquisas, são as características relacionadas a indivíduos. Esta questão pode levar a problemas de interpretação dos resultados das análises. Desta maneira, os possíveis efeitos da escala não podem ser ignorados, pois há riscos de causar sérios equívocos na interpretação dos resultados. Uma abordagem mais coerente na modelagem espacial seria manter a mesma escala e características nos agrupamentos de dados espaciais (OPENSHAW, 1976).

Como visto na seção anterior, verificou-se que a utilização da análise espacial passou a ganhar destaque, principalmente após a incorporação de análises realizadas através de SIG. No entanto, devem ser consideradas questões metodológicas na análise espacial em transportes (MILLER, 1999; WANG, ZHUANG e LIU, 2012). Páez e Scott (2005) discutem a importância de se considerar os efeitos de diferentes zonamentos nas análises.

Em análise espacial, dois fenômenos relacionados à escala vêm sendo bastante discutidos em diversas áreas de estudo. A existência do Problema da Unidade de Área Modificável (*Modifiable Areal Unit Problem – MAUP*) pode ocorrer devido ao nível de agregação dos dados, através da modificação da escala ou do nível de agregação de zonas. A próxima subseção apresenta algumas considerações teóricas sobre a ocorrência destes eventos.

2.2.1. O Problema da Unidade de Área Modificável – MAUP

Análises de mudança de escala, com base na mudança de tamanhos das unidades de área ou polígonos, já foram bastante discutidas na literatura através do conceito do Problema da Unidade de Área Modificável (*Modifiable Areal Unit Problem – MAUP*). O MAUP é decorrente de uma conhecida discussão da Geografia: o problema da escala. O efeito de escala está

relacionado à influência do nível de agregação de variáveis espaciais, onde os resultados, decorrentes de uma mesma técnica de análise, tendem a variar de acordo o nível de resolução espacial adotado (PÁEZ e SCOTT, 2005).

O Problema da Unidade de Área Modificável envolve duas diferentes questões. A relação com a localização dos dados e sua adequada representação que, depende do sistema de zoneamento adotado (FOTHERINGHAM, 1989). Mas, muitas vezes, a escolha de unidades de análise é feita sem critério e, a conveniência pode passar a ser mais importante do que a adequada representação do fenômeno geográfico no espaço. Esta falta de critério, sobre a natureza e definição dos níveis de agregação em análise espacial, tem como importante consequência o *MAUP* (OPENSHAW, 1984).

Em estudos urbanos, muitas vezes, os recortes espaciais não levam em conta fatores socioeconômicos da população. Desta forma, a mudança da escala de análise pode interferir nas características estatísticas de uma amostra e na variabilidade dos resultados. Em áreas com desigualdades socioeconômicas, como no caso de muitos municípios brasileiros, dados agregados em uma mesma área ou polígono podem estar representando diferentes grupos sociais. Estatisticamente, valores médios por áreas podem divergir essencialmente de dados desagregados ou individuais (HOLT, STEEL, TRANMER, e WRIGLEY, 1996). Como a análise por unidade de áreas, muitas vezes, pretende inferir dados pontuais com base em médias ou taxas de uma variável nos blocos ou zonas, o comportamento individual em certas áreas pode ser bem diferente da média encontrada em análises agregadas, podendo ocorrer erros grosseiros de interpretação.

Para exemplificar o *MAUP*, a Figura 2.2 (a) representa, hipoteticamente, o recorte de um município e as Figuras 2.2 (b) e (c) correspondem a uma agregação por bairros e por setores censitários, respectivamente.

Figura 2.2: Representação hipotética de diferentes níveis de agregação de um município.

Considerando os três diferentes recortes espaciais apresentados acima, supõe-se que um estudo sobre a variável renda familiar foi realizado neste município. As rendas mais altas a mais baixas são representadas por uma gradação da cor escura para mais clara (Figura 2.3). A representação da renda neste município varia de acordo com o nível de agregação dos dados, ou seja, o rendimento em cada área é representado por valores médios da renda das famílias incluídas nas diferentes escalas de representação.

Figura 2.3: Representação hipotética do *MAUP*.

Desta maneira, a interpretação de uma característica analisada através da visualização de um mapa coroplético depende das diferentes maneiras que o fenômeno pode ser representado através de seu agrupamento. Fotheringham e Wong (1991) alertam para a falta de confiabilidade de análises realizadas a partir de dados agregados por áreas. Para estes autores, uma solução mais óbvia seria, sempre que possível, evitar o uso de dados agregados. No entanto, isso não é uma solução viável em muitos tipos de análise, particularmente aquelas que usam dados do censo, onde a confidencialidade, em regra, deve ser mantida.

Diferentes níveis de agregação podem levar a incertezas na interpretação dos resultados, levando ao problema da falácia ecológica, conceito associado ao *MAUP*. A falácia ecológica ocorre quando as análises baseadas em dados agregados levam a conclusões diferentes das análises baseadas em dados individuais (GOTWAY e YOUNG, 2002). No entanto, é importante considerar que sua existência depende da natureza da agregação utilizada e do grau de homogeneidade dos agrupamentos (OPENSHAW, 1984).

Diversos sistemas de zoneamento podem ser utilizados para representar uma cidade em unidades de área menores e cada sistema de zoneamento representará, de diferentes maneiras, o mesmo atributo. Em análise espacial de estudos urbanos, grande parte dos zoneamentos são internamente heterogêneos e indivíduos com características distintas podem fazer parte da mesma zona. Uma consequência direta disso é a variação dos resultados em estudos que

utilizam a mesma técnica, mas diferentes sistemas de zoneamento na mesma área de estudo (PÁEZ e SCOTT, 2005).

Diferentes propostas para minimizar os efeitos do *MAUP* são encontradas na literatura. Uma solução proposta por Jelinski e Wu (1996) seria uma avaliação de diferentes estruturas espaciais (variando o zoneamento e a escala) como alternativa para realizar análises de sensibilidade. Para Páez e Scott (2005), uma maneira de mitigar os efeitos do *MAUP* é empregar as menores unidades de área disponíveis para verificar a precisão dos resultados sob diferentes escalas e verificar os consequentes efeitos de zoneamento. Caso seja demonstrado que os resultados não sofrem variação em diferentes escalas, pode-se considerar que os resultados são confiáveis e não simplesmente representações dependentes da resolução espacial adotada pelo pesquisador. Haynes, Daras, Reading e Jones (2007) abordam a necessidade de criar unidades de zoneamento com homogeneidade interna e manter uma quantidade considerável de heterogeneidade externa entre zonas.

Geralmente, pesquisas na área de saúde utilizam setores censitários como unidade de área em suas análises. Parenteau e Sawada (2011) criticaram o fato de que, em muitos estudos, nenhuma outra unidade de área é utilizada como alternativa para a verificação da existência do *MAUP*. Para entender como o nível de agregação interfere nos resultados, os autores analisaram três diferentes níveis de agregação de dados: os setores censitários, bairros e um zoneamento automático (agrupamentos de setores censitários, com escala equivalente aos bairros). A técnica de agrupamento automático delimitou unidades homogêneas a partir de fatores socioeconômicos da população como renda, taxa de desemprego e nível de escolaridade. Os autores concluíram que se o zoneamento leva em conta características da população, preservando a homogeneidade intrazonal, a modificação de áreas passa a não interferir nos resultados das análises, logo o *MAUP* e a *falácia ecológica* podem ser solucionados.

Goovaerts (2012) combinou três conjuntos de dados geográficos (município, setor censitário e códigos postais) para o mapeamento do risco de câncer de próstata utilizando uma abordagem geoestatística. Para manter a confidencialidade dos dados dos pacientes, os dados individuais foram representados através de setores censitários. Os mapas coropléticos dos três níveis de análise apresentaram diferentes padrões espaciais, o que configura *MAUP*. O autor realizou a modelagem espacial dos valores associados aos códigos postais, aplicando a técnica de deconvolução de semivariogramas. Este método consiste na obtenção de semivariogramas de suporte de pontos a partir de dados de área.

Na área de transportes, Wang, Zhuang e Liu (2012) analisaram como as técnicas de regressão aplicadas em análises espaciais na área de transportes são limitadas, apesar de muito utilizadas na previsão de demanda por viagens e acidentes de trânsito. Os autores discutiram estas limitações abordando entre elas a existência do *MAUP*, *falácia ecológica* e a dependência espacial.

Em geoestatística, o *MAUP* já vem sendo discutido em alguns trabalhos. Em exploração de mineração, por exemplo, amostras com apenas alguns metros cúbicos de volume são utilizadas para estimar o grau médio de minério em unidades de volume bem maiores. Cressie (1996) identifica esse problema na mineração como um caso especial de problema de mudança de suporte – *Change of Support Problem (COSP)*. Dado que a geoestatística busca realizar a inferência de informações em um domínio espacial contínuo, não considerar a mudança de suporte espacial pode resultar em previsões altamente tendenciosas.

Considerando a importância da mudança de escala na geoestatística, este estudo propõe a investigação da escala apropriada para variáveis de demanda por transportes, considerando distribuição populacional dos dados mais desagregados, além de homogeneidade de suporte. Sabe-se que resultados totalmente confiáveis, onde as soluções das análises não dependam do sistema de zoneamento utilizado, são assegurados a partir de dados individuais ou domiciliares. No entanto, dados individuais ou domiciliares, relacionados à demanda por transportes, quando disponíveis, apresentam elevado grau de aleatoriedade espacial, resultando em modelos com pouca estrutura espacial. Por isso, neste estudo, a análise por krigagem foi realizada para o suporte pontual e comparada a resultados de dados agregados através de dimensões de grade regular que representam a distribuição populacional domiciliar ou pontual.

A análise por zona de tráfego, usual na área de planejamento de transportes, está associada a problemas de não homogeneidade de suporte, suposição básica para aplicação da geoestatística, considerando que as zonas de tráfego são heterogêneas em forma e tamanho.

2.3. ESTUDO DA DEMANDA POR TRANSPORTES E A GEOESTATÍSTICA

A geoestatística é uma técnica de análise espacial que, recentemente, passou a ser utilizada em análises de demanda por transportes. Ela possibilita o desenvolvimento de estudos nos casos em que os valores das variáveis têm relação locacional. Um dos seus objetivos é caracterizar

uma variável através de sua distribuição espacial (WEBSTER e OLIVER, 2007; YAMAMOTO e LANDIM, 2013).

Desta maneira, a geoestatística permite, principalmente, analisar a variabilidade do fenômeno relacionado a uma Variável Regionalizada, através da função matemática estabelecida pelo modelo teórico do semivariograma. A análise estrutural, com base na variância do fenômeno em relação à distância, permite estimar o valor de uma variável em pontos sem informações sobre o atributo que se quer conhecer, além de obter o erro decorrente das estimativas (JOURNEL e HUIJBREGTS, 1978).

A partir dos trabalhos realizados pelo grupo de pesquisa no qual se insere este trabalho foram verificadas várias restrições relativas à aplicação da geoestatística para estimativas de variáveis relacionadas à demanda por transportes (GOMES, PITOMBO, ROCHA e SALGUEIRO, 2016; LINDNER, PITOMBO, ROCHA e QUINTANILHA, 2016; ROCHA, PITOMBO e SALGUEIRO, 2016). Dentre as restrições observadas, destacam-se aquelas relacionadas ao tipo de agregação ou o suporte dos dados. Em geoestatística o suporte define o método, espaçamento e/ou volume de obtenção das amostras (JOURNEL e HUIJBREGTS, 1978). Além disso, verificou-se que o comportamento individual pode variar bastante dentro de uma unidade de área que representa um atributo agregado, como por exemplo, a escolha do modo de transporte de um indivíduo ou o total de viagens por domicílio.

Uma outra constatação foi a existência de alta variabilidade a curtas distâncias para os casos de análise domiciliar ou por indivíduo (dados pontuais) (LINDNER, 2015; PITOMBO, SALGUEIRO, COSTA e ISLER, 2015; PITOMBO, COSTA e SALGUEIRO, 2015). Esta característica de dados desagregados, na maioria das vezes, estabelece problemas na representação do modelo de variabilidade, o que pode indicar que há alta aleatoriedade espacial da variável estudada.

A partir destes trabalhos, constatou-se que a escala de amostragem é um dos fatores que influenciam alguns parâmetros do modelo teórico do semivariograma (YAMAMOTO e LANDIM, 2013). No entanto, o método de interpolação, através de krigagem, só deve ser utilizado quando o semivariograma experimental apresentar estrutura espacial, ou seja, a variabilidade não pode ser totalmente aleatória (YAMAMOTO, 2001).

Assim, o uso da análise geoestatística em ciências sociais e/ou não naturais, como epidemiologia ou medicina geográfica, pressupõe algumas adaptações dos dados (GOOVAERTS, 2006; 2008; 2010). A adequação destes dados à análise geoestatística justifica-

se porque as variáveis não são naturais, como na mineralogia, geologia, geoquímica etc., onde a geoestatística surgiu e foi aprimorada.

Além disso, unidades de área usuais em análises de demanda por transporte, como zonas e setores censitários, possuem formas irregulares. Este é mais um obstáculo da aplicação da geoestatística a estudos de transportes, visto que esta técnica foi desenvolvida para ser aplicada a suportes regulares. Por isso, recomenda-se que dados de variáveis de demanda por transportes sejam agregados através de uma malha regular, de modo a adequá-los à aplicação de métodos geoestatísticos de estimação de dados.

2.3.1. Otimização para escolha de parâmetros da geoestatística

Em estudo que utilizam a geoestatística, além do processo de escolha da escala de obtenção das amostras, há uma etapa não menos importante: a descrição da estrutura espacial de uma Variável Regionalizada. Para demonstrar a autocorrelação espacial utiliza-se o semivariograma, um gráfico que explica a variabilidade de um atributo em função da distância. Para obter um semivariograma adequado para descrever tal aspecto de uma VR, é necessária a escolha de alguns parâmetros, tais como o *lag* (ou passo), a direção de anisotropia e a tolerância angular, além do ajuste a uma curva teórica.

O cálculo e ajuste automático do semivariograma experimental ainda é um desafio na geoestatística. Tradicionalmente, a escolha de tais parâmetros demanda tempo e amplo conhecimento sobre o comportamento estrutural de uma VR. A definição dos parâmetros para a obtenção do semivariograma experimental e o ajuste do modelo teórico, quando realizado de forma visual, acontece de acordo com a experiência e familiaridade do usuário, o que pode demandar tempo, pois a escolha destes parâmetros é realizada de forma empírica, por tentativa e erro.

Por outro lado, o uso de métodos de ajuste automático de semivariograma procura mesclar os princípios geoestatísticos e de técnicas de otimização. Por esta razão, métodos de inteligência computacional podem ser uma alternativa viável na substituição de técnicas tradicionais de cálculo e ajuste de um semivariograma experimental.

Poucos trabalhos que utilizam inteligência computacional aplicada à análise espacial, especificamente a geoestatística, são encontrados na literatura e, além disso, são muito recentes. Em estudos geológicos, alguns trabalhos já buscaram encontrar soluções adequadas para

problemas relacionados à geoestatística, por meio de recursos de otimização. Abedini, Nasseri e Burn (2012) propuseram o uso de Algoritmos Genéticos (AG) para encontrar o tamanho do suporte de domínio (raio ou elipse) para estimativas de dados referentes a recursos hídricos. Tal suporte determina quais dados seriam incluídos na predição de um ponto na etapa da krigagem. O emprego de técnicas de otimização possibilitou melhor performance da técnica em relação ao uso tradicional da geoestatística.

Segundo Cui, Yoneda, Shimada e Matsui (2016), existe grande dificuldade na escolha de pontos relativamente “ótimos” para obter a amostragem a ser utilizada na geoestatística. Além disso, a krigagem é uma etapa pouco prática, devido à complexidade em definir os valores dos parâmetros utilizados na fase de análise da estrutura espacial (etapa essencial para a estimação por krigagem), pois o semivariograma requer dados de amostragem adequados para gerar estimativas mais precisas. Neste trabalho, foi feita uma combinação entre geoestatística e AG para obter a localização de amostras adicionais.

São raríssimos os trabalhos que tratam, especificamente, do ajuste de semivariograma experimental. Li e Lu (2010) propõem um método para ajuste automático do semivariograma, através de programação linear, mesclando as vantagens do método polinomial ponderado e do método de programação linear. O intuito desta proposta de ajuste é utilizar o inverso da distância de *lag* como coeficiente de peso. Com isso, foi possível estimar, satisfatoriamente, parâmetros de um modelo esférico, que foi linearizado.

O estudo de Jiao e Lei (2011) criticou o uso de programação linear ponderada, devido à complexidade de implementação. Os autores também questionam o Algoritmo Genético tradicional, abordando o problema da taxa de convergência lenta ou prematura. Estes autores propuseram um Algoritmo Genético aprimorado para ajustar, automaticamente, o semivariograma experimental ao modelo teórico, através da introdução do método de otimização conhecido como *Simulated Annealing* (SA). O uso combinado destes métodos serviu para rever a taxa de seleção parental, aumentado para um a probabilidade de seleção do indivíduo que representa a solução ótima. O aprimoramento proposto tem como objetivo revisar a função *fitness* e mecanismos de sobrevivência, como cruzamento e mutação, para obter rápida convergência global. Vale ressaltar que em nenhum dos trabalhos citados foi discutido o uso de técnicas de otimização para o cálculo do semivariograma experimental.

Com isso, considera-se que métodos de otimização, integrados à modelagem geoestatística, ainda necessitam de bastante aprimoramento. Na área de análise da demanda por transportes,

esta abordagem pode trazer diversas contribuições práticas. Por isso, esta Tese de doutoramento busca conhecer algumas características espaciais de variáveis relacionadas à demanda por transportes através do uso conjunto de técnicas de inteligência computacional e geoestatística. A aplicação do método de otimização e seus resultados será discutida nos Capítulos 6 e 7 deste texto, respectivamente.

Este capítulo abordou diversos temas teóricos, relacionados aos problemas que motivaram a realização desta pesquisa. O Quadro 2.1 resume os problemas e as soluções sugeridas no presente trabalho.

Quadro 2.1: Problemas detectados e soluções sugeridas.

Problemas detectados	Soluções propostas
Modelos tradicionais para previsão de demanda por transportes, como Regressão Linear Múltipla, trazem pressupostos que são “quebrados” por conta da representação das variáveis de demanda	Uso de Modelos Lineares Generalizados (MLGs)
Modelos Lineares Generalizados podem apresentar melhor ajustes das variáveis de demanda, no entanto, não representam o fator espacial	Uso de técnicas de Estatística Espacial para previsão de demanda por transportes
Modelos espaciais confirmatórios, em geral, estimam a variável em locais de valores conhecidos	Uso da geoestatística para previsão de demanda por transportes
Escolha de escala adequada para modelagem geoestatística aplicada à demanda por transportes	Considerar distribuição de probabilidade dos dados mais desagregados na agregação de grades regulares
Cálculo e ajuste de semivariogramas experimentais (muitas decisões são baseadas na experiência do usuário; algumas limitações nos critérios de ajuste automático vigentes)	Proposta de uso de AG para cálculo e ajuste de semivariogramas

Capítulo

3

MÉTODOS DE OTIMIZAÇÃO

Este capítulo apresenta, sucintamente, alguns princípios fundamentais dos Algoritmos Evolutivos e introduz conceitos básicos sobre um dos seus ramos, os Algoritmos Genéticos, método de otimização utilizado neste trabalho.

3.1. ALGUMAS CONSIDERAÇÕES SOBRE TÉCNICAS DE OTIMIZAÇÃO

Técnicas de otimização buscam encontrar a melhor solução viável, considerando o objetivo do problema e um conjunto de restrições. Tais problemas podem ser originados de situações associadas a maximização ou minimização de determinada função (HAMAWAKI, 2005). Encontrar soluções “ótimas”, ou aproximadas, para um problema, muitas vezes não é trivial. Por isso, métodos heurísticos de otimização surgem como uma ferramenta eficiente.

Métodos heurísticos de otimização podem ser definidos como sendo um conjunto de técnicas de busca por boas soluções ou, ao menos, resultados próximos do ideal, para resolução de um problema (MARTÍ e REINELT, 2011; PEARL, 1984). No entanto, métodos heurísticos de otimização não estão preocupados em alcançar a convergência para solução ótima, mas garantir a melhor solução entre as possíveis. Por isso, o desafio de técnicas de otimização que utilizam métodos heurísticos é produzir soluções próximas à solução ótima dentro do seu espaço de busca.

Dentre os métodos heurísticos de otimização existem os de solução única e de população. Um destes métodos é o *Simulated Annealing (SA)*, que é empregado neste trabalho para ajustar o semivariograma experimental à curva teórica. Dentre os métodos que apresentam como resposta uma população, ou seja, um conjunto de soluções possíveis, destacam-se os Algoritmos Evolutivos e Colônia de Formigas. Os Algoritmos Evolutivos, sobretudo os Genéticos (ferramenta utilizada neste trabalho), serão discutidos nas próximas seções deste capítulo.

3.2. ALGORITMOS EVOLUTIVOS

Os algoritmos evolutivos fazem parte de um conjunto de técnicas de inteligência computacional. Os Algoritmos Genéticos (AGs) são um campo de estudo dos Algoritmos Evolucionários e, como tal, são baseados na Teoria de Seleção Natural, proposta no século XIX por Charles Darwin (BANZHAF, NORDIN, KELLER e FRANCONI, 1998). A principal ligação dos Algoritmos Evolucionários com as teorias de seleção e evolução de organismos naturais é o funcionamento das “estruturas” e dos “operadores genéticos” (LINDEN, 2008).

3.2.1. Algoritmos Genéticos (AG)

A introdução de inteligência computacional a processos genéticos teve início nos anos 1950 e 1960, quando introduziu-se os primeiros conceitos da biologia ao Aprendizado de Máquinas (GOLDBERG, 1989; GOLDBERG e HOLLAND, 1988). Desde então os AGs passaram a ser utilizados para solucionar problemas de otimização e Aprendizado de Máquinas.

A ideia básica do AG é, analogamente, simular processos naturais de adaptação, sobrevivência e reprodução dos indivíduos, mecanismos essenciais para a evolução das espécies, de acordo com a Teoria da Seleção Natural (BANZHAF, NORDIN, KELLER e FRANCONI, 1998).

A mesma ideia de competição pela sobrevivência entre indivíduos da mesma espécie foi introduzida na teoria do AG. Os indivíduos com maior chance de sobrevivência, ou seja, aqueles mais aptos, possivelmente terão uma maior chance de reproduzir. Do mesmo modo, indivíduos menos aptos tendem a desaparecer.

Os AGs podem ser considerados algoritmos de busca exaustiva, nos quais um conjunto de indivíduos de uma população é submetido aos processos de seleção e operadores genéticos para garantir a evolução e sobrevivência da espécie. A seguir encontra-se um conjunto de definições básicas dos elementos componentes dos AGs. A Figura 3.1 representa, sinteticamente, o funcionamento de um AG. A terminologia utilizada em AGs será apresentada, sucintamente, nesta seção. Tais definições básicas foram baseadas em Affenzeller, Wagner, Winkler e Beham (2009) e Sivanandam e Deepa (2007).

Figura 3.1: Esquema básico de funcionamento do AG.

População

O AG tem início a partir da geração de uma população inicial. A população é um conjunto de indivíduos que representam possíveis soluções de um problema, e que serão utilizados (ou não) para reproduzir um novo conjunto de indivíduos. Quanto maior o tamanho da população, maior a chance de diversidade genética e, conseqüentemente, maiores chances de sobrevivência. Caso contrário, populações muito pequenas podem diminuir a diversidade, essencial para a conversão a uma boa da função de avaliação, ou solução. Entretanto, se a população tiver muitos indivíduos, o algoritmo poderá perder grande parte de sua eficiência pela demora em avaliar a função objetivo de todo o conjunto a cada iteração.

Cromossomos ou indivíduos

A representação do problema que se deseja otimizar é uma etapa fundamental no AG. A forma de representação dos indivíduos traduz a informação do problema que se quer otimizar de maneira que se torne mais fácil de ser entendida pelo AG (LINDEN, 2008).

Os indivíduos podem ser representados por valores reais ou em forma binária. No AG binário, uma variável tem seu valor representado por uma sequência de *bits* (0 ou 1). Um número real é codificado através de um valor binário de n *bits*. Uma variável x binária pode assumir 2^n diferentes valores (HAUPT e HAUPT, 2004). Se uma variável do AG tem o número de *bits* igual a 4, por exemplo, ela pode assumir 2^4 valores, ou seja, é possível haver 16 valores diferentes para x entre os intervalos máximo e mínimo definidos.

Uma sequência binária **000..00**, representa o valor mínimo da variável x , enquanto um gene formado pela sequência **111..11** representa o limite superior do intervalo de busca.

Para calcular a função objetivo, muitas vezes é necessário conhecer os valores contínuos das variáveis. Por isso, a decodificação é realizada. Por exemplo, o valor que se deseja representar em forma binária é 51. A Figura 3.2 demonstra como é realizada a decodificação do valor binário para real.

Figura 3.2: Exemplo de representação binária no AG.

Gerações do AG

Cada novo conjunto de indivíduos originado a partir da população anterior dará origem a uma nova “geração” do AG, ou seja, um novo conjunto de soluções. A execução de uma grande quantidade de gerações torna possível a obtenção de possíveis soluções para o problema a ser otimizado. Cada indivíduo é composto por um conjunto de genes, que são os parâmetros a serem investigados pelo algoritmo.

Operadores genéticos

Os operadores genéticos são necessários para obter diversidade de características da espécie, através de sucessivas gerações, com a finalidade de obter a evolução da população, em outras palavras, se chegar a uma solução ótima. Dentre o conjunto de indivíduos de uma população, alguns são selecionados para o processo de reprodução e mutação. Os operadores genéticos são importantes para que a população se diversifique e mantenha características de adaptação adquiridas pelas gerações anteriores, onde, os indivíduos mais aptos são selecionados para reproduzirem, resultando em uma nova geração.

Na etapa de **seleção** ocorre a escolha dos indivíduos “pais” que irão compor a nova geração de indivíduos. O objetivo do AG é obter indivíduos com mais chances de sobrevivência. Sendo assim, os indivíduos mais aptos (com melhor avaliação) tendem a ser selecionados para a reprodução de descendentes. Os métodos de seleção mais utilizados são baseados nos princípios

de proporcionalidade, onde a chance de cada indivíduo ser sorteado, para fazer parte da etapa de cruzamento, é proporcional à sua aptidão. Os métodos de seleção devem procurar manter a diversidade populacional para evitar a convergência prematura (SIVANANDAM e DEEPA, 2007).

O **cruzamento** (*crossover*) é um operador genético essencial, pois através dele são criados novos indivíduos, a partir da troca de informações genéticas (características) de dois indivíduos “pais”. Em analogia a processos biológicos, os genes são responsáveis pelas características dos indivíduos e carregam toda a informação genética dos mesmos.

No *crossover*, as características de um indivíduo são trocadas a partir de um ponto de corte ou de cruzamento, que é selecionado aleatoriamente, entre o primeiro e o último gene do indivíduo “pai” (HAUPT e HAUPT, 2004) (Figura 3.3). O resultado desta operação é um indivíduo que, potencialmente, resulte em melhores características dos indivíduos usados como base.

Figura 3.3: Representação do *crossover*. Fonte: Affenzeller, Wagner, Winkler e Beham (2009).

Mutação é um operador que modifica, aleatoriamente, alguma característica do indivíduo (Figura 3.4). Esta alteração é importante para diversificar alguma característica, com o objetivo de produzir indivíduos mais aptos. É operador necessário para a introdução e manutenção da diversidade genética da população. Geralmente, assim como na natureza, ele é aplicado a uma taxa menor aos indivíduos do que o *crossover*.

Figura 3.4: Representação do processo de mutação.

O **elitismo** é um caso particular de seleção, onde parte dos indivíduos mais aptos da população sempre é preservada e nenhum dos indivíduos menos aptos é selecionado para as próximas gerações (LINDEN, 2008; SIVANANDAM e DEEPA, 2007). O elitismo pode ser usado para evitar a perda indesejada de informações durante a etapa de mutação. Teoricamente, a inclusão do elitismo no AG irá melhorar a performance do processo de otimização (SIVANANDAM e DEEPA, 2007).

Função objetivo

A função de avaliação (*fitness*) calcula a aptidão de cada indivíduo da população. Este é o componente mais importante do AG, pois indica se um indivíduo está apto a se tornar a provável solução de problema. A partir desta função se mensura a qualidade de uma solução e avalia-se, portanto, se um indivíduo está próximo da solução desejada. Assim, cada indivíduo recebe uma “avaliação” que é comparada a um valor “ótimo” encontrado. Em métodos de otimização, o melhor valor *fitness*, ou seja, a melhor resposta possível para um problema, é chamado de *solução ótima*.

Critério de parada do AG

Os procedimentos iterativos requerem a definição de um critério de parada para interromper a execução do algoritmo (RODRIGUES, LEITE, SANTOS, SOUZA e SILVA, 2004). Em problemas de otimização, o ideal seria se parar a execução assim que encontrar a solução ótima. Por outro lado, como AG é um método heurístico de otimização, várias soluções possíveis podem existir. Na prática, é difícil afirmar com total certeza, se um ponto ótimo é uma solução ótima, por isso, se faz necessário definir um critério de parada para o AG. Entre os critérios mais utilizados, em geral, são usados a aptidão do melhor indivíduo em conjunto com a limitação do número de gerações. Outro critério de parada bastante utilizado é o tempo de execução do algoritmo. Neste estudo, o AG executa um número n de gerações, para tentar obter uma solução ótima. Portanto, o AG tenta encontrar várias soluções viáveis, escolhendo aquela de melhor valor para a função objetivo.

3.3. SIMULATED ANNEALING (SA)

Simulated Annealing é uma técnica de otimização simples, que faz uma alusão ao processo de resfriamento de metais da termodinâmica. Dentro de um conjunto de soluções, a solução

“ótima” é encontrada através de um processo de perturbação ou busca na vizinhança. Esse processo tem início com uma “atribuição” ou “palpite” aleatório de um vetor. A cada iteração os valores do vetor são alterados e, se os resultados melhoram, as alterações são mantidas (KINSER, 2010).

A técnica SA é utilizada em problemas de otimização combinatória, para $\min_x f(x), x \in S$ ($S \rightarrow \mathbb{R}$). O procedimento de otimização é realizado simulando os níveis de temperatura T no processo de resfriamento. Inicialmente, T_0 e L_0 são fixados e atribui-se um ponto u inicial a partir de um espaço de S . Sendo $u \in S$, a cada nível de perturbação, vários pontos próximos de u serão gerados e o respectivo valor de f será calculado. Um novo valor de u é aceito ou rejeitado de acordo com uma certa probabilidade. Essa probabilidade de aceitação diminui de acordo com a temperatura (HAESER e GOMES RUGGIERO, 2008).

Quanto maior for a temperatura, maior a componente aleatória que será incluída na próxima solução escolhida, ou seja, T_0 deve ser grande o suficiente para que o algoritmo realize o maior número de iterações, para escapar de um minimizador local. A temperatura vai diminuindo aos poucos até atingir um estado de energia mínimo. O objetivo do SA é permitir que as soluções encontradas na vizinhança, a partir do processo de perturbações, sejam melhores que as iniciais. A probabilidade de uma solução pior ser selecionada vai diminuindo devido à redução de T .

3.4 PARTICULARIDADES DOS DOIS MÉTODOS

Apesar de se tratar de dois métodos estocásticos de otimização, o SA é um algoritmo baseado na busca de uma solução única, enquanto o AG é um algoritmo baseado em população. Isso significa que a SA começa com apenas uma solução e tenta aprimorá-la, enquanto o AG tem várias soluções possíveis, a partir de um determinado tamanho da população. Diante disso, deve-se considerar algumas particularidades destes métodos (Quadro 3.1).

Quadro 3.1: Particularidades do AG e SA

	Algoritmos Genéticos (AG)	<i>Simulated Annealing</i> (SA)
Inspiração	Biologia	Física
Tipo de solução	População	Única
Solução inicial	Aleatório	Aleatório
Processo de iteração	Combinação de diferentes soluções (cruzamento e mutação)	Modificação da solução inicial, controlada por T
Critério para escolha da melhor solução	Função de avaliação	$T \cong 0$

Capítulo

4

GEOESTATÍSTICA

Este capítulo apresenta os principais conceitos da técnica de análise espacial utilizada neste estudo. Serão apresentadas algumas abordagens da geoestatística importantes para o entendimento desta pesquisa, tais como: descrição da estrutura espacial de dados (caracterização das variáveis regionalizadas), semivariogramas, validação cruzada e Krigagem Ordinária (técnica de interpolação e mapeamento utilizada neste trabalho). Por fim será discutido o problema da mudança de escala na geoestatística.

4.1. CONCEITOS BÁSICOS DA GEOESTATÍSTICA

A modelagem de fenômenos espaciais através da geoestatística surgiu a partir da associação de fenômenos naturais e de embasamentos teóricos da estatística, principalmente da teoria das funções aleatórias (SOARES, 2000). Conceitualmente, a geoestatística pode ser definida como a aplicação a Teoria das Variáveis Regionalizadas (VR) para a compreensão da distribuição espacial de dados (JOURNEL, 1986).

A teoria geoestatística associa a aleatoriedade do fenômeno usando um modelo estocástico, onde a Variável Regionalizada é considerada como sendo uma entre muitas realizações possíveis de uma função aleatória (CHILÈS e DELFINER, 2012). Em síntese, a geoestatística consiste em um conjunto de métodos que quantificam o comportamento espacial de um conjunto amostral, com base na sua variabilidade estrutural (SOARES, 2000).

A origem da geoestatística se deu a partir do estudo estatístico de um fenômeno natural, levando-se em consideração suas características estruturais (JOURNEL e HUIJBREGTS, 1978). Em meados da década de 1960, o geólogo Georges Matheron utilizou o termo *geoestatística* para denominar um ramo da ciência que associa duas características intrínsecas na análise de depósitos de minério – de um lado a estrutura, e do outro a aleatoriedade do objeto de estudo (JOURNEL e HUIJBREGTS, 1978; MATHERON, 1970).

A partir dos anos 1980, as ciências ambientais, em áreas que envolviam poluição atmosférica, qualidade da água e do solo, por exemplo, começaram a aplicar técnicas de geoestatística

(WEBSTER e MCBRATNEY, 1987, 1989). Em outra fase, métodos geoestatísticos passaram a ser utilizados na área de recursos petrolíferos (HOHN, 1998). A partir dos anos 2000, esta técnica passou a ser utilizada em áreas como a geografia da saúde (GOOVAERTS, 2005, 2008) e no planejamento de transportes (GOMES, PITOMBO, ROCHA e SALGUEIRO, 2016; LINDNER, PITOMBO, ROCHA e QUINTANILHA, 2016; PITOMBO, SALGUEIRO, COSTA e ISLER, 2015; PITOMBO, COSTA e SALGUEIRO, 2015; ROCHA, PITOMBO e SALGUEIRO, 2016).

As próximas seções apresentam alguns aspectos teóricos básicos das etapas da análise geoestatística, incluindo a definição e cálculo do semivariograma, ferramenta fundamental para a estimação de valores de uma variável e objeto central deste estudo.

4.1.1. Variáveis Regionalizadas (VR)

As Variáveis Regionalizadas (VR) representam um conjunto de dados distribuídos espacialmente, por exemplo, dados de minério, pluviométricos ou demográficos (JOURNEL e HUIJBREGTS, 1978). Para caracterizar-se como uma VR, um conjunto amostral, necessariamente, apresenta uma forte dependência em relação à localização espacial. Portanto, a VR pode ser entendida como uma variável que possui dependência espacial e seu entendimento pode representar a distribuição espacial do fenômeno (YAMAMOTO e LANDIM, 2013).

De acordo com o formalismo geoestatístico, cada ponto não representa um único valor no espaço amostral, mas uma distribuição de probabilidade de ocorrência de valores, sendo que o conjunto de valores observados de $Z(x)$ é conhecido como VR (YAMAMOTO e LANDIM, 2013). Portanto, teoricamente, as VR representam características estruturais de uma forma matematicamente adequada para, na prática, estimar dados a partir de uma pequena amostragem (MATHERON, 1970).

De acordo com a Teoria das VR, os valores de uma variável não se distribuem espacialmente de forma aleatória, mas de acordo com determinada distância e direção. Por causa do seu duplo caráter, simultaneamente aleatório e estruturado, um valor $z(x)$ de uma VR é considerado como sendo uma realização de um processo aleatório de $Z(x)$ (CHILÈS e DELFINER, 2012).

Por outro lado, para se conhecer ou inferir o comportamento de uma variável a partir de uma única realização, a suposição de estacionaridade deve ser admitida. A hipótese de estacionaridade de uma VR permite supor que existe o mesmo grau de variação de uma VR no

espaço amostral (OLIVER e WEBSTER, 2014) e o grau de dependência entre dois pontos é influenciado apenas pela distância que os separa, não possuindo relação com sua localização na área de estudo (JOURNEL e HUIJBREGTS, 1978; SOARES, 2000). Por isso, a estacionaridade está relacionada à continuidade do processo aleatório $Z(x)$ (COSTA e FERNANDES FILHO, 2004). Portanto, segundo a Teoria das VR, é possível descrever uma variável levando em consideração sua localização no espaço para resolver problemas de estimativas.

4.1.2. Comportamento de uma VR

A partir da estrutura espacial de uma VR pode-se considerar que, quando a função de variabilidade $\gamma(h)$ depende apenas da distância entre as observações (h), o fenômeno é caracterizado como isotrópico (Figura 4.1a). Neste caso, o semivariograma experimental tem o mesmo comportamento em todas as direções do campo amostral, ou seja, é considerado omnidirecional.

Um fenômeno espacial é dito anisotrópico, quando o comportamento espacial da VR varia dependendo da direção angular (Figura 4.1b). A anisotropia estrutural representa a continuidade espacial atrelada às direções no espaço. Esta direção pode ser conhecida a partir da análise estrutural dos dados, sendo importante seu reconhecimento para a representação da variabilidade através do semivariograma. Desta forma, o alcance α varia de acordo a direção, sendo $\gamma_{\alpha1} \neq \gamma_{\alpha2}$ (JOURNEL e HUIJBREGTS, 1978).

Figura 4.1: Comportamento de uma VR. Fonte: Adaptado de Yamamoto e Landim, 2013.

O tipo mais comum de anisotropia é a geométrica (OLEA, 1999; SOARES, 2000). Na anisotropia geométrica (Figura 4.2), diferentes direções resultam em semivariogramas com mesmo patamar, mas com amplitudes diferentes, onde a máxima e a mínima amplitude são representadas por direções perpendiculares, respectivamente.

Figura 4.2: Representação anisotropia geométrica. Fonte: Soares, 2000.

É importante salientar que a consideração da existência de anisotropia pode acarretar o problema da redução da quantidade de pares considerados para a estimativa, o que pode prejudicar uma análise direcional, no caso de pequenos conjuntos de dados (OLEA, 1999).

4.1.3. Semivariograma

A variabilidade entre dois pontos $z(x)$ e $z(x + h)$ pode ser expressa pela função do variograma.

$$2\gamma(x, h) = E\{[Z(x) - Z(x + h)]^2\} \quad (4.1)$$

Quando esta função é representada por $\gamma(x, h)$ é conhecida como semivariograma. O estimador da função do semivariograma da variável regionalizada z entre dois pontos, x_i e x_j , separados por uma distância h é:

$$\gamma(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} [Z(x_i) - Z(x_i + h)]^2 \quad (4.2)$$

O cálculo da variância $\gamma(h)$ para um determinado valor de *lag* (h) é feito através somatório do quadrado das diferenças entre pares de pontos e dividido por duas vezes o número de pares de pontos utilizados. Desta forma, obtém-se a variância de um conjunto de dados, até uma distância máxima que separa os pares de pontos. A variabilidade do fenômeno é então representada graficamente pelo semivariograma experimental. Em seguida, a modelagem da variabilidade dos dados é feita através de uma curva de ajuste (semivariograma teórico). A

Figura 4.3 apresenta os parâmetros do semivariograma teórico, que serão utilizados na etapa de krigagem.

Figura 4.3: Estrutura do semivariograma. Fonte: Adaptado de DRUCK, CARVALHO, CÂMARA e MONTEIRO, 2004.

O semivariograma expressa variabilidade em função da distância e direção, além do comportamento da variabilidade a pequenas distâncias próximo à origem (YAMAMOTO, 2001). A seguir são definidos os parâmetros do semivariograma apresentados na Figura 4.3. Tais parâmetros serão utilizados nas equações dos modelos teóricos (Quadro 4.1).

- Efeito Pepita (C_0) é uma aparente descontinuidade junto à origem, onde a curva do modelo toca o eixo γ . O efeito pepita caracteriza o resíduo correspondente à variabilidades a pequena escala, ou seja, a variabilidade não detectada devido ao vetor h ser maior que as distâncias entre as observações (JOURNEL e HUIJBREGTS, 1978). Um valor alto deste parâmetro, também pode indicar a presença de grandes variações do conjunto dos dados a curtas distâncias.
- Alcance (α) é distância em que as amostras são correlacionadas espacialmente. Corresponde à distância a partir da qual o patamar é atingido (DRUCK, CARVALHO, CÂMARA e MONTEIRO, 2004).
- Patamar ($C = C_0 + C_1$) é o valor de γ máximo da curva correspondente à amplitude. C_1 é a contribuição ou variância espacial do modelo. A partir do patamar, considera-se que

não existe mais dependência entre as amostras, porque a variância da diferença entre pares de amostras torna-se, aproximadamente, constante (DRUCK, CARVALHO, CÂMARA e MONTEIRO, 2004).

As equações e a representação gráfica dos modelos teóricos de semivariograma mais utilizados na análise de fenômenos espaciais estão apresentados no Quadro 4.1. Nos modelos apresentados a semivariância atinge um patamar, ou seja, a variabilidade dos dados passa a ser constante a partir de uma distância h . No entanto, existem semivariogramas sem patamar, onde a variância não se estabiliza. Os modelos sem patamar mais conhecidos são o linear e de potência (YAMAMOTO e LANDIM, 2013).

Quadro 4.1: Modelos teóricos de semivariogramas. Fonte: Soares, 2000.

<p style="text-align: center;">Modelo Esférico</p> <p style="text-align: center;">$y(h)$</p> <p style="text-align: center;">Distância (h)</p>	$y(h) = C_0 + C_1 \left[1,5 \frac{h}{\alpha} - 0,5 \left(\frac{h}{\alpha} \right)^3 \right], h < \alpha \quad (4.3)$ $\gamma(h) = C_0 + C_1, h > \alpha$
<p style="text-align: center;">Modelo Exponencial</p> <p style="text-align: center;">$y(h)$</p> <p style="text-align: center;">Distância (h)</p>	$y(h) = C_0 + C_1 \left[1 - \exp \frac{-3h}{\alpha} \right] \quad (4.4)$
<p style="text-align: center;">Modelo Gaussiano</p> <p style="text-align: center;">$y(h)$</p> <p style="text-align: center;">Distância (h)</p>	$y(h) = C_0 + C_1 \left[1 - \exp \left(\frac{-3h^2}{\alpha^2} \right) \right] \quad (4.5)$

onde,

$\gamma(h)$ = semivariância;

C_0 = efeito pepita;

C_1 = contribuição do modelo;

$|h|$ = distância de *lag*,

α = alcance, ou seja, o limite da correlação espacial.

A escolha do modelo teórico é uma etapa importante na modelagem do semivariograma, pois, a sua função será utilizada para o cálculo dos ponderadores na etapa da krigagem.

4.1.4. Validação Cruzada

A Validação Cruzada, ou teste do ponto fictício, é a técnica que obtém a estimativa dos pontos amostrados, desconsiderando o valor verdadeiro da observação. Para estimar o valor de um ponto, são utilizados os valores de amostras vizinhas. Com isso, para cada ponto de valor conhecido inicial, existirá o valor verdadeiro (valor observado) e um outro valor estimado, podendo-se assim obter medidas de desempenho dos modelos (WACKERNAGEL, 2003). A partir da validação cruzada, podem ser analisados erro e variância da estimação, além do erro quadrático médio (EQM) e o desvio quadrático médio (DQM) (Equações 4.6 e 4.7).

$$\text{EQM} = \frac{1}{N} \sum_{i=1}^N \{z(x_i) - \hat{z}(x_i)\}^2 \quad (4.6)$$

$$\text{DQM} = \frac{1}{N} \sum_{i=1}^N \frac{\{z(x_i) - \hat{z}(x_i)\}^2}{\hat{\sigma}^2(x_i)} \quad (4.7)$$

onde,

$z(x_i)$ é o valor observado no ponto x_i ;

$\hat{z}(x_i)$ é o valor previsto pelo modelo para x_i e,

$\hat{\sigma}^2(x_i)$ é a variância da krigagem em x_i .

O DQM é utilizado para avaliar o quanto o erro da estimação real é igual, em média, ao erro previsto pelo modelo (WACKERNAGEL, 2003). Se o modelo de semivariograma está bem ajustado, então o valor do EQM deve estar próximo à variância de krigagem, logo o DQM deve

ser aproximadamente um (WEBSTER e OLIVER, 2007). Portanto, esta medida serve para analisar a qualidade do ajuste dos parâmetros do modelo. As medidas de desempenho, referentes aos resultados deste trabalho, serão discutidas no Capítulo 7.

4.2. ESTIMATIVAS POR GEOESTATÍSTICA

A geoestatística é um método de estimativas local, e seu objetivo é a modelagem da distribuição e variabilidade de um fenômeno espacial (YAMAMOTO e LANDIM, 2013).

Existem métodos mais simples de interpolação espacial como, por exemplo, o inverso do quadrado das distâncias. Este método consiste na determinação do peso de cada amostra a estimar um ponto, através do inverso do quadrado das distâncias que cada observação possui do ponto a ser estimado. No entanto, este método possui algumas limitações, como por exemplo, desconsiderar agrupamentos de amostras, ou seja, estando a uma mesma distância do ponto a ser estimado, qualquer agrupamento de amostras tem o mesmo peso que uma amostra isolada (SOARES, 2000).

A geoestatística evita estes problemas pois considera a estrutura espacial do fenômeno. A krigagem é um algoritmo que realiza a interpolação espacial de dados através dos pesos associados a cada amostra. Esses pesos são atribuídos a partir dos parâmetros do semivariograma teórico (WACKERNAGEL, 2003). Quanto mais próximo uma amostra estiver de um ponto a ser estimado, maior será sua influência no valor deste ponto. Por outro lado, quanto mais agrupados estiverem os dados, menor será o peso individual de cada amostra no sistema de estimação de um ponto (SOARES, 2000).

Na krigagem, a interpolação é realizada estabelecendo uma grade regularmente espaçada, onde cada nó da grade corresponde a um ponto x_0 a ser estimado. Como resultado da interpolação, tem-se um mapa de superfície de valores estimados de $Z(x)$. Além disso, como método estatístico, a krigagem fornece medidas de incerteza como, por exemplo, o erro e a variância da estimação (WACKERNAGEL, 2003). A próxima seção apresenta o método de interpolação espacial utilizado neste trabalho, a Krigagem Ordinária.

4.2.1. Krigagem Ordinária (KO)

Krigagem é uma técnica geoestatística de estimação linear que obedece dois critérios em relação ao seu erro: o não-enviesamento e variância mínima da estimação. A primeira condição

significa que a média dos valores estimados seja igual à média dos dados observados, de forma que a média de erros seja igual a zero. A variância dos valores estimados também determina a qualidade da interpolação espacial (SOARES, 2000).

A Krigagem Ordinária (KO) é utilizada quando se admite que as médias dos atributos são constantes e desconhecidas. O valor desconhecido $Z^*(x_0)$ (Equação 4.8) pode ser estimado a partir de uma combinação linear dos pesos λ_i com n valores vizinhos $Z(x_i)$ do ponto a ser estimado (JOURNAL, 1989).

$$Z_{x_0}^* = \sum_{i=1}^n \lambda_i Z(x_i) \quad (4.8)$$

Os pesos λ_i são calculados de modo a garantir que o estimador seja não-enviesado e a variância de estimação seja mínima. Por isso, o estimador KO é bastante utilizado. Os pesos ótimos e variância mínima são obtidos pelos parâmetros de Lagrange (JOURNAL e HUIJBREGTS, 1978). A esperança de um erro,

$$E[Z_{x_0} - Z_{x_0}^*] = 0 \quad (4.9)$$

é dada pela condição de não-enviesamento:

$$\sum_{i=1}^n \lambda_i = 1 \quad (4.10)$$

Atendendo a esta condição a variância das estimativas será minimizada. Tal restrição é obtida construindo-se um sistema de n equações com n incógnitas, então passa-se a $n + 1$ equações, introduzindo uma incógnita, também desconhecida, para satisfazer a condição de variância mínima estabelecida pela Krigagem Ordinária. Esta incógnita é o multiplicador de Lagrange (μ) (LANDIM, 2006).

Logo, a partir do sistema de equações da krigagem ($n + 1$), são calculados os ponderadores λ_n e a variância da estimação:

$$\begin{cases} \sum_{i=1}^n \lambda_i \gamma(x_i - x_j) + \mu = \gamma(x_0 - x_j), i = 1, \dots, n \\ \sum_{j=1}^n \lambda_j = 1 \end{cases} \quad (4.11)$$

$$\sigma_{KO}^2(x_0) = \sum_{i=1}^n \lambda_i \gamma(x_0 - x_i) + \mu \quad (4.12)$$

O sistema de equações da Krigagem Ordinária pode ser de representado de forma matricial,

$$[K] \cdot [\lambda] = [M] \quad (4.13)$$

onde,

K é a matriz da variância entre as amostras,

λ_n são os ponderadores da krigagem e

M é matriz da variância entre as amostras e o ponto a ser previsto.

Desenvolvendo os termos da Equação 4.14, monta-se a matriz da Krigagem Ordinária:

$$\begin{bmatrix} \gamma(x_1 - x_1) & \gamma(x_1 - x_2) & \dots & \gamma(x_1 - x_n) & 1 \\ \gamma(x_2 - x_1) & \gamma(x_2 - x_2) & \dots & \gamma(x_2 - x_n) & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \gamma(x_n - x_1) & \gamma(x_n - x_2) & \dots & \gamma(x_n - x_n) & 1 \\ 1 & 1 & \dots & 1 & 0 \end{bmatrix} \cdot \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_n \\ \mu \end{bmatrix} = \begin{bmatrix} \gamma(x_0 - x_1) \\ \gamma(x_0 - x_2) \\ \vdots \\ \gamma(x_0 - x_n) \\ 1 \end{bmatrix} \quad (4.14)$$

Para obter os valores dos ponderadores $[\lambda]$, são obtidos os produtos entre a matriz inversa de K e o segundo membro da matriz (JOURNEL e HUIJBREGTS, 1978).

$$[\lambda] = [K]^{-1} \cdot [M] \quad (4.15)$$

4.3. MUDANÇA DE SUPORTE E GEOESTATÍSTICA

Na geoestatística, o termo suporte define o método para obtenção de um conjunto amostral (JOURNEL e HUIJBREGTS, 1978). O conceito de suporte relaciona-se às dimensões (espaçamento e/ou volume) nos quais os dados são coletados e a sua influência na representação espacial de uma variável. Mudança de suporte de dados é tratada teoricamente, em muitas áreas de pesquisa, como o *Problema da Unidade de Área Modificável*, conforme discutido no Capítulo 2. Geralmente, o suporte representa valores de pontos, áreas ou blocos. Alterar o suporte de análise gera um novo conjunto de dados, relacionado ao original, mas com características espaciais diferentes.

Na análise de uma Variável Regionalizada, o cálculo da variância depende intimamente do suporte estabelecido. Como a geoestatística trabalha com Variáveis Regionalizadas, é necessário analisar a disposição espacial dos pontos, superfícies ou volumes a ser estudada através da sua variabilidade.

Em geoestatística, o *MAUP* foi discutido por Cressie (1996). O autor analisa que em mineração, por exemplo, amostras com apenas alguns metros cúbicos de volume são utilizadas para estimar o grau médio de minério em unidades de volume bem maiores. O autor associa tal problema como um caso especial de problema de mudança de suporte – *Change of Support Problem (COSP)*. Como a geoestatística realiza a inferência de dados em uma superfície contínua, não considerar a mudança de suporte espacial pode resultar em erros de estimativas, ou seja, pode envolver problemas como a *falácia ecológica*, ocasionando previsões altamente tendenciosas.

O formalismo para a mudança de suporte tem grandes aplicações práticas. Os dados de reservas de minério, geralmente, são associados a blocos, baseado na coleta de amostras pontuais. Portanto, são estimados valores médios por bloco através de suportes menores (LANDIM, 2006). Este procedimento é conhecido na geoestatística como regularização. Em geoestatística, técnicas de mudanças de suporte são capazes de avaliar a variabilidade espacial de grades médias em blocos por meio do variograma $2\gamma_V(h)$, definido pelo suporte de bloco (grades), e podem ser utilizadas para calcular a variância da estimação de um suporte Z_V para um suporte médio Z_v (JOURNEL e HUIJBREGTS, 1978).

Outro problema comum, relacionado à escala de análise, é a combinação de amostras obtidas por diferentes suportes espaciais. Por isso, devido às características dos dados das ciências sociais, a mudança de suporte tem como objetivo tornar possível a estimação de valores de uma variável em pontos ou blocos diferentes daqueles que foram observados (GELFAND, ZHU e CARLIN, 2001; GOOVAERTS, 2011).

Dados das ciências sociais e geografia médica, muitas vezes, estão associados a taxas da variável investigada. Nos dados de saúde, a grande preocupação é em relação à confidencialidade dos dados. Então, geralmente, tais dados são expressos pela razão entre o número de ocorrências a cada 100.000 habitantes, por exemplo, ou associados a áreas com forma e tamanho irregulares, como os setores censitários. Porém, na geoestatística se assume implicitamente que as áreas possuem o mesmo tamanho e forma, através da suposição de suportes geométricos homogêneos (GOOVAERTS, 2008).

No caso do planejamento de transportes, geralmente, os dados só estão disponíveis de forma agregada (na maioria das vezes a unidade utilizada é a zona de tráfego). Assim, devido à sua natureza, é necessário fazer algumas transformações para adequá-los ao uso de ferramentas geoestatísticas. Dentre estas transformações, estão o uso de relações como densidade ou proporções. Estes procedimentos metodológicos são completamente diferentes aos das análises

de atributos físicos pontuais, como minério ou contaminação do solo, para os quais as ferramentas geoestatísticas foram originalmente desenvolvidas. Por isso, a necessidade de entender técnicas de mudança de escalas de análise na geoestatística e quando elas são necessárias.

Vale ressaltar, ainda, que em planejamento de transportes é importante utilizar dados desagregados (pontuais) relativos a viagens, para se conhecer o comportamento individual, ao contrário do que é geralmente feito, utilizando-se informações agregadas. No entanto, verificou-se alguns entraves relacionados ao comportamento individual de variáveis relacionadas a viagens, como a escolha do modo de transporte, por exemplo. Esta característica dos residentes ou famílias provoca, nas análises geoestatísticas, alta variabilidade a curtas distâncias, principalmente nas cidades brasileiras, onde há forte presença de desigualdade socioeconômica. Esta característica, na maioria das vezes, ocasiona problemas na representação gráfica do modelo de variabilidade, resultando em semivariogramas com efeito pepita puro.

Por isso, neste trabalho, busca-se encontrar formas de agregar dados de viagens, através de grades regulares, com correspondência ao suporte de pontos (dados desagregados). O emprego de um método que represente dados pontuais de forma agregada, pode ser uma alternativa para obter melhores modelos variográficos (com menor variância e menor efeito pepita) e, conseqüentemente, melhores estimativas na área de planejamento de transportes.

Capítulo**5****MATERIAIS E MÉTODO:
DADOS E ÁREA DE
ESTUDO**

Este capítulo apresenta a área de estudo e a pesquisa Origem-Destino de 2007 da Região Metropolitana de São Paulo. As variáveis utilizadas neste trabalho serão apresentadas e descritas estatisticamente. Também são apresentados os procedimentos metodológicos que antecedem a etapa de otimização. Assim, um método para a definição da grade, que melhor representa o domicílio, será apresentado. Também será feita uma breve apresentação das ferramentas computacionais utilizadas. As etapas metodológicas, relativas à otimização por AG, e posteriores a estas, serão descritas no Capítulo 6.

5.1. APRESENTAÇÃO DA ÁREA DE ESTUDO

Este trabalho analisa o percentual de viagens domiciliares realizadas por automóvel (a razão entre o número de viagens por automóvel e o total de viagens realizadas por domicílio) e a renda média familiar, variáveis positivamente correlacionadas (com coeficiente de correlação igual a 0,46), no município de São Paulo (Figura 5.1).

Figura 5.1: Mapa de localização do município de São Paulo.

São Paulo é o município brasileiro com maior população, e sua região metropolitana é uma das mais populosas do mundo, com mais de 20 milhões de habitantes. Deste total, cerca de 11,5 milhões de pessoas vivem na capital, que apresenta densidade demográfica de aproximadamente sete mil habitantes por km² (IBGE, 2010).

A Figura 5.2 apresenta as formas de uso e ocupação do solo no município de São Paulo. A demanda por transportes está relacionada com atividades cotidianas, por isso, é necessário conhecer o uso do solo urbano (MCFADDEN, 1974). Usos mistos ou segregado do espaço urbano, como por exemplo, a densidade de atividades ou o espraiamento urbano, podem influenciar no quantitativo de viagens geradas em determinado local ou na escolha modal dos indivíduos (ORTÚZAR e WILLUMSEN, 2011; TSEKERIS e TSEKERIS, 2011).

Figura 5.2: Uso e ocupação do solo do município de São Paulo. Fonte: SMDU, 2016.

Na Figura 5.2 observa-se a predominância de área urbanizada, com cerca de 45% de ocupação da área do município. Áreas de proteção ambiental equivalem a aproximadamente 26% da extensão territorial, estando presente, principalmente, nos extremos norte e sul do município.

O padrão econômico das residências varia de alto para baixo, da área central para as bordas do município. Este padrão também é observado na análise da renda familiar e na taxa de uso do automóvel nas residências. Na área central (Figura 5.3), recorte estabelecido nesta pesquisa (descrito na Seção 5.2.1), observa-se a predominância de comércio e serviços (ocupação de 20% em relação a área total) e residencial de médio a alto padrão (ocupando cerca de 40% da área central). Áreas de proteção ambiental ocupam apenas 3% da área central.

Figura 5.3: Uso e ocupação do solo da área de estudo. Fonte: SMDU, 2016.

5.2. PESQUISA ORIGEM-DESTINO 2007

Os dados utilizados neste trabalho são provenientes da Pesquisa Origem-Destino (OD) da Região Metropolitana de São Paulo, realizada em 2007 pela Diretoria de Planejamento e Expansão dos Transportes Metropolitanos – Metrô/SP (METRÔ, 2007). Esta pesquisa foi realizada em 469 zonas de tráfego (Figura 5.4a), com uma amostra de 30.000 domicílios (Figura 5.4b).

(a) Zonas de tráfego

(b) Domicílios

Figura 5.4: Dados da Pesquisa Origem/Destino 2007. Fonte: Metrô, 2007.

A Pesquisa OD é importante para o planejamento de transportes, pois fornece características relacionadas aos deslocamentos realizados pela população em um dia útil. É possível identificar as viagens realizadas através do georreferenciamento das origens e destinos das viagens, possibilitando a previsão dos fluxos de deslocamentos mais frequentes.

Além de dados relacionados a viagens, características socioeconômicas dos indivíduos também são coletadas. Estas informações permitem estabelecer prováveis relações entre as características das viagens e aspectos econômicos da população. Estas relações podem ser utilizadas em estimativas de modelos para a previsão de viagens futuras, por exemplo.

A Figura 5.5 apresenta a divisão modal para as viagens realizadas na RMSP em 2007. O modo mais utilizado no período analisado foi o não motorizado, seguido pelo modo público.

Figura 5.5: Divisão modal da RMSP. Fonte: Metrô, 2007.

5.2.1. Seleção da área de estudo e apresentação dos dados

Os dados desagregados, por domicílio, relacionados a viagens da RMSP, não apresentaram estrutura espacial adequada para a aplicação da geoestatística em trabalhos anteriores. Diante disso, a área central do município de São Paulo foi selecionada, por apresentar maior padrão espacial aparente.

A geoestatística considera o pressuposto de que um evento ocorre uma única vez na mesma localização (mesmo par de coordenadas geográficas). Para os casos onde uma localização geográfica registou dois ou mais eventos, foi obtida a média aritmética dos valores das variáveis com mesma localização, passando-se a considerar apenas uma realização do evento. Assim, o banco de dados deste trabalho passou a ser composto por 8294 observações (domicílios) (Figura 5.6a).

(a) Localização dos domicílios

(b) Percentual de viagens por automóvel (%)

(c) Renda familiar (SM)

Figura 5.6: Localização dos domicílios e distribuição espacial das variáveis estudadas.
Fonte: Metrô, 2007.

Foram utilizadas as variáveis relacionadas aos domicílios da área de estudo: a) *Percentual de viagens realizadas por automóvel* (Figura 5.6b) e, b) *Renda familiar*, em salários mínimos (Figura 5.6c).

A Figura 5.7 apresenta a divisão modal para a área central do município de São Paulo. Nesta área, o modo mais utilizado é o automóvel, seguido pelo não-motorizado. Esta área possui renda mais elevada, o que pode explicar o aumento do percentual de viagens realizadas por automóvel em relação à RMSP.

Figura 5.7: Divisão modal da área de estudo. Fonte: Metrô, 2007.

5.2.2. Análise das variáveis estudadas

Esta seção descreve algumas características estatísticas das variáveis estudadas no presente trabalho. A descrição das variáveis, por domicílio, é apresentada na Tabela 5.1. Do total de 8294 domicílios, para a variável relacionada à escolha do automóvel, 34% (2830 casos) não realizou nenhuma viagem utilizando este modo.

Tabela 5.1: Estatísticas descritivas das variáveis por domicílio

	<i>Renda Familiar (SM)</i>	<i>Viagens por automóvel (%)</i>	<i>Viagens por automóvel (%)</i> [sem 'zero']
Média	13,42	47,19	71,62
Mediana	10,26	50,00	75,00
Moda	5,26	0,00	100,00
Desvio Padrão	10,84	40,98	28,27
Variância	117,58	1679,70	799,05
Mínimo	0,66	0,00	1,11
Máximo	92,10	100,00	100,00
1º Quartil	5,53	0,00	50,00
3º Quartil	18,42	100,00	100,00
N	8294	8294	5464

O elevado número de domicílios com esta característica causou alguns problemas iniciais nas análises, especialmente no código de escolha da grade. Por isso, estes domicílios foram

eliminados do banco de dados da variável relacionada à escolha modal. Assim, as análises sequenciais foram realizadas apenas com domicílios que utilizaram o automóvel (Figura 5.8). As Figura 5.9 e 5.10 apresentam a distribuição de frequência das variáveis pontuais utilizadas nesta pesquisa.

Figura 5.8: Percentual de viagens por automóvel (sem valores ‘zero’).

Figura 5.9: Distribuição de frequência da variável *Percentual de viagens por automóvel* (domicílio).

Figura 5.10: Distribuição de frequência da variável *Renda familiar* (SM) por domicílio.

5.3. PROCEDIMENTO PARA ESCOLHA DO TAMANHO DA GRADE

Devido a problemas relacionados a amostragem pontual em problemas de transportes, como descrito no Capítulo 2, e considerando a importância de se conhecer o comportamento de viagens de forma desagregada, este trabalho propõe a escolha de uma dimensão de grade regular que corresponda ao comportamento (distribuição de probabilidades) dos dados domiciliares. Para isso, foi utilizado o teste não-paramétrico de *Mann – Whitney*. Este teste compara a distribuição de frequência acumulada entre duas amostras independentes e verifica se elas são similares, sem a necessidade do conhecimento de distribuição populacional de ambas as amostras. Para um nível de confiança de 95%, a hipótese nula é rejeitada para o valor $p > 0,05$ (DINNEEN e BLAKESLEY, 1973; HOLLANDER e WOLFE, 1999).

O programa *RStudio* (através da função *wilcox.test*) foi utilizado para a elaboração de um código (Apêndice I) para verificar qual o tamanho de grade (para agregar os dados domiciliares pontuais) representaria melhor o comportamento do domicílio. Os testes de *Mann – Whitney* e o *Wilcoxon* são equivalentes (FIELD, 2013). O código gerou, através de *loops*, grades de cinco em cinco metros e obteve valores médios dos pontos inseridos em cada quadrícula gerada. Com isso, o teste não-paramétrico para cada configuração de grade testou se a distribuição acumulada dos dois pares de amostra (pontual x grade) eram similares. Através do p -valor, foi “tomada a decisão” se a hipótese nula foi aceita ou não aceita para cada grade. Depois desta etapa, foi determinado o limite de dimensão para a grade regular que considerasse o comportamento individual.

Assim, a grade selecionada foi aquela de maior tamanho (consequentemente menor erro de estimativa de krigagem) e que tivesse distribuição populacional similar à pontual. Foram selecionadas as grades de 85 e 75 metros para as variáveis *Percentual de viagens por Automóvel*

e *Renda familiar*, respectivamente. Apesar da grade de 75 metros ter sido aceita pelo teste de hipótese para a variável referente à escolha modal, utilizou-se a grade de 85 metros, devido a necessidade de escolha da maior grade possível para agregação de dados, a fim de minimizar os erros das estimativas no procedimento de krigagem.

O Quadro 5.1 resume os resultados do teste de *Mann-Whitney*, como descrito acima. A Figura 5.11 compara as distribuições de frequências acumuladas das variáveis agregadas e desagregadas de acordo o valor da grade determinada pelo teste não-paramétrico.

Quadro 5.1: Resultados do teste não paramétrico de *Mann – Whitney*.

Hipótese nula	<i>Viagens por automóvel (%)</i>			<i>Renda familiar</i>		
	Valor da grade (m)	p-valor	Decisão	Valor da grade (m)	p-valor	Decisão
“A distribuição é a mesma para as duas amostras”	5	0,91	aceita H_0	5	0,91	aceita H_0
	10	0,85	aceita H_0	10	0,92	aceita H_0
	15	0,72	aceita H_0	15	0,82	aceita H_0
	20	0,64	aceita H_0	20	0,72	aceita H_0
	25	0,52	aceita H_0	25	0,36	aceita H_0
	30	0,40	aceita H_0	30	0,57	aceita H_0
	35	0,25	aceita H_0	35	0,60	aceita H_0
	40	0,34	aceita H_0	40	0,25	aceita H_0
	45	0,27	aceita H_0	45	0,17	aceita H_0
	50	0,16	aceita H_0	50	0,25	aceita H_0
	55	0,17	aceita H_0	55	0,11	aceita H_0
	60	0,16	aceita H_0	60	0,10	aceita H_0
	65	0,08	aceita H_0	65	0,07	aceita H_0
	70	0,05	aceita H_0	70	0,09	aceita H_0
	75	0,07	aceita H_0	75	0,09	aceita H_0
	80	0,05	não aceita H_0	80	0,02	não aceita H_0
	85	0,07	aceita H_0	85	0,01	não aceita H_0
	90	0,01	não aceita H_0	90	0,03	não aceita H_0
	95	0,03	não aceita H_0	95	0,01	não aceita H_0
	100	0,03	não aceita H_0	100	0,00	não aceita H_0

Figura 5.11: Distribuição de probabilidade acumulada para os suportes pontual x grade.

A partir dos gráficos, pode-se inferir que, tanto amostra pontual da variável porcentagem de viagens por automóvel, quanto a de renda familiar, apresentam a mesma distribuição acumulada para o banco de dados onde foi aplicada uma agregação por meio da grade escolhida.

Após a escolha da dimensão das grades para cada variável, os dados pontuais foram agregados através de um *join* espacial. Com isso, foram obtidas as médias dos valores das variáveis para cada quadrícula da respectiva grade. A Tabela 5.2 apresenta um resumo estatístico das variáveis agregadas por grades (de acordo com o resultado do teste não-paramétrico).

Tabela 5. 2: Estatísticas descritivas das variáveis por grades.

	<i>Renda Familiar (SM) grade 75m</i>	<i>Viagens realizadas por automóvel (%) grade 85m</i>
Média	13,40	71,20
Mediana	10,53	75,00
Moda	5,26	100
Desvio Padrão	10,31	27,12
Variância	106,35	735,63
Mínimo	0,66	1,11
Máximo	84,21	100
1º Quartil	5,79	50,00
3º Quartil	18,42	100,00
N	6229	4320

Os gráficos de distribuição de frequências das variáveis agregadas pelas grades regulares definidas são apresentados na Figura 5.12.

Figura 5.12: Distribuição de frequência das variáveis para as grades selecionadas.

Verifica-se que os dados não obedecem à condição de normalidade. Em geoestatística, esta condição influencia a análise estrutural variográfica, pois o variograma experimental depende não apenas dos parâmetros escolhidos, mas também da distribuição de frequência dos dados (YAMAMOTO e LANDIM, 2013). No próximo capítulo, serão apresentadas as análises para as variáveis agregadas pela grade escolhida e para os dados pontuais (domicílios), através da geoestatística.

5.4. FERRAMENTAS UTILIZADAS

Neste trabalho propõe-se a utilização de técnicas de estatística espacial para analisar características domiciliares relacionadas a viagens urbanas. Um conjunto de ferramentas foi utilizado para realização das etapas metodológicas deste trabalho:

Caracterização estatística das variáveis

Os resultados de análises de estatísticas descritivas são essenciais para conhecer os dados e saber se os mesmos são apropriados para a análise geoestatística. A etapa de análises estatísticas convencionais foi realizada pela ferramenta Análises de Dados do *software Microsoft Excel* e pela linguagem R, através do *software RStudio* (sistema computacional estatístico livre).

Etapas de otimização

O *RStudio* também foi utilizado para desenvolvimento de um código para escolha da grade que agregou os dados utilizados neste trabalho, como descrito neste capítulo. Posteriormente a esta etapa, a otimização da escolha dos parâmetros dos semivariogramas experimentais e modelos teóricos foram determinados através de Algoritmos Genéticos, utilizando a linguagem C# (Apêndice II). Esta etapa será descrita detalhadamente no próximo capítulo.

Análise geoestatística

Após a escolha do “melhor” modelo através do AG, as etapas elementares da geoestatística, como a geração dos semivariogramas experimentais, os ajustes dos modelos teóricos e interpolação por Krigagem foram realizadas no *Stanford Geostatistical Modeling Software (SGeMS)*. A etapa da validação cruzada foi realizada no *software GeoMS 1.0*, desenvolvido pelo Centro de Recursos Naturais e Ambientais – CERENA, do Instituto Superior Técnico de Lisboa. Estes pacotes computacionais também foram utilizados para cálculo de semivariograma experimental e ajuste do modelo teórico de forma tradicional “*by eye*”, a fim de comparação dos resultados da etapa, realizada através do AG, e o método já consagrado entre os pesquisadores da área.

Sistemas de Informação Geográfica (SIG) foram indispensáveis para este trabalho, pois além de permitir a visualização e análise de atributos geográficos estudados, permitiram a elaboração de mapas de estimação através da krigagem. O ArcGIS 10.3 foi utilizado na etapa de geração de mapas de superfície de estimativas geoestatísticas. Foi possível a elaboração e edição gráfica de mapas, através de uma grade regularmente espaçada elaborada na etapa da krigagem.

5.5. CONSIDERAÇÕES METODOLÓGICAS

Este capítulo apresentou as etapas que antecederam a otimização e modelagem geoestatística, bem como apresenta alguns resultados (preliminares), referentes aos testes estatísticos para escolha do suporte geográfico. A Figura 5.13 apresenta o fluxograma metodológico seguido no presente trabalho. As caixas na cor cinza correspondem às etapas descritas neste capítulo. As caixas nas cores rosa e amarelo correspondem às etapas metodológicas descritas nos Capítulos 6 e 7.

Figura 5.13: Fluxograma das etapas do método.

Capítulo

6

APLICAÇÃO DO AG EM GEOESTATÍSTICA

Este capítulo procura abordar as principais etapas do processo de otimização, utilizado neste trabalho. Nele, serão discutidas as etapas para a obtenção dos parâmetros dos semivariogramas, correspondentes aos ótimos de cada execução do AG. A função objetivo do algoritmo e o critério de parada serão discutidos.

6.1. CONSIDERAÇÕES INICIAIS

A computação evolutiva vem ganhando destaque em diversas áreas de pesquisa. O método empregado para solucionar as questões apresentadas nesta pesquisa é o Algoritmo Genético (AG). O AG é um método de busca exaustiva e, faz parte da família dos Algoritmos Evolucionários. Como já mencionado no capítulo introdutório deste texto, esta pesquisa discute o problema de cálculo e ajuste de semivariogramas experimentais para estimação de variáveis relacionadas à demanda por transportes, a partir de ferramentas de otimização e geoestatística, levando em consideração os erros das estimativas.

A modelagem do semivariograma, além de exigir amplo conhecimento sobre o comportamento estrutural de uma Variável Regionalizada, é uma tarefa que demanda tempo. Geralmente, o ajuste do semivariograma é realizado visual e intuitivamente, a partir da experiência do pesquisador. Muitas vezes, este “empirismo” deixa a análise geoestatística extremamente cansativa e trabalhosa para o pesquisador iniciante, exigindo maior dedicação a esta etapa. Por esta razão, algoritmos de inteligência artificial podem ser uma alternativa viável para obtenção e ajuste de semivariogramas.

Considerando estas duas abordagens, esta tese busca propor um método com o intuito de buscar alternativas à modelagem tradicional, a partir da otimização da modelagem espacial de variáveis de transportes.

Etapas metodológicas, preliminares às discutidas neste capítulo, foram realizadas e discutidas anteriormente. A Figura 5.14 apresentou os procedimentos necessários para a execução deste trabalho, através de um fluxograma resumido. Após o tratamento dos dados e definição das

variáveis a serem estudadas, foi realizada a etapa para a definição de duas grades para agregar dados referentes a domicílios, conforme descrito no capítulo anterior.

Este capítulo discute os passos sequenciais, referentes à etapa de otimização, para a definição dos parâmetros dos semivariogramas experimentais e ajuste teórico dos mesmos. Os resultados do processo de otimização e análise dos resultados serão apresentados no Capítulo 7 deste texto.

A Figura 6.1 apresenta um esquema ilustrativo do procedimento sequencial do AG e da geoestatística para estimar dados relativos a viagens utilizado neste trabalho.

Figura 6.1: Ilustração do método proposto.

6.2. DETERMINAÇÃO DOS PARAMETROS DO SEMIVARIOGRAMA

O primeiro passo para aplicar o AG a este tipo de otimização, foi definir quais os parâmetros do semivariograma deveriam vir a compor os cromossomos ou indivíduos. Como visto no Capítulo 3, assim como na biologia, um indivíduo é formado por um conjunto de genes que configuram possíveis soluções de um problema. O Algoritmo Genético utilizado neste trabalho é binário, logo, cada variável do problema é representada por um conjunto de genes.

Neste trabalho, um indivíduo é composto por quatro parâmetros com oito genes binários cada. Desta maneira, um indivíduo possui 32 genes. Assim, como discutido no Capítulo 3, um indivíduo com estrutura binária possui 2^n valores possíveis. Logo, cada parâmetro do indivíduo pode encontrar 2^8 ou 256 valores diferentes, dentro dos intervalos (valores mínimos e máximos) estabelecidos para cada variável.

A Figura 6.2 ilustra a estrutura dos cromossomos utilizados neste trabalho. A próxima subseção detalha como foi calculado cada um destes parâmetros ou genes de cada indivíduo.

Figura 6.2: Representação hipotética de um cromossomo.

Sendo:

Gene 1: Direção angular

Gene 2: Tolerância angular

Gene 3: Largura de banda

Gene 4: Tamanho do *lag* ou passo

Os parâmetros de ajuste do semivariograma experimental foram determinados pela regressão não linear realizada pelo *Simulated Annealing*, discutido na Subseção 6.2.2.

6.2.1. Estrutura dos indivíduos

Primeiramente, foram determinados quais os parâmetros do semivariograma experimental seriam otimizados pelo AG. No caso de amostras irregulares, além de parâmetros como o *lag* e a direção são necessários parâmetros adicionais, como tolerância angular e largura de banda. Os passos para definição dos parâmetros dos cromossomos do AG utilizado encontram-se descritos mais detalhadamente a seguir.

Direção angular

Nos casos onde a continuidade espacial varia em diferentes direções no espaço (anisotropia), a estrutura espacial de uma Variável Regionalizada corresponde a uma **direção angular** particular. Tal direção define o ângulo ou azimuth no qual haverá a busca por pares de pontos para o cálculo do semivariograma experimental para os n lags ou passos (h), até uma distância de corte e número de lags pré-definidos. Neste trabalho, foram definidos limites para os valores de direção entre 0 e 180°. As direções definidas pelo AG foram comparadas aos mapas de semivariância (Seção 7.3). Esta ferramenta calcula a continuidade do fenômeno para todas as direções, onde o centro do mapa representa a distância “zero”. A principal função do mapa de semivariância é auxiliar na definição da maior e menor direções de continuidade espacial (PYRCZ e DEUTSCH, 2014).

Tolerância angular

Limitada entre 1 e 90 graus. Maior que 90° corresponde a um semivariograma omnidirecional (isotrópico).

Largura de banda

Geralmente, amostras irregulares necessitam de alguns critérios para a captura de pontos. Em muitos casos, a **largura de banda** (*bandwidth*) é utilizada para obter um limite máximo para o desvio do vetor da direção angular. Quando os dados são suficientes, a largura de banda pode ser pequena, em aproximadamente 1 a 3 vezes a distância de *lag* (DEUTSCH e JOURNEL, 1998; PYRCZ e DEUTSCH, 2014). Segundo Yamamoto e Landim (2013), este parâmetro permite delimitar a abertura infinita da janela de pesquisa, definida pela tolerância angular. A Figura 6.3 ilustra o esquema para determinar este parâmetro.

Figura 6.3: Esquema do cálculo da largura de banda.

Sendo d a distância de corte e LB a largura de banda, dada em metros, P_0 corresponde ao primeiro ponto x_0 , a partir do qual se abre uma “janela” para iniciar a pesquisa por pares de pontos e calcular os desvios quadráticos para a posterior obtenção do semivariograma experimental. Os pontos P_1, P_2, P_3, P_4 e P_5 são obtidos pelas Equações 6.1, 6.2, 6.3, 6.4 e 6.5, respectivamente.

$$\mathbf{P}_1 \Rightarrow x_1 = x_0 + d * \cos \theta; y_1 = y_0 + d * \sin \theta \quad (6.1)$$

$$\mathbf{P}_2 \Rightarrow x_2 = x_1 + LB * \sin \theta; y_2 = y_1 - LB * \cos \theta \quad (6.2)$$

$$\mathbf{P}_3 \Rightarrow x_3 = x_1 - LB * \sin \theta; y_3 = y_1 + LB * \cos \theta \quad (6.3)$$

$$\mathbf{P}_4 \Rightarrow x_4 = x_0 + t_4 * \cos(\theta + \alpha); y_4 = y_0 + t_4 * \sin(\theta + \alpha) \quad (6.4)$$

$$\text{onde, } t_4 = \frac{(y_3 - y_0) * \cos \theta - (x_3 - x_0) * \sin \theta}{\sin \alpha}$$

$$\mathbf{P}_5 \Rightarrow x_5 = x_0 + t_5 * \cos(\theta - \alpha); y_5 = y_0 + t_5 * \sin(\theta - \alpha) \quad (6.5)$$

$$\text{onde, } t_5 = \frac{(x_2 - x_0) * \sin \theta - (y_2 - y_0) * \cos \theta}{\sin \alpha}$$

Para identificar se um ponto está dentro ou fora da janela definida, um segmento de reta é traçado entre um ponto x_0 e os pontos que definem os limites do polígono ou janela (Figura 6.4). Um ponto insere-se na janela se,

$$\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 + \alpha_5 = 360^\circ$$

Figura 6.4: Esquema ilustrativo para identificar se um ponto pertence à janela de busca.

Tamanho do lag

O tamanho do **lag ou passo** é definido levando-se em consideração uma tolerância de *lag* que, geralmente, é dada por $\pm$ duas vezes o tamanho de h . A Figura 6.5 ilustra o esquema de aquisição de amostra para a construção do semivariograma experimental para amostras irregulares.

Figura 6.5: Ilustração da janela de busca para amostras irregulares. Fonte: Adaptado de Pyrcz e Deutsch, 2014.

6.2.2. Etapa de otimização do ajuste dos semivariogramas

Esta etapa corresponde a etapa de obtenção dos parâmetros do ajuste do modelo teórico do semivariograma, ou seja, **efeito pepita, contribuição e alcance**, descritos no Capítulo 4, além da definição da equação do modelo (**esférico, exponencial e gaussiano**). O ajuste foi definido utilizando o método de *Simulated Annealing (SA)*, conforme discutido no Capítulo 3.

A técnica de *SA* foi utilizada para ajuste da curva do modelo através do método dos mínimos quadrados ponderados, para a estimativa dos parâmetros do semivariograma teórico. Segundo este método, o melhor modelo é aquele que minimiza a diferença média quadrática entre os pontos do semivariograma experimental e a curva do modelo teórico. Muitas vezes, este método produz resultados comparáveis ao método manual (JIAN, OLEA e YU, 1996).

Na falta de experiência ou conhecimento prévio sobre o semivariograma experimental para definir um valor inicial para o vetor contendo os parâmetros do modelo teórico, recomenda-se o uso de valores *default* para iniciar o procedimento de iteração. Com isso, segundo Jian, Olea e Yu (1996), o “palpite” inicial pode ser encontrado aplicando-se as Equações 6.6, 6.7, e 6.8.

$$C_0 \text{ inicial} = \text{Max} \left[0, \gamma(h_1) - \frac{h_1}{h_2 - h_1} (\hat{\gamma}(h_2) - (\hat{\gamma}(h_1))) \right] \quad (6.6)$$

$$\alpha_{inicial} = \frac{h_k}{2} \quad (6.7)$$

$$C_{1\ inicial} = \frac{\hat{\gamma}(h_{k-2}) + \hat{\gamma}(h_{k-1}) + \hat{\gamma}(h_k)}{3} - C_{0\ inicial} \quad (6.8)$$

- $C_{0\ inicial}$, $C_{1\ inicial}$ e $\alpha_{inicial}$ correspondem aos valores *default* dos parâmetros efeito pepita, contribuição e alcance, respectivamente;
- γ representa a semivariância do modelo teórico e $\hat{\gamma}$ a semivariância experimental nas distâncias $h_1, h_2, \dots$, até uma distância de corte h_k .

A equação do modelo teórico (esférica, exponencial ou gaussiana) é fundamental para estimar a superfície de valores desconhecidos, além de prever dados em locais já amostrados na etapa de krigagem, o que permite avaliar os resultados da estimação. O método de otimização SA também foi utilizado para definir a equação do modelo teórico do semivariograma.

6.3. ETAPA DE OTIMIZAÇÃO POR AG

Após a determinação da estrutura cromossomial do algoritmo genético, foi definida a função objetivo (Equação 6.9) para avaliar cada indivíduo ou cromossomo.

$$f = \min \left\{ \frac{\sum_{i=1}^n |z(x_i) - \hat{z}(x_i)|}{n} \right\} \quad (6.9)$$

onde,

$\hat{z}(x_i)$ é o valor estimado por krigagem;

$z(x_i)$: valor conhecido e,

n : é o número de observações.

Com a definição da função de avaliação, foram realizadas dez execuções do algoritmo genético para obter as soluções ótimas (melhor resultado) para cada variável. Cada execução fornece os melhores resultados de cada geração, correspondentes aos menores valores da função objetivo. Neste trabalho, chamou-se de solução ótima o melhor valor de função *fitness*, por execução do AG. Dentre as dez melhores soluções obtidas, utilizou-se a melhor (indivíduo com menor *fitness*) dentre todos os indivíduos, para obter o semivariograma experimental e teórico das variáveis estudadas.

O AG proposto utilizou como critério de parada o número de gerações, ou seja, após encontrar a primeira solução viável (isto é, o valor mínimo da Função Objetivo), o algoritmo executou uma quantidade n de gerações. Ao término das n gerações, estipuladas para cada performance, o AG foi finalizado. A maioria das execuções utilizou como critério de parada a quantidade de 50 gerações. No entanto, diferentes números de gerações foram determinados, em alguns casos, para avaliar se houve relação entre a convergência da função objetivo e a quantidade de gerações.

Vale ressaltar que, para as variáveis desagregadas, por domicílio, foram realizadas apenas duas execuções do AG, pois o tempo de processamento do algoritmo é maior, devido à grande quantidade de pontos a serem considerados para o cálculo de semivariograma teórico e à etapa de validação cruzada (mensuração dos erros da etapa de krigagem), necessária para obtenção do valor *fitness* de cada indivíduo.

Após a avaliação de cada indivíduo e escolhido aquele que representa a solução ótima (com menor erro dentre a população de todas as gerações obtidas), a próxima etapa foi a construção e ajuste dos semivariogramas. A partir daí, foi possível estimar superfícies de dados de renda familiar e de percentual de viagens por automóvel por domicílio na região central de São Paulo, através das etapas descritas no capítulo anterior.

As Figuras 6.7 e 6.8 apresentam a relação entre o valor da função objetivo e a quantidade de gerações em cada execução. Com estes gráficos, foi possível notar que, de forma geral, o valor *fitness* já convergiu para a minimização da função objetivo entre a 10ª e 20ª gerações. Conclui-se, portanto, que o critério de parada pode ser diminuído, minimizando o tempo de execução do algoritmo.

Figura 6.6: Gráfico de evolução do AG – *Fitness* x Geração – *Percentual de viagens por automóvel*.

Figura 6.7: Gráfico de evolução do AG – *Fitness* x Geração – *Renda (SM)*.

A partir da análise dos gráficos de evolução dos AGs, verifica-se que na primeira execução para a variável Renda familiar, provavelmente, houve convergência prematura da função fitness, pois esse valor aparece quase como uma constante a cada geração do algoritmo.

O capítulo seguinte apresenta um resumo e uma breve discussão dos resultados obtidos através dos métodos utilizados neste trabalho. Nele também será apresentada a etapa de validação cruzada e suas respectivas medidas de desempenho, como os erros médio e quadrático. Os mapas de valores estimados a partir dos semivariogramas teóricos também serão apresentados. Além disso, o próximo capítulo apresenta uma comparação entre os modelos de ajuste automático por AG e manual, como também uma avaliação dos dois métodos de modelagem geoestatística. Finalmente, o Capítulo 8 aborda os principais problemas e desafio encontrados através da aplicação dos métodos discutidos nos Capítulos 5 e 6, além das recomendações necessárias e propostas para trabalhos futuros, provenientes dos desafios encontrados nesta pesquisa.

Capítulo **7** *RESULTADOS DA ETAPA DE OTIMIZAÇÃO*

Os resultados das etapas de modelagem (parâmetros dos semivariogramas resultantes da etapa de otimização) e estimação por krigagem serão analisados neste capítulo. Algumas medidas de desempenho dos modelos, obtidas através da validação cruzada, serão discutidas. Este capítulo traz, ainda, uma importante discussão relativa a uma proposta de “validação” do método utilizado neste trabalho: o mapa de semivariância.

7.1. OBTENÇÃO DE MODELOS VARIOGRÁFICOS ATRAVÉS DE AG

7.1.1. Distribuição de parâmetros de semivariogramas

Após a aplicação do AG, diversos semivariogramas experimentais foram calculados e ajustados. Em cada execução, a quantidade de resultados salvos dependeu da quantidade de gerações estabelecidas. No AG empregado foi estabelecido um valor fixo de 100 indivíduos para cada geração. Assim, para uma execução com critério de parada igual a 50 gerações, por exemplo, foram obtidos aproximadamente 5000 novos indivíduos. Tais resultados foram associados a soluções quase ótimas. Neste trabalho, ‘solução ótima’ é o termo dado ao melhor resultado, associado à função objetivo, dentre todas as execuções realizadas.

Alguns dados foram gerados e serviram para verificar as respostas mais frequentes relativas aos parâmetros dos semivariogramas (*genes dos cromossomos* – como foi visto no capítulo anterior). Foram selecionados os 10 melhores cromossomos de cada execução do AG para compor um conjunto de 100 indivíduos mais aptos (menor erro), para as variáveis por grades. No AG executado, para o suporte domiciliar, foram selecionados os 100 melhores indivíduos para a execução que apresentou melhor *fitness*. Com isso, foi possível encontrar diversas soluções quase ótimas, além das soluções ótimas, escolhidas para representar o fenômeno espacial analisado através de semivariogramas. As Figuras 7.1, 7.2, 7.3 e 7.4 apresentam os intervalos de distribuição dos parâmetros dos semivariogramas experimentais e teóricos, relativos aos 100 melhores indivíduos obtidos em cada execução do AG.

Figura 7.1: Distribuição de frequência dos parâmetros do AG para a variável *Percentual de viagens por automóvel – domicílio*.

Figura 7.2: Distribuição de frequência dos parâmetros do AG para a variável *Renda familiar – domicílio*.

Figura 7.3: Distribuição de frequência dos parâmetros do AG para a variável *Percentual de viagens por automóvel* – grade regular.

Figura 7.4: Distribuição de frequências dos parâmetros do AG para a variável *Renda familiar – grade regular*.

A partir da análise dos histogramas apresentados, verifica-se que os valores mais frequentes dos parâmetros dos semivariogramas são, aproximadamente, os mesmos para os dois diferentes suportes de cada variável analisada. Assim, estes intervalos podem ser considerados como, relativamente, mais adequados para analisar estas variáveis por meio da geoestatística. Essa é uma das importantes contribuições associadas ao uso do AG para cálculo e ajuste de semivariogramas. É possível obter uma distribuição de valores de parâmetros para o cálculo de semivariogramas experimentais e ajuste a uma curva teórica. A partir das execuções é possível, também, obter um intervalo para cada um dos parâmetros, conforme apresentado na Tabela 7.1.

Tabela 7.1: Intervalos dos parâmetros para cálculo e ajuste de semivariogramas (soluções quase ótimas e ótimas)

Parâmetros AG	Auto dom. (%)		Auto grade (%)		Renda dom. (SM)		Renda grade (SM)	
	<i>Mín.</i>	<i>Máx.</i>	<i>Mín.</i>	<i>Máx.</i>	<i>Mín.</i>	<i>Máx.</i>	<i>Mín.</i>	<i>Máx.</i>
<i>Direção</i>	91,76	139,06	87,53	132,71	104,50	110,80	81,88	111,53
<i>Tol. angular</i>	1,00	75,69	1,35	86,86	3,44	88,60	3,09	87,91
<i>Larg. de banda</i>	10,00	500,00	10,00	459,65	10,00	209,84	19,61	250,20
<i>Lag</i>	10,00	253,07	10,00	310,25	10,00	322,88	10,00	282,95
<i>Efeito pepita</i>	448,50	621,40	488,80	562,10	74,02	84,37	0,00	72,51
<i>Contribuição</i>	132,40	304,30	136,20	202,10	26,33	37,15	28,80	108,29
<i>Alcance</i>	550,50	1065,00	674,00	1394,00	642,40	1135,00	987,70	1810,70

A Tabela 7.2, em seguida, apresenta os parâmetros escolhidos para a modelagem dos semivariogramas (soluções ótimas). Esses valores foram utilizados na etapa posterior para obtenção dos “semivariogramas ótimos”, descrita na subseção subsequente.

Tabela 7.2: Parâmetros dos indivíduos – soluções ótimas do AG

Parâmetro	Variável			
	Auto dom.	Auto grade	Renda dom.	Renda grade
<i>Fitness</i>	25,34	24,46	7,43	7,71
Direção principal (°)	115,06	109,41	108,00	104,47
Tol. Angular (°)	33,11	46,37	32,06	56,84
Larg. de banda (m)	15,76	10,00	33,06	21,53
<i>Lag</i> (m)	96,53	182,50	50,34	145,90
Efeito pepita (C_0)	598,66	504,36	74,88	55,04
C_1 Contribuição	153,47	182,36	36,28	46,01
Alcance (m)	550,47	674,00	833,86	987,65
Modelo teórico	Exponencial	Exponencial	Exponencial	Exponencial

7.1.2. Geração de semivariogramas experimental e teórico

A Figura 7.5 apresenta os “semivariogramas ótimos” calculados e ajustados a partir dos melhores resultados do AG. Os parâmetros ou *genes* da solução ótima, ou seja, o indivíduo com menor *fitness* dentre as execuções realizadas, foram escolhidos para elaboração dos semivariogramas no *software SGeMS*.

Figura 7.5: Semivariogramas obtidos por AG. a) *Percentual de viagens por automóvel – domicílio*; b) *Percentual de viagens por automóvel – grade regular*; c) *Renda familiar*; d) *Renda familiar – grade regular*.

Em relação à mudança de escala e sua influência nos semivariogramas, observa-se que houve uma diminuição do efeito pepita para as duas variáveis analisadas neste estudo (Figura 7.5). No caso da variável relacionada ao uso do automóvel, houve uma diminuição de 18% para este parâmetro. A mudança de escala ocasionou uma diminuição de 36% no efeito pepita para a análise da variável renda familiar.

7.2. ESTIMAÇÃO DE DADOS DE VIAGENS POR KRIGAGEM ORDINÁRIA

Os mapas de valores estimados referentes aos melhores modelos obtidos por AG são apresentados na Figura 7.7. Observa-se que os valores mais elevados da variável *Percentual de viagens realizadas por automóvel* podem ser comparados, visualmente, aos valores de renda média (SM) nas zonas de tráfego (Figura 7.6). Esta constatação mostra a influência da renda na escolha do automóvel, no período analisado. Regiões paulistas, tradicionalmente consideradas nobres, apresentam as maiores taxas de viagens por automóvel nos domicílios. O contrário também pode ser observado nas áreas que apresentam baixa porcentagem de viagens por automóvel, como na região da Sé, centro antigo da cidade de São Paulo, que possui menor renda média.

Figura 7.6: Renda familiar média por zona de tráfego.

Percentual de viagens realizadas por automóvel – domicílio

Percentual de viagens realizadas por automóvel – grade 85m

Renda – domicílio

Renda – grade 75m

Figura 7.7: Mapas estimados através dos modelos otimizados.

7.3. VALIDAÇÃO DA OTIMIZAÇÃO ATRAVÉS DE PROCEDIMENTOS TRADICIONAIS

A fim de comparar os resultados das otimizações (“semivariogramas ótimos” e krigagem), foram elaborados mapas de semivariância (Figura 7.9 e 7.10). Estes mapas indicam a semivariância para todas as direções, permitindo detectar a direção angular para cálculo do semivariograma experimental (para os casos com presença de anisotropia), além de fornecer informações como a presença de efeito pepita.

No mapa de semivariância (*varmap*), o centro do mapa representa o *lag* zero (Figura 7.8) e a quantidade de *pixels* representa a quantidade de *lags* adotada. A partir da construção destes mapas, os valores da semivariância podem ser representados por uma escala de cores, sendo possível também identificar as direções de maior e menor continuidade espacial (CHILÈS e DELFINER, 2012).

Sendo assim, mapas de semivariância foram utilizados para comparar os resultados das soluções ótimas, obtidos através da etapa de otimização, para verificar, sobretudo, se a direção escolhida pelo AG está próxima à direção de maior continuidade espacial, evidenciada pelos referidos mapas.

Figura 7.8: Esquema de um mapa de semivariância. Fonte: Adaptado de Pyrcz e Deutsch, 2014.

Para fins comparativos, os mapas das Figuras 7.9a, 7.9b, 7.10a e 7.10b foram obtidos pelo pacote *gstat* do programa *RStudio*. Os mapas de semivariância das Figuras 7.9c, 7.9d, 7.10c e 7.10d foram obtidos através do *software WinGslib*.

A partir das informações obtidas nestas análises, a etapa de análise geoestatística referente ao cálculo e ajuste teórico de semivariograma foram feitos da maneira tradicionalmente realizada, especialmente na área de ciências ambientais: manualmente ou “*by eye*”, considerando conhecimento sobre a Variável Regionalizada e, também, a experiência do pesquisador.

Figura 7.9: Mapas de semivariância da variável *Percentual de viagens por automóvel*.

Figura 7.10: Mapas de semivariância da variável *Renda familiar* (SM).

A análise dos mapas de semivariância, apresentados nas figuras acima, permitiu verificar alguns aspectos relacionados à estrutura espacial das variáveis analisadas, dentre eles:

- As variáveis relacionadas à escolha do automóvel por domicílio ou por grade regular possuem uma estrutura anisotrópica com um efeito pepita considerável. Esta estrutura tem sua continuidade máxima, aproximadamente, na direção 30°E , com um intervalo entre 15 e 20 unidades de *lag*. No entanto, para variável 'auto 85m', o alcance é menor. É possível observar ainda que, para estas duas variáveis, possivelmente há uma outra estrutura, também anisotrópica, na direção 15°O .
- As variáveis relacionadas à renda têm um comportamento isotrópico, com alcance maior para variáveis relacionadas a domicílios, do que aquela representada por grade regular de 75m.

Diante destes pontos apresentados acima, considerou-se importante, para este trabalho, analisar as variáveis estudadas da forma tradicionalmente abordada na geoestatística, utilizando o ajuste manual em um *software* da área. Com isso, foram feitos o cálculo e modelagem do semivariograma experimental, considerando as informações contidas nas Figuras 7.9 e 7.10. Os semivariogramas resultantes desta etapa de análise convencional da geoestatística são apresentados na Figura 7.11. A Figura 7.12 apresenta os mapas de estimação desta análise.

Ajuste manual

Figura 7.11: Semivariogramas obtidos por ajuste manual. a) *Percentual de viagens por automóvel – domicílio*; b) *Percentual de viagens por automóvel – grade regular*; c) *Renda familiar*; d) *Renda familiar – grade regular*.

Percentual de viagens realizadas por
automóvel – domicílio

Percentual de viagens realizadas por
automóvel – grade 85m

Renda – Domicílio

Renda – Grade 75m

Figura 7.12: Mapas estimados através de ajuste variográfico manual.

Para verificar a performance do AG, nos casos das variáveis que apresentaram anisotropia, foi verificado, dentre os resultados obtidos com a otimização, se houve a definição da direção principal de anisotropia, referente à detectada através dos mapas de semivariância. Com isso, foram calculados e ajustados semivariogramas na direção de aproximadamente 30° , para a variável *Percentual de viagens por automóvel* (desagregadas e por grade) (Figura 7.13), a partir de resultados do AG.

Figura 7.13: Semivariogramas referentes a anisotropia do *varmap* obtidos com a otimização.

Verifica-se que os semivariogramas são relativamente diferentes aos obtidos manualmente, inclusive para a mesma direção de anisotropia. Os demais parâmetros definidos pelo AG foram bem diversos aos utilizados no ajuste manual. A próxima subseção apresenta algumas medidas de desempenho provenientes da validação cruzada para ambos os métodos.

7.4. MEDIDAS DE DESEMPENHO DOS MODELOS – VALIDAÇÃO CRUZADA

Após a etapa de análise espacial através dos semivariogramas, foi realizada a etapa de validação cruzada para as variáveis analisadas (desagregadas e agregadas por grades). A Tabela 7.3 apresenta um resumo das medidas de desempenho de cada um dos modelos obtidos através do ajuste otimizado (AG) e manual (ajuste geoestatístico convencional). Vale ressaltar que as

medidas de desempenho foram calculadas com base nos valores estimados (pelo AG e manual) e valores observados.

Tabela 7.3: Medidas de desempenho dos modelos geoestatísticos.

Suporte	Variável	Erro médio	EQM	Variância média	DQM
<i>Domicílios</i>	Auto "manual"	23,14	746,06	715,12	1,04
	Auto AG	23,17	750,97	729,86	1,03
	Renda "manual"	6,89	92,93	94,05	0,99
	Renda AG	6,89	93,25	94,18	0,99
<i>Grade regular</i>	Auto "manual"	21,88	680,72	653,69	1,04
	Auto AG	21,90	683,27	650,40	1,05
	Renda "manual"	6,51	82,82	82,28	1,01
	Renda AG	6,53	83,17	76,85	1,08

Através da análise das medidas de desempenho, aparentemente não houve diferenças significativas nas estimativas obtidas através dos semivariogramas ajustados pelo AG ou de forma manual. Por outro lado, as variáveis agregadas por grade regular apresentaram melhores resultados das medidas como erro médio, EQM e variância média da krigagem. Outros produtos relacionados às medidas de desempenho da validação cruzada, como histogramas de erros e dispersão entre valores observados e estimados, encontram-se nos Apêndices IV, V e VI. Os resultados da validação cruzada comprovam a eficiência do método de otimização proposto, aplicado à modelagem geoestatística.

Capítulo

8

CONSIDERAÇÕES

FINAIS

O último capítulo desta tese procura levantar as principais conclusões obtidas a partir deste estudo. Nele serão abordados os problemas de pesquisa que motivaram este trabalho, a fim de discutir as principais conclusões e pontuar as restrições encontradas ao longo do seu desenvolvimento. Por fim, com base nos entraves encontrados, serão apresentadas algumas sugestões que podem vir a fundamentar pesquisas futuras.

8.1. SÍNTESE DO ESTUDO

A compreensão do comportamento individual relativo a viagens é indispensável na análise da demanda por transportes. Além disso, dados espaciais aumentam a capacidade de previsão dos modelos, por considerarem a importância da localização das amostras. Considerando estes dois pressupostos, este trabalho procurou propor a análise de variáveis relacionadas à demanda por transportes, levando-se em conta duas abordagens diferentes e complementares:

- a) Otimizar a busca de parâmetros para o cálculo e ajuste do semivariograma experimental (procedimento tradicionalmente realizado de forma manual, por tentativa e erro);
- b) Entender o comportamento individual através da agregação de dados pontuais (domicílios ou indivíduos), com o intuito de contornar problemas relacionados à mudança de escala, como os relacionados à *falácia ecológica*, buscando minimizar problemas ocasionados pela alta variabilidade de dados desagregados por domicílios.

Com isso, este estudo foi norteado através de dois principais problemas de pesquisa:

1. O cálculo e ajuste de semivariogramas (que é uma etapa importante da modelagem geostatística, podendo ser morosa para o caso de ajuste manual e para um pesquisador iniciante, por exemplo, tendo restrições associadas a alguns ajustes automáticos utilizados comumente);

2. Escolha da unidade de área ou suporte geográfico (sabendo-se que o domicílio representa bastante aleatoriedade espacial à análise e a agregação por Zona de Tráfego não respeita o pressuposto de suporte geográfico homogêneo).

Desta forma, um problema está estritamente relacionado ao objetivo central e outro ao secundário. Com isso, o principal ponto de partida desta tese foi propor e analisar um método de otimização para a obtenção de semivariogramas experimentais e teóricos e posterior estimativa por krigagem, com o objetivo de analisar variáveis relacionadas à escolha modal. Além disso, levou-se em consideração a importância do suporte ou escala de análise (o que interfere diretamente nos resultados dos modelos). A Figura 8.1 sintetiza as principais etapas deste trabalho.

Figura 8.1: Síntese da tese.

Por se tratar de análise espacial, através da geoestatística, a questão chave deste estudo é o semivariograma, ferramenta indispensável em análise espacial através da krigagem. Por isso, foi implementado um código de otimização para a obtenção de semivariogramas experimentais e teóricos para obter estimativas, também, em locais sem informação. Vale ressaltar, que esta é uma importante vantagem da krigagem em relação a outros métodos de análise espacial.

Com a execução do AG, foram obtidas diversas soluções para a construção e modelagem do semivariograma experimental, a partir de combinação de possíveis valores dos parâmetros envolvidos na etapa de análise variográfica. No Apêndice III encontra-se disponibilizado um tutorial contendo o link de acesso e alguns exemplos dos passos necessários para a execução do código do AG utilizado neste trabalho.

A análise dos valores da função de avaliação ou objetivo (menor erro dos modelos) gerou outra importante contribuição deste trabalho. Entende-se, com isso, que a geração de intervalos de valores mais frequentes dos parâmetros dos semivariogramas pode vir a corroborar a compreensão da estrutura espacial e escolha dos parâmetros definitivos dos modelos de semivariograma para uma dada VR.

A definição de uma malha regular influenciou positivamente a qualidade do ajuste dos semivariogramas. Através da análise geoestatística para esse suporte, observou-se que houve um significativo declínio do efeito pepita para as duas variáveis analisadas neste estudo (diminuição de 18% para a variável relacionada ao percentual do uso do automóvel e de 36% para a variável *Renda familiar*).

Como forma de corroborar os resultados da aplicação de otimização para modelagem de semivariograma, os resultados obtidos, pela otimização, foram analisados e comparados sob o ponto de vista da análise geoestatística mais tradicional. No entanto, os mapas de semivariância (*varmaps*) indicaram direção de continuidade espacial (anisotropia) para ângulos de azimuth diferentes dos observados pelos intervalos de frequência das soluções quase ótimas. Um aspecto relevante a ser considerado, é que a variável sobre renda familiar apresentou comportamento isotrópico para os dois suportes. No entanto, o intervalo de valores observados para o parâmetro “direção angular”, por exemplo, foi bastante parecido aos obtidos para a variável sobre a escolha do automóvel. Esta constatação atesta a relação entre as duas variáveis, já verificada anteriormente.

A seguir, discute-se as principais conclusões referentes aos resultados apresentados no capítulo anterior. Por fim, algumas sugestões para trabalhos posteriores serão apresentadas.

8.2. CONCLUSÕES E RECOMENDAÇÕES

8.2.1. Conclusões e contribuições

Ferramentas de otimização podem ser uma importante alternativa para a construção de modelos variográficos. Este trabalho é um dos pioneiros neste tipo de análise na área de demanda por transportes, sobretudo em relação à definição dos parâmetros para o cálculo do semivariograma experimental. No entanto, é necessário fazer algumas ressalvas concernentes à aplicação do método proposto, principalmente considerando-se as hipóteses de pesquisa que motivaram este trabalho:

1. *O uso de Algoritmos Genéticos otimiza a modelagem geoestatística aplicada a variáveis de demanda por transportes;*
2. *O conhecimento de distribuição e intervalo de parâmetros do variograma teórico, a partir de soluções quase ótimas, oriundas do AG, pode colaborar para o entendimento da estrutura espacial da variável regionalizada.*

Em relação à hipótese 1, foi verificado que a modelagem com parâmetros otimizados não interfere na qualidade dos modelos, fazendo-se uma comparação com a modelagem tradicional visual. Portanto, a partir desta comparação entre o método proposto nesta tese e um método mais tradicional, julga-se que a grande vantagem do uso do AG é a economia de esforços para modelar adequadamente um semivariograma.

Outra importante questão a ser levantada, é sobre a obtenção de intervalos de parâmetros mais frequentes para o cálculo e ajuste teórico dos modelos variográficos. Entende-se que estes resultados podem representar uma importante contribuição metodológica para os usuários da geoestatística, principalmente por permitir o entendimento de uma provável tendência do comportamento estrutural de uma Variável Regionalizada. Por outro lado, a ampliação do uso de tais métodos ainda precisa ser bastante explorada, especialmente em estudos de demanda por transportes.

Verifica-se ainda, como relevante contribuição deste trabalho, a implementação de um método para representar valores de variáveis em diferentes suportes ou escala, com o objetivo de minimizar o conhecido problema da mudança de escala em análise espacial. No entanto, apesar de ter sido observada uma sutil diminuição do efeito pepita, especialmente para a variável sobre a escolha do automóvel, os semivariogramas ainda continuaram apresentando valores elevados deste parâmetro.

Apesar das relevantes contribuições metodológicas para o estudo de variáveis de transportes através da geoestatística, alguns obstáculos foram encontrados. O Quadro 8.1 resume os problemas detectados tanto em relação ao método, quanto a problemas inerentes a variáveis de demanda por transportes.

Quadro 8.1: Resumo dos entraves relacionados à tese

	Variáveis	Método
Problema verificado	<ul style="list-style-type: none"> ✓ Baixa dependência espacial; ✓ Efeito pepita continuou alto mesmo após a agregação das amostras. 	<ul style="list-style-type: none"> ✓ Tempo de execução do algoritmo; ✓ Estrutura espacial divergentes das identificadas pelos <i>varmaps</i>; ✓ Valores negativos de semivariância para Funções Objetivo menores.

Em relação à análise de mapas de semivariância, os resultados do AG foram de encontro às informações fornecidas pelos mapas de semivariância. Esperava-se que as direções angulares, melhor avaliadas pela função objetivo estabelecida no AG, estivessem próximas àquelas apresentadas pelos mapas de análise de variância (*varmap*).

No entanto, mesmo constatada uma frequência maior das direções entre 100°-115°, para as quatro variáveis analisadas pelo método de otimização, a análise da continuidade espacial, obtida pela análise convencional, indicou que a variável renda possui comportamento isotrópico, ou seja, varia uniformemente em todas as direções, sendo, portanto, omnidirecional. Já a variável referente à porcentagem de viagens realizadas por automóvel apresentou direção de continuidade espacial próximo dos 30°. Ainda assim, após a análise de medidas de desempenho e mapas estimados através dos modelos manual e otimizados, constatou-se que os resultados derivados da análise da validação cruzada são bastante próximos.

Outra grande vantagem conferida pelo método proposto é o conhecimento dos intervalos dos parâmetros utilizados para a modelagem, o que valida, de certa forma, a escolha de um modelo em relação a outro, já que determinado parâmetro recorrente atribuiu melhores resultados às estimativas.

Diante desta breve síntese dos resultados e problemas encontrados nesta pesquisa, a próxima subseção busca indicar possíveis sugestões para a continuidade de estudos relacionados à otimização de semivariogramas na análise de variáveis relacionadas ao planejamento de transportes e em relação a escolha do suporte de análise.

8.2.2. Recomendações para trabalhos futuros

Com base nos principais problemas e críticas pontuadas na subseção anterior, algumas sugestões precisam ser colocadas. Acredita-se que problemas relacionados à baixa dependência espacial e a alta variabilidade de dados de transportes podem ser minimizados através do uso de variáveis transformadas. Com isso, espera-se obter melhores modelos de semivariograma, desta forma, melhorando as estimativas.

Outra importante observação é em relação a problemas encontrados na execução do código. Em relação ao tempo de execução do AG, pode-se considerar que, definindo um critério de parada menor como, por exemplo, um menor número de gerações, a execução seria mais rápida, sem comprometer a qualidade dos resultados, já que na maioria dos casos, o menor erro já foi

obtido entre a 10^a e a 30^a geração. Com isso, observa-se que não houve necessidade de executar o algoritmo com o valor do critério de parada tão elevado. No entanto, neste estudo, o número de gerações para cada execução foi definido de forma empírica, para a observação de quantidade de gerações que proporcionasse a conversão para do melhor modelo.

Quanto aos valores negativos de semivariância explicada (contribuição do modelo), para uma possível melhor solução do algoritmo, há a possibilidade de introdução de algumas restrições em relação ao valor do parâmetro referente à contribuição do modelo. Este valor limitante poderia ser zero, o que representa total aleatoriedade da variável regionalizada, ou seja, representaria um semivariograma conhecido como efeito pepita puro. Esta alternativa impossibilita que o algoritmo encontre soluções “ótimas” com problemas de representação da estrutura espacial da Variável Regionalizada.

REFERÊNCIAS

- ABEDINI, M. J.; NASSERI, M.; BURN, D. H. The use of a genetic algorithm-based search strategy in geostatistics: application to a set of anisotropic piezometric head data. **Computers & Geosciences**, v. 41, p. 136–146, 2012.
- ADJEMIAN, M. K.; CYNTHIA LIN, C.-Y.; WILLIAMS, J. Estimating spatial interdependence in automobile type choice with survey data. **Transportation Research Part A: Policy and Practice**, v. 44, n. 9, p. 661–675, 2010.
- ALVES, V. F. B.; SILVA, A. N. R. DA; VAN DER WAERDEN, P. Técnicas exploratórias para localizar potenciais usuários de transporte público urbano. **Journal of Transport Literature**, v. 6, n. 3, p. 181–203, 2012.
- AFFENZELLER, M.; WAGNER, S.; WINKLER, S.; BEHAM, A. Genetic algorithms and genetic programming: modern concepts and practical applications. Chapman and Hall/CRC, 2009.
- ANSELIN, L. Local Indicators of Spatial Association-LISA. **Geographical Analysis**, v. 27, n. 2, p. 93–115, 1995.
- ANSELIN, L.; GETIS, A. Spatial statistical analysis and geographic information systems. **The Annals of Regional Science**, v. 26, n. 1, p. 19–33, 1992.
- BANDYOPADHYAY, S.; MAULIK, U. Genetic clustering for automatic evolution of clusters and application to image classification. **Pattern Recognition**, v. 35, n. 6, p. 1197–1208, 2002.
- BANZHAF, W.; NORDIN, P.; KELLER, R. E; FRANCONI, F. D. **Genetic programming: an introduction**. San Francisco: Morgan Kaufmann, 1998.
- BHAT, C.; ZHAO, H. The spatial analysis of activity stop generation. **Transportation Research Part B: Methodological**, v. 36, n. 6, p. 557–575, 2002.
- BLAINEY, S. Trip end models of local rail demand in England and Wales. **Journal of Transport Geography**, v. 18, n. 1, p. 153–165, 2010.
- BLAINEY, S.; MULLEY, C. Using Geographically Weighted Regression to forecast rail demand in the Sydney Region. In: **Australasian Transport Research Forum**, Brisbane 2013. Disponível em: <https://atrf.info/papers/2013/2013_blainey_mulley.pdf>. Acesso em: 18 jan. 2019.
- BRUTON, M. J. **Introdução ao planejamento dos transportes**. 2. ed. Rio de Janeiro: Editora Interciência, 1979.
- CHANG, J. S.; JUNG, D.; KIM, J.; KANG, T. Comparative analysis of trip generation models: results using home-based work trips in the Seoul Metropolitan Area. **Transportation Letters**, v. 6, n. 2, p. 78–88, 2014.
- CHILÈS, J. P.; DELFINER, P. **Geostatistics Modeling Spatial Uncertainty**. 2. ed. New Jersey: John Wiley & Sons, 2012.
- CHOI, J.; LEE, Y. J.; KIM, T.; SOHN, K. An analysis of Metro ridership at the station-to-station level in Seoul. **Transportation**, v. 39, n. 3, p. 705–722, 2012.

- CHU, X. Ridership models at the stop level. **World Transit Research**, 2004.
- CLARK, S. D. Estimating local car ownership models. **Journal of Transport Geography**, v. 15, n. 3, p. 184–197, 2007.
- COSTA, T. C.; FERNANDES FILHO, E. I. Aplicação da geoestatística para seleção de unidades de amostra independentes. **Embrapa Solos-Documents (INFOTECA-E)**, 2004.
- CRESSIE, N. Fitting variogram models by weighted least squares. **Journal of the International Association for Mathematical Geology**, v. 17, n. 5, p. 563–586, 1985.
- CRESSIE, N. Change of support and the modifiable areal unit problem. **Geographical Systems**, v. 3, n. 2–3, p. 159–180, 1996.
- CUI, Y. Q.; YONEDA, M.; SHIMADA, Y.; MATSUI, Y. Cost-Effective Strategy for the Investigation and Remediation of Polluted Soil Using Geostatistics and a Genetic Algorithm Approach. **Journal of Environmental Protection**, v. 07, n. 01, p. 99–115, 2016.
- CUNHA, A. L.; BESSA, J. E.; SETTI, J. R. Genetic Algorithm for the Calibration of Vehicle Performance Models of Microscopic Traffic Simulators. **BT - Progress in Artificial Intelligence**, v. 5816, p. 3–14, 2009.
- DE CASTRO, I. E.; CORRÊA, R. L.; GOMES, P. C. G. **Geografia, conceitos e temas**. Rio de Janeiro: Bertrand Brasil, 1995.
- DEUTSCH, C. V.; JOURNEL, A. G. Geostatistical software library and user's guide. **Oxford University Press, New York**, 1998.
- DIMATULAC, T.; MAOH, H. The spatial distribution of hybrid electric vehicles in a sprawled mid-size Canadian city: Evidence from Windsor, Canada. **Journal of Transport Geography**, v. 60, p. 59–67, 2017.
- DINNEEN, L. C.; BLAKESLEY, B. C. Algorithm AS 62: a generator for the sampling distribution of the Mann-Whitney U statistic. **Journal of the Royal Statistical Society. Series C (Applied Statistics)**, v. 22, n. 2, p. 269–273, 1973.
- DRUCK, S.; CARVALHO, M.S.; CÂMARA, G.; MONTEIRO, A. V. M. **Análise espacial de dados geográficos**. Brasília: Empraba Cerrados, 2004.
- EGAMI, C. Y.; MON-MA, M. L.; SETTI, J. R.; RILETT, L. R. **Automatic calibration of two-lane highway traffic simulation models using a genetic algorithm**. In: Applications of Advanced Technology in Transportation-Proceedings of the Ninth International Conference on Applications of Advanced Technology in Transportation (pp. 510-515), 2006.
- EOM, J. K.; PARK, M. S.; HEO, T.-Y.; HUNTSINGER, L. F. Improving the prediction of annual average daily traffic for nonfreeway facilities by applying a spatial statistical method. **Transportation Research Record: Journal of the Transportation Research Board**, v. 1, n. 1968, p. 20–29, 2006.
- FIELD, A. **Discovering statistics using IBM SPSS statistics: and sex and drugs and rock 'n' roll**. 4 ed. Los Angeles: Sage, 2013.
- FISCHER, M. M. Spatial Analysis in Geography. **International Encyclopedia of the Social & Behavioral Sciences**, p. 14752–14758, 2001.
- FOTHERINGHAM, S. Scale-independent spatial analysis. **Accuracy of spatial databases**, p. 221–228, Londres, 1989.

- FOTHERINGHAM, A. S.; WONG, D. W. S. The Modifiable Areal Unit Problem in Multivariate Statistical Analysis. **Environment and Planning A: Economy and Space**, v. 23, n. 7, p. 1025–1044, 2006.
- GELFAND, A. E.; ZHU, L.; CARLIN, B. P. On the change of support problem for spatio-temporal data. **Biostatistics**, v. 2, n. 1, p. 31–45, 2001.
- GOLDBERG, D. E. **Genetic algorithms in search, optimization, and machine learning**. Addison-Wesley Longman Publishing Co., Inc., 1989.
- GOLDBERG, D. E.; HOLLAND, J. H. Genetic Algorithms and Machine Learning. **Machine Learning**, v. 3, n. 2/3, p. 95–99, 1988.
- GOMES, V. A.; PITOMBO, C. S.; ROCHA, S. S.; SALGUEIRO, A. R. (2016). Kriging geostatistical methods for travel mode choice: a spatial data analysis to travel demand forecasting. **Open Journal of Statistics**, v. 6, n. 3, p. 514-527, 2016.
- GOOVAERTS, P. Geostatistical analysis of disease data: estimation of cancer mortality risk from empirical frequencies using Poisson kriging. **International journal of health geographics**, v. 4, p. 31, 2005.
- GOOVAERTS, P. Geostatistical analysis of disease data: accounting for spatial support and population density in the isopleth mapping of cancer mortality risk using area-to-point Poisson kriging. **International Journal of Health Geographics**, v. 5, p. 52, 2006.
- GOOVAERTS, P. Kriging and Semivariogram Deconvolution in the Presence of Irregular Geographical Units. **Mathematical Geology**, v. 40, n. 1, p. 101–128, 2008.
- GOOVAERTS, P. Combining Areal and Point Data in Geostatistical Interpolation: Applications to Soil Science and Medical Geography. **Mathematical Geosciences**, v. 42, n. 5, p. 535–554, 2010.
- GOOVAERTS, P. Applications to Soil Science and Medical Geography. **Mathematical Geosciences**, v. 42, n. 5, p. 535–554, 2011.
- GOOVAERTS, P. Geostatistical analysis of health data with different levels of spatial aggregation. **Spatial and Spatio-temporal Epidemiology**, v. 3, n. 1, p. 83–92, 2012.
- GOTWAY, C. A.; YOUNG, L. J. Combining Incompatible Spatial Data. **Journal of the American Statistical Association**, v. 97, n. 458, p. 632–648, 2002.
- HADAYEGHI, A.; SHALABY, A. S.; PERSAUD, B. Macrolevel Accident Prediction Models for Evaluating Safety of Urban Transportation Systems. **Transportation Research Record: Journal of the Transportation Research Board**, v. 1840, n. 1, p. 87–95, 2003.
- HAESER, G.; GOMES RUGGIERO, M. Aspectos Teóricos de Simulated Annealing e um Algoritmo duas Fases em Otimização Global. **TEMA - Tendências em Matemática Aplicada e Computacional**, v. 9, n. 3, p. 395–404, 2008.
- HAMAWAKI, C. D. L. **Geração automática de grade horária usando algoritmos genéticos: o caso da Faculdade de Engenharia Elétrica da UFU**. Universidade Federal de Uberlândia, 2005.
- HAUPT, R. L.; HAUPT, S. E. **Practical Genetic Algorithms**. 2. ed. New Jersey: Wiley-Interscience, 2004.
- HAYNES, R.; DARAS, K.; READING, R.; JONES, A. Modifiable neighbourhood units,

- zone design and residents' perceptions. **Health & Place**, v. 13, n. 4, p. 812–825, 2007.
- HOHN, M. **Geostatistics and petroleum geology**. Springer Science & Business Media, 1998.
- HOLLANDER, M.; WOLFE, D. A. **Nonparametric statistical methods**. 2. ed. New York: Wiley, 1999.
- HOLT, D.; STEEL, D. G.; TRANMER, M.; WRIGLEY, N. Aggregation and ecological effects in geographically based data. **Geographical analysis**, v. 28, n. 3, p. 244–261, 1996.
- IBGE – INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. cidades@. Disponível em: <<http://cidades.ibge.gov.br/xtras/home.php>>. Acesso em: 25 out. 2016.
- ISAAKS, E. H.; SRIVASTAVA, M. R. **Applied geostatistics**. 1989.
- JELINSKI, D. E.; WU, J. The modifiable areal unit problem and implications for landscape ecology. **Landscape Ecology**, v. 11, n. 3, p. 129–140, 1996.
- JIAN, X.; OLEA, R. A.; YU, Y.-S. Semivariogram modeling by weighted least squares. **Computers & Geosciences**, v. 22, n. 4, p. 387–397, 1996.
- JIAO, L.; LEI, H. **The application of improved genetic algorithm in fitting the spatial variogram**. Proceedings of 2011 International Conference on Computer Science and Network Technology. **Anais...**2011.
- JOSHUA, S. C.; GARBER, N. J. Estimating truck accident rate and involvements using linear and Poisson regression models. **Transportation Planning and Technology**, v. 15, n. 1, p. 41–58, 1990.
- JOURNEL, A. **Fundamentals of Geostatistics in Five Lessons**. v. 8. Washington, DC: American Geophysical Union, 1989.
- JOURNEL, A. G. Geostatistics: Models and tools for the earth sciences. **Mathematical Geology**, v. 18, n. 1, p. 119–140, 1986.
- JOURNEL, A. G.; HUIJBREGTS, C. J. **Mining geostatistics**. Londres: Academic press London, 1978.
- KINSER, J. M. **Python for bioinformatics**. Jones & Bartlett Publishers, 2010.
- LANDIM, P. M. B. Sobre Geoestatística e mapas. **Terra e Didática**, v. 2, n. 1, p. 19–33, 2006.
- LI, S.; WU, H.; WAN, D.; ZHU, J. An effective feature selection method for hyperspectral image classification based on genetic algorithm and support vector machine. **Knowledge-Based Systems**, v. 24, n. 1, p. 40–48, 2011.
- LI, S.; LU, W. **Automatic Fit of the Variogram**. Third International Conference on Information and Computing. **Anais...IEEE**, 2010. Disponível em: <<http://ieeexplore.ieee.org/document/5514035/>>. Acesso em: 19 dez. 2018
- LINDEN, R. **Algoritmos genéticos**. 2. ed. Brasport, 2008.
- LINDNER, A. **Análise desagregada de dados de demanda por transportes através de modelagem geoestatística e tradicional**. 106p. Dissertação de mestrado. Departamento de Engenharia de Transportes, Escola de Engenharia de São Carlos, Universidade de São Paulo, São Carlos 2015.

- LINDNER, A.; PITOMBO, C. S.; ROCHA, S. S.; QUINTANILHA, J. A. Estimation of transit trip production using Factorial Kriging with External Drift: an aggregated data case study. **Geo-Spatial Information Science**, v. 19, n. 4, 2016.
- LINDNER, A.; PITOMBO, C. S. A Conjoint Approach of Spatial Statistics and a Traditional Method for Travel Mode Choice Issues. **Journal of Geovisualization and Spatial Analysis**, v. 2, n. 1, p. 1, 2018.
- LOPES, S.; BRONDINO, N.; DA SILVA, A. GIS-Based Analytical Tools for Transport Planning: Spatial Regression Models for Transportation Demand Forecast. **ISPRS International Journal of Geo-Information**, v. 3, n. 2, p. 565–583, 2014.
- MAHER, M. J.; SUMMERSGILL, I. A comprehensive methodology for the fitting of predictive accident models. **Accident Analysis & Prevention**, v. 28, n. 3, p. 281–296, 1996.
- MARCEAU, D. J. The scale issue in social and natural sciences. **Canadian Journal of Remote Sensing**, v. 25, n. July, p. 347–356, 1999.
- MARTÍ, R.; REINELT, G. **The linear ordering problem: exact and heuristic methods in combinatorial optimization**. Springer Science & Business Media, v. 175, 2011.
- MASCHIO, C.; VIDAL, A. C.; SCHIOZER, D. J. A framework to integrate history matching and geostatistical modeling using genetic algorithm and direct search methods. **Journal of Petroleum Science and Engineering**, v. 63, n. 1–4, p. 34–42, 2008.
- MATHERON, G. La théorie des variables régionalisées et ses applications. **Les Cahiers de Morphologie Mathématique de Fontainebleau**, v. 5, p. 1–212, 1970.
- MCFADDEN, D. The measurement of urban travel demand. **Journal of public economics**, v. 3, n. 4, p. 303–328, 1974.
- MELLO, J.C. **Planejamento dos transportes**. São Paulo, McGraw-Hill do Brasil, 1975.
- METRÔ. COMPANHIA DO METROPOLITANO DE SÃO PAULO. **Pesquisa Origem e Destino 2007**. Disponível em: <<http://www.metro.sp.gov.br/>>. Acesso em: 10 ago. 2016.
- MIAOU, S. P. The relationship between truck accidents and geometric design of road sections: Poisson versus negative binomial regressions. **Accident Analysis & Prevention**, v. 26, n. 4, p. 471–482, 1994.
- MIAOU, S.-P.; LUM, H. Modeling vehicle accidents and highway geometric design relationships. **Accident Analysis & Prevention**, v. 25, n. 6, p. 689–709, 1993.
- MILLER, H. J. Potential contributions of spatial analysis to geographic information systems for transportation (GIS-T). **Wiley Online Library**, 1999.
- MIURA, H. A study of travel time prediction using universal kriging. **Top**, v. 18, n. 1, p. 257–270, 2010.
- NELDER, J. A.; WEDDERBURN, R. W. M. Generalized Linear Models. **Journal of the Royal Statistical Society: Series A (General)**, v. 135, n. 3, p. 370–384, 1972.
- NICHOLSON, A. Analysis of spatial distributions of accidents. **Safety Science**, v. 31, n. 1, p. 71–91, 1998.
- OLEA, R. A. **Geostatistics for Engineers and Earth Scientists**. New York: Springer US, 1999.

- OLIVER, M. A.; WEBSTER, R. A tutorial guide to geostatistics: Computing and modelling variograms and kriging. **CATENA**, v. 113, p. 56–69, 2014.
- OPENSHAW, S. An Empirical Study of Some Spatial Interaction Models. **Environment and Planning A: Economy and Space**, v. 8, n. 1, p. 23–41, 1976.
- OPENSHAW, S. The modifiable areal unit problem. **Concepts and techniques in modern geography**, 1984.
- ORTÚZAR, J.D.; WILLUMSEN, L.G. **Modelling Transport**. Wiley, 4. ed., 2011.
- PÁEZ, A.; SCOTT, D. M. Spatial statistics for urban analysis: a review of techniques with examples. **GeoJournal**, v. 61, n. 1, p. 53–67, 2005.
- PARENTEAU, M. P.; SAWADA, M. C. The modifiable areal unit problem (MAUP) in the relationship between exposure to NO₂ and respiratory health. **International Journal of Health Geographics**, v. 10, n. 58, p. 15, 2011.
- PEARL, J. **Heuristics: Intelligent search strategies for computer problem solving**. Addison-Wesley Pub. Co., Inc., Reading, Boston, 1984. Disponível em: <<https://www.osti.gov/biblio/5127296>>. Acesso em: 17 jan. 2019.
- PITOMBO, C. S.; SOUSA, A. J.; BIRKIN, M. **Comparing different spatial data analysis to forecast trip generation**. World Conference on Transport Research Society. **Anais...2010**.
- PITOMBO, C. S.; SALGUEIRO, A. R.; COSTA, A. S. G.; ISLER, C. A. A two-step method for mode choice estimation with socioeconomic and spatial information. **Spatial Statistics**, v. 11, p. 45–64, 2015.
- PITOMBO, C. S.; COSTA, A. S. G. DA; SALGUEIRO, A. R. Proposal of a sequential method for spatial interpolation of mode choice. **Boletim de Ciências Geodésicas**, v. 21, n. 2, p. 274–289, 2015.
- PITOMBO, C. S.; KAWAMOTO, E.; SOUSA, A. J. An exploratory analysis of relationships between socioeconomic, land use, activity participation variables and travel patterns. **Transport Policy**, v. 18, n. 3, p. 347–357, 2011.
- PYRCZ, M. J.; DEUTSCH, C. V. **Geostatistical reservoir modeling**. Oxford University Press, 2014.
- ROCHA, S. S.; LINDNER, A.; PITOMBO, C. S. Proposal of a geostatistical procedure for transportation planning field. **Boletim de Ciências Geodésicas**, v. 23, n. 4, 2017.
- ROCHA, S. S.; PITOMBO, C. S.; SALGUEIRO, A. R. Interpolação espacial de viagens urbanas por transporte coletivo através de amostra sistemática desagregada artificial. **Revista Brasileira de Cartografia**, v. 68, n. 4, 2016.
- RODRIGUES, F. L.; LEITE, H. G.; SANTOS, H. D. N.; SOUZA, A. L. D.; SILVA, G. F. D. Genetic algorithm metaheuristic to solve forest planning problem with integer constraints. **Revista Árvore**, v. 28, n. 2, p. 233–245, 2004.
- SMDU - SECRETARIA MUNICIPAL DE DESENVOLVIMENTO URBANO. **Mapa digital da cidade de São Paulo**. Disponível em: <http://geosampa.prefeitura.sp.gov.br/PaginasPublicas/_SBC.aspx>. Acesso em: 1 set. 2016.
- SELBY, B.; KOCKELMAN, K. M. Spatial prediction of traffic levels in unmeasured locations: applications of universal kriging and geographically weighted regression. **Journal**

of **Transport Geography**, v. 29, p. 24–32, 2013.

SHAFABAKHSH, G. A.; FAMILI, A.; BAHADORI, M. S. GIS-based spatial analysis of urban traffic accidents: Case study in Mashhad, Iran. **Journal of Traffic and Transportation Engineering (English Edition)**, v. 4, n. 3, p. 290–299, 2017.

SIVANANDAM, S. N.; DEEPA, S. N. **Introduction to genetic algorithms**. Springer Science & Business Media, 2007.

SOARES, A. **Geoestatística para as ciências da terra e do ambiente**. Lisboa: Instituto Superior Técnico, 2000.

SOUSA, D. C. B. D.; PITOMBO, C. S.; ROCHA, S. S.; SALGUEIRO, A. R.; DELGADO, J. P. M. Violence in public transportation: an approach based on spatial analysis. **Revista de Saúde Pública**, v. 51, p. 127, 2017.

THOMPSON, G. L. Achieving Suburban Transit Potential: Sacramento Revisited. **Transportation Research Record: Journal of the Transportation Research Board**, v. 1571, n. 1, p. 151–160, 1997.

TOBLER, W. R. A Computer Movie Simulating Urban Growth in the Detroit Region. **Economic Geography**, v. 46, p. 234, 1970.

TSEKERIS, T.; TSEKERIS, C. Demand forecasting in transport: Overview and modeling advances. **Economic research-Ekonomska istraživanja**, v. 24, n. 1, p. 82–94, 2011.

TSUTSUMI, M.; SEYA, H. Measuring the impact of large-scale transportation projects on land price using spatial statistical models. *Papers in Regional Science*, v. 87, n. 3, p. 385–401, 2008.

WACKERNAGEL, H. **Multivariate geostatistics: an introduction with applications**. 3. ed. Fontainebleau: Springer Science & Business Media, 2003.

WANG, Q. J. The Genetic Algorithm and Its Application to Calibrating Conceptual Rainfall-Runoff Models. **Water Resources Research**, v. 27, n. 9, p. 2467–2471, 1991.

WANG, Y.; ZHUANG, D.; LIU, H. Spatial Distribution of Floating Car Speed. **Journal of Transportation Systems Engineering and Information Technology**, v. 12, n. 1, p. 36–41, 2012.

WEBSTER, R.; MCBRATNEY, A. B. Mapping soil fertility at Broom's Barn by simple kriging. **Journal of the Science of Food and Agriculture**, v. 38, n. 2, p. 97–115, 1987.

WEBSTER, R.; MCBRATNEY, A. B. On the Akaike Information Criterion for choosing models for variograms of soil properties. **Journal of Soil Science**, v. 40, n. 3, p. 493–496, 1989.

WEBSTER, R.; OLIVER, M. A. **Geostatistics for environmental scientists**. John Wiley & Sons, 2007.

YAMAMOTO, J. K. **Avaliação e Classificação de Reservas Minerais**. São Paulo: Edusp, 2001.

YAMAMOTO, J. K.; LANDIM, P. M. B. **Geoestatística: conceitos e aplicações**. São Paulo: Oficina de Textos, 2013.

YONG, W.; ZHUANG, D.; HAODE, L. I. U. Spatial distribution of floating car speed.

Journal of Transportation Systems Engineering and Information Technology, v. 12, n. 1, p. 36–41, 2012.

ZHANG, D.; WANG, X. C. Transit ridership estimation with network Kriging: a case study of Second Avenue Subway, NYC. **Journal of Transport Geography**, v. 41, p. 107–115, 2014.

ZHAO, F.; PARK, N. Regression Models to Estimate Annual Average Daily Traffic. **Transportation Research**, p. 99–107, 2004.

APÊNDICE I

Código para escolha da grade

```
fun.AUTO.MW <- function(n){

  setwd("C:\\Users\\Samille\\Dropbox\\TESE 2018\\Variaveis por grade\\freq
acumulada\\função teste Mann whitney\\banco completo")
  dados.comp<-read.table("domicilios_completo_porcent.csv", h=T, sep=";")
  library(raster)
  library(sp)
  valores <- matrix(ncol=4, nrow=n)

  coordinates(dados.comp) = ~x+y
  for(b in 1:n){
    raster.auto <- raster(ext=extent(dados.comp), resolution = b*5)
    resol2<- res(raster.auto) #retorna o tamanho da grade
    rasOut.auto<-rasterize(dados.comp, raster.auto, dados.comp$AUTO, fun = mean)
    poly.auto<-rasterToPolygons(rasOut.auto)
    coord.pol <- coordinates(poly.auto)
    df.auto <- data.frame(coord.pol, poly.auto)
    ord.x.auto <- df.auto[order(df.auto$X1),] #ordenar dados através do valor da
coord x
    write.table(ord.x.auto, "poly_auto.csv", sep = ";")

    #test mann whitney

    { teste.MW <- wilcox.test(dados.comp$AUTO, ord.x.auto$layer, alternative =
"two.sided", paired = F, exact = NULL, conf.int = FALSE, conf.level = 0.95)
    p <- teste.MW$p.value
    w <- teste.MW$statistic
    decisao <- if(p >= 0.05) {
      print("aceita H0")

    } else {

      print("não aceita H0")
    }
  }

  valores[b,1] <- resol2[1]
  valores[b,2] <- p
  valores[b,3] <- w
  valores[b,4] <- decisao

  write.table(valores,print("testMW_MODAL_AUTO.txt")) #cria tabela com os
resultados
}

print(valores)
}
```

APÊNDICE II

Código para otimização do semivariograma experimental e teórico

```

using System;
using System.Collections.Generic;
using System.Linq;
using System.Text;
using System.Threading;
using System.Threading.Tasks;
using OfficeOpenXml;
using OfficeOpenXml.Drawing;
using OfficeOpenXml.Drawing.Chart;
using System.IO;
using Accord.Math;

namespace ag
{
    public static class geo
    {
        static int ndir,npontos,sim_np,npares; //numero de direções e de pontos
        static double sim_T0;
        public static double[] vdirecao = new double[2];
        public static double[] vtang = new double[2];
        public static double[] vlarg_banda = new double[2];
        public static double[] vlag = new double[2];
        public static double[,] mbest = new double[101, 5];
        public static double pop, taxac, taxam, taxae,dcorte;
        public static int geracao;

        //static int idir;//direção que está sendo calculada
        //static double[] mparametros;
        //matriz ndir x 7 , cada linha uma direção cada col um parametro (angulo, tol
angular,...)
        static double[] vretorno;
        static double[,] mpontos;//matriz que receberá todos os pontos (num de pontos x
3)
        static double[,] mregressao;
        static double[,] mpontosf,mvar,mvarf;
        //matriz 3D com os pontos filtrados após grid ( direção x npontos x 3)
        static double[] vlimite = new double[5];//xmin,ymin,xmax,ymax
        static string[] vtipo = new string[4];
        public static double largcorte;
        public static List<string> listasol = new List<string>();
        //vetor contendo os valores max e min para x e y (será usado na construção do
grid)
        static Random aleat = new Random((int)DateTime.Now.Ticks);

        public static void carregar()
        {
            ExcelPackage arquivoexcel = new ExcelPackage(new FileInfo(@"input.xlsx"));
            ExcelWorkbook planilha = arquivoexcel.Workbook;
            int naba = planilha.Worksheets.Count;
            ExcelWorksheet aba;
            aba = planilha.Worksheets[1];
            string st1 = "";
            st1 = aba.Cells[3, 6].Value.ToString();
            vdirecao[0] = Convert.ToDouble(st1);
            st1 = aba.Cells[3, 7].Value.ToString();
            vdirecao[1] = Convert.ToDouble(st1);
            st1 = aba.Cells[4, 6].Value.ToString();

```

```

vtang[0] = Convert.ToDouble(st1);
st1 = aba.Cells[4, 7].Value.ToString();
vtang[1] = Convert.ToDouble(st1);
st1 = aba.Cells[5, 6].Value.ToString();
vlarg_banda[0] = Convert.ToDouble(st1);
st1 = aba.Cells[5, 7].Value.ToString();
vlarg_banda[1] = Convert.ToDouble(st1);
st1 = aba.Cells[6, 6].Value.ToString();
vlag[0] = Convert.ToDouble(st1);
st1 = aba.Cells[6, 7].Value.ToString();
vlag[1] = Convert.ToDouble(st1);
pop = Convert.ToDouble(aba.Cells[2, 10].Value.ToString());
taxac = Convert.ToDouble(aba.Cells[3, 10].Value.ToString());
taxam = Convert.ToDouble(aba.Cells[4, 10].Value.ToString());
taxae = Convert.ToDouble(aba.Cells[5, 10].Value.ToString());
geracao = Convert.ToInt32(aba.Cells[6, 10].Value.ToString());

//largcorte = Convert.ToDouble(aba.Cells[13, 5].Value.ToString());

npares = Convert.ToInt32(aba.Cells[13, 5].Value.ToString());
st1 = aba.Cells[9, 6].Value.ToString();
sim_T0 = Convert.ToDouble(st1);
st1 = aba.Cells[10, 6].Value.ToString();
sim_np = Convert.ToInt32(st1);
mregressao = new double[4, 4];
vretorno = new double[5];
//captando os valores de entrada e passando para mparametros
vtipo[1] = "Esférico"; vtipo[2] = "Exponencial"; vtipo[3] = "Gaussiana";

//mparametros = new double[ndir+1,14];
//for (var i = 3; i <= 8; i++) for (var j=6;j<=6+ndir-1;j++)
//{
//    st1 = aba.Cells[i, j].Value.ToString();
//    mparametros[j - 5, i - 2] = Convert.ToDouble(st1);
//} ag

//verificando a ultima linha preenchida para saber quantidade de pontos
st1 = "1";
int count = 0;
st1 = (aba.Cells[1, 3].Value ?? "").ToString();
while (st1!="")
{
    count++;
    st1 = (aba.Cells[count, 1].Value ?? "").ToString();
}
npontos = count-2;
mpontos = new double[npontos + 1, 18];
mpontos[0, 0] = npontos;
//lendo os x,y e valores dos pontos + coordenadas da janela

for (var i = 2; i <= npontos+1; i++)
{
    st1 = aba.Cells[i, 1].Value.ToString();
    mpontos[i - 1, 1] = Convert.ToDouble(st1);
    st1 = aba.Cells[i, 2].Value.ToString();
    mpontos[i - 1, 2] = Convert.ToDouble(st1);
    st1 = aba.Cells[i, 3].Value.ToString();
    mpontos[i - 1, 3] = Convert.ToDouble(st1);
}
//dimensionando matriz 3D
//mpontosf = new double[ndir+1,npontos+1,18];
mvar = new double[ 1001, 5];
mvarf = new double[ 1001, 6];
//identificando maior x e menor x, maior y e menor y

double xmax, xmin,ymax,ymin;

```

```

    xmax = mpontos[1, 1];
    xmin = mpontos[1, 1];
    ymax = mpontos[1, 2];
    ymin = mpontos[1, 2];
    for (var i = 1; i <= npontos; i++)
    {
        if (xmax < mpontos[i, 1]) xmax = mpontos[i, 1];
        if (xmin > mpontos[i, 1]) xmin = mpontos[i, 1];
        if (ymax < mpontos[i, 2]) ymax = mpontos[i, 2];
        if (ymin > mpontos[i, 2]) ymin = mpontos[i, 2];
    }
    //passando para o vetor e dando uma folga de largura do grid dividido por 2
    largcorte = Math.Max((xmax - xmin), (ymax - ymin)) / 2;

    //vlimite[1] = xmin - mparametros[7] / 2;
    //vlimite[2] = ymin - mparametros[ 7] / 2;
    //vlimite[3] = xmax + mparametros[ 7] / 2;
    //vlimite[4] = ymax + mparametros[ 7] / 2;
    //ag

    Console.WriteLine("Quantidade de pontos: {0}", npontos);
    Console.WriteLine("Largura na horizontal: {0}", (xmax-xmin).ToString());
    Console.WriteLine("Largura na vertical: {0}", (ymax - ymin).ToString());
    Console.WriteLine("Distância de corte: {0}", largcorte);
    //Console.WriteLine("Digite Enter para iniciar o AG...");
    Console.WriteLine("-----");
    //Console.ReadKey();
}

public static double[] Fobj(double ang,double tolang, double lb, double lag)
{
    // DateTime t1, t2;
    // t1 = DateTime.Now;

    janela(ang, tolang, lb, largcorte);
    int m=variograma(lag,largcorte);
    if (m >= npares)
    {
        regressao();
        vretorno[0] = validacao();
    }
    else
    {
        vretorno[0] = 100;
    }

    //1-ang, 2 - tol ang , 3 - largura banda , 4 - dist corte , 5 - lag
    // t2 = DateTime.Now;
    //var duracao = (t2 - t1).TotalSeconds;

    //mparametros[8] = vaux[0]; //LOSS
    //mparametros[9] = vaux[1]; //C0
    //mparametros[10] = vaux[2]; //C
    //mparametros[11] = vaux[3]; //a
    //mparametros[12] = k; //tipo equação

    return vretorno;
}

public static void janela(double ang, double tolang, double LB,double dcorte)
{

```

```

//convertendo para radianos
ang = (ang / 180) * Math.PI; //Beta
double cosb = Math.Cos(ang);
double senb = Math.Sin(ang);
tolang = (tolang / 180) * Math.PI; //alpha
double sena = Math.Sin(tolang);
double senab = Math.Sin(ang + tolang); //sen(beta + alpha)
double cosab = Math.Cos(ang + tolang); // cos (beta + alpha)
double sena_b = Math.Sin(ang - tolang); //sen(beta - alpha)
double cosa_b = Math.Cos(ang - tolang); // cos (beta - alpha)
int np= Convert.ToInt32(mpontos[ 0, 0]); //pegando quant de pontos
double t;
//preenchendo os pontos da janela para cada ponto na próxima matriz mpontof,
a partir da coluna 4
for (var i = 1; i <= np; i++)
{
    //x1=x0+d*cos(beta)
    mpontos[ i, 4] = mpontos[ i, 1] + dcorte * cosb;
    //y1=x0+d*sen(beta)
    mpontos[ i, 5] = mpontos[ i, 2] + dcorte * senb;
    //x2=x1+LB*sen(beta)
    mpontos[ i, 6] = mpontos[ i, 4] + LB*senb;
    //y2=y1-LB*cos(beta)
    mpontos[ i, 7] = mpontos[ i, 5] - LB * cosb;
    //x3=x1-LB*sen(beta)
    mpontos[ i, 8] = mpontos[ i, 4] - LB * senb;
    //y3=y1+LB*cos(beta)
    mpontos[ i, 9] = mpontos[ i, 5] + LB * cosb;
    // t para o ponto 4
    t = ((mpontos[ i, 9] - mpontos[ i, 2]) * cosb - (mpontos[ i, 8] -
mpontos[ i, 1]) * senb) / sena;
    //x4=x0+t*cos(beta+alpha)
    mpontos[ i, 10] = mpontos[ i, 1] + t * cosab;
    //y4=y0+t*sen(beta+alpha)
    mpontos[ i, 11] = mpontos[ i, 2] + t * senab;
    // t para o ponto 5
    t = ((mpontos[ i, 6] - mpontos[ i, 1]) * senb - (mpontos[ i, 7] -
mpontos[ i, 2]) * cosb) / sena;
    //x5=x0+t*cos(beta-alpha)
    mpontos[ i, 12] = mpontos[ i, 1] + t * cosa_b;
    //y5=y0+t*sen(beta-alpha)
    mpontos[ i, 13] = mpontos[ i, 2] + t * sena_b;
}
}

public static int variograma(double lag, double lcorte)
{
    List<double> lx = new List<double>();
    List<double> ly = new List<double>();
    List<double> lz = new List<double>();
    double dist,soma,ang1, ang2, ang3, ang4,
ang5,x0,y0,z0,x1,y1,x2,y2,cosang,aux1,aux2;
    int np = Convert.ToInt32(mpontos[ 0, 0]);
    int nh = Convert.ToInt32(Math.Round(largcorte / lag,0))+1;

    //preenchendo as distancias na primeira coluna da matriz mvar
    mvar[ 1, 1] = 0;
    for (var i = 2; i <= nh; i++)
        mvar[ i, 1] = (i-1) * lag;
    // zerando toda a coluna 2 da matriz mvar, pois irei somar os [z(x) -
z(x+h)]^2
    for (var i = 1; i <= 1000; i++) mvar[ i, 2] = 0;
    // zerando toda a coluna 3 da matriz mvar, pois irei somar a quantidade de
pares
    for (var i = 1; i <= 1000; i++) mvar[ i, 3] = 0;

```

```

for (var i = 1; i <= np; i++)
{
    //Console.WriteLine(i);
    //deixando a coluna 14 todas com zero para iniciar a busca
    //for (var j = 1; j <= np; j++) mpontos[ j, 14] = 0;
    //limpando a lista de coordenadas x e y
    lx.Clear();
    ly.Clear();
    lz.Clear();
    for (var k = 1; k <= np; k++) if (k != i)
    {
        //ponto a ser analisado
        x0 = mpontos[ k, 1];
        y0 = mpontos[ k, 2];
        z0 = mpontos[ k, 3];
        //P3 e P2
        x1 = mpontos[ i, 8];
        y1 = mpontos[ i, 9];
        x2 = mpontos[ i, 6];
        y2 = mpontos[ i, 7];
        aux1 = (x1 - x0) * (x2 - x0) + (y1 - y0) * (y2 - y0);
        aux2 = Math.Sqrt(Math.Pow(x1 - x0, 2) + Math.Pow(y1 - y0, 2)) *
Math.Sqrt(Math.Pow(x2 - x0, 2) + Math.Pow(y2 - y0, 2));
        cosang = aux1 / aux2;

        ang1 = Math.Acos(cosang);
        //P2 e P5
        x1 = mpontos[ i, 6];
        y1 = mpontos[ i, 7];
        x2 = mpontos[ i, 12];
        y2 = mpontos[ i, 13];
        aux1 = (x1 - x0) * (x2 - x0) + (y1 - y0) * (y2 - y0);
        aux2 = Math.Sqrt(Math.Pow(x1 - x0, 2) + Math.Pow(y1 - y0, 2)) *
Math.Sqrt(Math.Pow(x2 - x0, 2) + Math.Pow(y2 - y0, 2));
        cosang = aux1 / aux2;
        ang2 = Math.Acos(cosang);
        //P5 e P0
        x1 = mpontos[ i, 12];
        y1 = mpontos[ i, 13];
        x2 = mpontos[ i, 1];
        y2 = mpontos[ i, 2];
        aux1 = (x1 - x0) * (x2 - x0) + (y1 - y0) * (y2 - y0);
        aux2 = Math.Sqrt(Math.Pow(x1 - x0, 2) + Math.Pow(y1 - y0, 2)) *
Math.Sqrt(Math.Pow(x2 - x0, 2) + Math.Pow(y2 - y0, 2));
        cosang = aux1 / aux2;
        ang3 = Math.Acos(cosang);
        //P0 e P4
        x1 = mpontos[ i, 1];
        y1 = mpontos[ i, 2];
        x2 = mpontos[ i, 10];
        y2 = mpontos[ i, 11];
        aux1 = (x1 - x0) * (x2 - x0) + (y1 - y0) * (y2 - y0);
        aux2 = Math.Sqrt(Math.Pow(x1 - x0, 2) + Math.Pow(y1 - y0, 2)) *
Math.Sqrt(Math.Pow(x2 - x0, 2) + Math.Pow(y2 - y0, 2));
        cosang = aux1 / aux2;
        ang4 = Math.Acos(cosang);
        //P4 e P3
        x1 = mpontos[ i, 10];
        y1 = mpontos[ i, 11];
        x2 = mpontos[ i, 8];
        y2 = mpontos[ i, 9];
        aux1 = (x1 - x0) * (x2 - x0) + (y1 - y0) * (y2 - y0);
        aux2 = Math.Sqrt(Math.Pow(x1 - x0, 2) + Math.Pow(y1 - y0, 2)) *
Math.Sqrt(Math.Pow(x2 - x0, 2) + Math.Pow(y2 - y0, 2));
        cosang = aux1 / aux2;
        ang5 = Math.Acos(cosang);
    }
}

```

```

// se o ponto tiver dentro da janela adicionando em uma lista
soma = ang1 + ang2 + ang3 + ang4 + ang5;
if (Math.Round(soma/(2*Math.PI), 2) > 0.99)
{
    lx.Add(x0);
    ly.Add(y0);
    lz.Add(z0);
}
}
if (lx.Count>0) for (var k = 0; k <= lx.Count-1; k++)
{
    aux1 = Math.Pow(lx[k] - mpontos[ i, 1],2);
    aux2 = Math.Pow(ly[k] - mpontos[ i, 2],2);
    dist = Math.Sqrt(aux1 + aux2);
    mvar[0, 0] = nh;
    //primeiro lag
    if (dist<lag/2)
    {
        aux1 = mpontos[ i, 3] - lz[k];
        aux2 = Math.Pow(aux1, 2);
        //somando para cada distância
        mvar[1, 2] = mvar[ 1, 2] + aux2;
        mvar[ 1, 3]++;
    }
    //lags intermediarios
    for (var j = 2; j <= nh-1; j++)
        if (dist>mvar[j,1]-lag/2 && dist<mvar[j,1]+lag/2)
        {
            aux1 = mpontos[i,3] - lz[k];
            aux2 = Math.Pow(aux1, 2);
            //somando para cada distância
            mvar[ j, 2] = mvar[ j, 2] + aux2;
            mvar[ j, 3]++;
        }
    //ultimo lag
    if (dist > lcorde-lag / 2)
    {
        aux1 = mpontos[ i, 3] - lz[k];
        aux2 = Math.Pow(aux1, 2);
        //somando para cada distância
        mvar[ nh, 2] = mvar[ nh, 2] + aux2;
        mvar[ nh, 3]++;
    }
}
}
//preenchendo o y(h) na coluna 4

int m = 0;
for (var i = 1; i <= nh; i++) if (mvar[ i, 3] > 0)
{
    m++;
    mvar[ i, 4] = mvar[ i, 2] / (2 * mvar[ i, 3]);
    mvarf[ m, 1] = mvar[ i, 1];
    mvarf[ m, 2] = mvar[ i, 2];
    mvarf[ m, 3] = mvar[ i, 3];
    mvarf[ m, 4] = mvar[ i, 4];
}
mvarf[ 0, 0] = m;
return m;
//if (m==0)
//{
//    Console.WriteLine("m=0");
//    Console.ReadKey();
//}

```

```

    }

    public static void output()
    {
        ExcelPackage arquivoexcel = new ExcelPackage(new FileInfo(@"input.xlsx"));
        ExcelWorkbook planilha = arquivoexcel.Workbook;
        //int naba = planilha.Worksheets.Count;
        planilha.Worksheets.Add("soluções_viaveis");
        // planilha.Worksheets.MoveAfter(nomeaba, "input");
        ExcelWorksheet aba;
        aba = planilha.Worksheets["soluções_viaveis"];
        int linhai, inicio, fim;
        string st1;
        linhai = 1;
        aba.Cells[linhai,1].Value = "Fobj";
        aba.Cells[linhai, 2].Value = "Direção";
        aba.Cells[linhai, 3].Value = "Tol. Ang";
        aba.Cells[linhai, 4].Value = "Larg Banda";
        aba.Cells[linhai, 5].Value = "Lag";
        aba.Cells[linhai, 6].Value = "C0";
        aba.Cells[linhai, 7].Value = "C";
        aba.Cells[linhai, 8].Value = "a";
        aba.Cells[linhai, 9].Value = "Equação";

        foreach (string st in listasol)
        {
            linhai++;
            string[] vst = st.Split(' ');
            for (int i=0;i<=7;i++) aba.Cells[linhai, i + 1].Value =
Math.Round(Convert.ToDouble(vst[i]),4);
            aba.Cells[linhai, 9].Value = vtipo[Convert.ToInt32(vst[8])];
        }
        arquivoexcel.Save();
    }

    public static double[] simulated_esferico(double t0, int nite, double[] param,
double[] x, double[] y, double[] vz)
    {
        double a0, a1;
        a0 = param[0];
        a1 = param[1];

        double[] vp = new double[2];
        vp[0] = a0;
        vp[1] = a1;
        return vp;
    }

    public static double funcoes(int tipo, double C0, double C, double a, double h)
    {
        double y=0;
        //esferico
        if (tipo == 1 && h>a) y = C0 + C;
        if (tipo ==1 && h <= a) y = C0 + C * (1.5 * (h / a) - 0.5 * Math.Pow(h / a,
3));
        if (tipo == 2) y = C0 + C * (1 - Math.Exp((-3 * h )/ a));
        if (tipo == 3) y = C0 + C * (1 - Math.Exp((-3 * h * h) / (a * a)));
        return y;
    }

    public static double loss(int tipo, double[] vx, double[] vy, double[] vz, double

```



```

C0, double C, double a)
{
    double soma,aux,ycalc,yobs;
    int n = vx.Length-1;

    soma = 0;
    for (var i = 1; i <= n; i++)
    {
        ycalc = funcoes(tipo, C0, C, a, vx[i]);
        yobs = vy[i];
        aux = Math.Pow(yobs - ycalc, 2);
        soma = soma + vz[i] * aux;
    }

    return soma;
}

public static void simulated(int tipo, int np, double T0)
{
    double C01, C02, C1, C2, a1, a2, f1, f2,aux1,aux2;
    double C0f=0, Cf=0, af=0, ff=0;
    int nh = Convert.ToInt32(mvarf[ 0, 0]);
    double[] vx = new double[nh+1];
    double[] vy = new double[nh + 1];
    double[] vz = new double[nh + 1];

    for (var i = 1; i <= nh; i++)
    {
        vx[i] = mvarf[ i, 1];
        vz[i] = mvarf[ i, 3];
        vy[i] = mvarf[ i, 4];
    }

    //inicializando parametros
    //inicio:
    aleat = new Random((int)DateTime.Now.Ticks);

    double T = T0;
    aux1=vy[1]-(vx[1]/(vx[2]-vx[1]))*(vy[2]-vy[1]);
    C01 = Math.Max(0,aux1);

    a1 = vx[nh] / 2;
    aux1 = (vy[nh - 2] + vy[nh - 1] + vy[nh]) / 3 - C01;
    if (aux1 < 0) aux1 = vy[nh] - C01;
    C1 = Math.Max(1, aux1);

    while (T > 0.1)
    {
        for (var i = 1; i <= np; i++)
        {
            //gerando vizinho (perturbação) até 10% para mais ou para menos
            C02 = (0.9 + aleat.NextDouble() / 5) * C01;
            C2 = (0.9 + aleat.NextDouble() / 5) * C1;
            a2 = (0.9 + aleat.NextDouble() / 5) * a1;
            f1 = loss(tipo, vx, vy, vz, C01, C1, a1);

            f2 = loss(tipo, vx, vy, vz, C02, C2, a2);
            if (f2 < f1)
            {
                C01 = C02;
                C1 = C2;
                a1 = a2;
                // C0f = C01;Cf = C1;af = a1;
            }
            if (f2 >= f1 && aleat.NextDouble() < Math.Exp((f1 - f2) / T))

```

```

        {
            C01 = C02;
            C1 = C2;
            a1 = a2;
        }
    }
    T = 0.95 * T;
}

//caso os parâmetros fiquem iguais a zero , repetir procedimento...

f1 = loss(tipo, vx, vy, vz, C01, C1, a1);
//if (f1 < 0.01)
//{
//    Console.WriteLine(" erro regressao...");
//    goto inicio;
//}
mregressao[tipo, 0] = f1;
mregressao[tipo, 1] = C01;
mregressao[tipo, 2] = C1;
mregressao[tipo, 3] = a1;
}

public static void regressao()
{
    double[] vs;
    double[] vaux = new double[4];
    double loss, C0, C, a,min;
    double db1 = 1000000000000;
    double aux;

    simulated(1, sim_np, sim_T0);
    simulated(2, sim_np, sim_T0);
    simulated(3, sim_np, sim_T0);

    min = mregressao[1, 0];
    int k = 1;
    for (var i = 1; i <= 3; i++) if (mregressao[i,0]<min)
    {
        min = mregressao[i, 0];
        k = i;
        // for (var j = 0; j <= 3; j++) vaux[j] = mregressao[i, j];
    }

    //mparametros[ 8] = vaux[0]; //LOSS
    //mparametros[ 9] = vaux[1]; //C0
    //mparametros[ 10] = vaux[2]; //C
    //mparametros[ 11] = vaux[3]; //a
    //mparametros[ 12] = k; //tipo equação

    for (var j = 1; j <= 3; j++) vretorno[j] = mregressao[k, j];
    //vretorno[1] = vaux[1]; vretorno[2] = vaux[2]; vretorno[3] = vaux[3];
    vretorno[4] = k;

    int m = Convert.ToInt32(mvarf[ 0, 0]);
    //preenchendo a 5a coluna da matriz com ycalc pela equação
    for (var i = 1; i <= m; i++)
        mvarf[i,5] = funcoes(k, vaux[1], vaux[2], vaux[3], mvarf[ i, 1]);
}

```

```

}

public static double[,] pt_prox(int index_p0,int np)
{
    int p1, p2, p3, p4,k;
    double ang, tang;
    //double x1, y1, x2, y2,xq1,yq1, xq2, yq2, xq3, yq3, xq4, yq4,ang,tgang;
    double[] x = new double[5];
    double[] y = new double[5];
    double[] z = new double[5];
    double[] d = new double[5];
    //for (var i = 0; i <= 4; i++) d[i] = mparametros[ 4]; //larg corte
    for (var i = 0; i <= 4; i++) d[i] = largcorte;
    x[0] = mpontos[ index_p0, 1];
    y[0] = mpontos[ index_p0, 2];
    z[0] = mpontos[ index_p0, 3];

    int[] vc = new int[5];
    for (var i = 0; i <= 3; i++) vc[i]=0;
    for (var i=1;i<=np;i++) if (i != index_p0)
    {

        double x_analisado = mpontos[ i, 1];
        double y_analisado = mpontos[ i, 2];
        double z_analisado= mpontos[ i, 3];

        //ang = Math.Atan2(y_analisado - y[0], y_analisado - x[0]);
        //ang = ang * (180 / Math.PI);
        //if (ang < 0) ang = 360 + ang;
        d[0] = Math.Sqrt(Math.Pow(x_analisado - x[0], 2) +
Math.Pow(y_analisado - y[0], 2));
        //if (ang>=0 && ang<90) if (d[0]<d[1])
        if (x_analisado > x[0] && y_analisado>y[0]) //quadrante 1
        {
            vc[1]++;
            d[1] = d[0];
            x[1] = x_analisado;
            y[1] = y_analisado;
            z[1] = z_analisado;
        }
        //if (ang > 90 && ang <= 180) if (d[0] < d[2]) //quadrante 2
        if (x_analisado < x[0] && y_analisado > y[0])
        {
            vc[2]++;
            d[2] = d[0];
            x[2] = x_analisado;
            y[2] = y_analisado;
            z[2] = z_analisado;
        }
        //if (ang > 180 && ang <= 270) if (d[0] < d[3])
        if (x_analisado < x[0] && y_analisado < y[0])
        {
            vc[3]++;
            d[3] = d[0];
            x[3] = x_analisado;
            y[3] = y_analisado;
            z[3] = z_analisado;
        }
        //if (ang > 270 && ang < 360) if (d[0] < d[4])
        if (x_analisado > x[0] && y_analisado < y[0])
        {
            vc[4]++;
            d[4] = d[0];
            x[4] = x_analisado;
            y[4] = y_analisado;
            z[4] = z_analisado;
        }
    }
}

```

```

    }

}

int n = 0;
for (var i = 1; i <= 4; i++) if (vc[i] == 0) n++;
n = 4 - n;
double[,] mreturn = new double[n+1, 4];
mreturn[0, 0] = n;

mreturn[0, 1] = x[0];
mreturn[0, 2] = y[0];
mreturn[0, 3] = z[0];
k = 0;
for (var i = 1; i <=4; i++) if (vc[i]!=0)
{
    k++;
    mreturn[k, 1] = x[i];
    mreturn[k, 2] = y[i];
    mreturn[k, 3] = z[i];
}
return mreturn;
}

public static double[] mcovarianca(double[,] mpts,int np)
{
    double C0, C, a,h,d;
    int tipoeq;
    //C0 = mparametros[ 9];
    //C= mparametros[ 10];
    //a= mparametros[ 11];
    //tipoeq = Convert.ToInt32(mparametros[ 12]);
    C0 = vretorno[1];
    C = vretorno[2];
    a = vretorno[3];
    tipoeq = Convert.ToInt32(vretorno[4]);
    double[,] meuc = new double[np + 1, np + 1];
    //preenchendo matriz euclidiana a partir da matriz mpts
    for (var i = 1; i <= np; i++) for (var j = 1; j <= np; j++)
        meuc[i,j] = Math.Sqrt(Math.Pow(mpts[i, 1] - mpts[j, 1], 2) +
Math.Pow(mpts[i, 2] - mpts[j, 2], 2));
    double[] veuc = new double[np + 1];
    //preenchendo a coluna n+1 contendo a distancia euclidiana entre os dados e o
ponto a estimar
    for (var i = 1; i <= np; i++)
        veuc[i] = Math.Sqrt(Math.Pow(mpts[i, 1] - mpts[0, 1], 2) +
Math.Pow(mpts[i, 2] - mpts[0, 2], 2));

    double[,] mcov = new double[np + 2, np + 2];
    //double[,] mcov = new double[6,6];
    for (var i = 1; i <= np; i++) for (var j = 1; j <= np; j++)
    {
        h = meuc[i, j];
        mcov[i, j] = funcoes(tipoeq, C0, C, a, h);
    }
    for (var i = 1; i <= np; i++) mcov[np+1, i] = 1;
    for (var i = 1; i <= np; i++) mcov[i, np+1] = 1;
    mcov[np + 1, np + 1] = 0;
    double[] vcov = new double[np + 1];
    for (var i = 1; i <= np; i++) vcov[i] = funcoes(tipoeq, C0, C, a, veuc[i]);

    double[,] mk = new double[np+1, np+1];
    for (var i = 0; i <= np; i++) for (var j = 0; j <= np; j++)
        mk[i, j] = mcov[i+1, j+1];

    double[] vm = new double[np+1];

```

```

        for (var i = 0; i <= np-1; i++) vm[i] = vcov[i + 1];
        vm[np] = 1;
        var x = mk.Solve(vm);
        double zreal = mpts[0,3];
        double soma = 0;
        for (var i = 1; i <= np; i++) soma = soma + x[i - 1] * mpts[i,3];
        double zestimado = soma;
        double erro = Math.Abs(zestimado - zreal);
        double[] vreturn = new double[3];
        vreturn[0] = zreal;
        vreturn[1] = zestimado;
        vreturn[2] = erro;
        return vreturn;
    }

    public static double validacao()
    {
        int np = Convert.ToInt32(mpontos[ 0, 0]);
        double soma = 0;
        for (var i = 1; i <= np; i++)
        {
            double[,] m_pontosp = pt_prox(i,np);
            int k = Convert.ToInt32(m_pontosp[0, 0]);
            double[] vz = mcovarianca(m_pontosp, k);
            mpontos[ i, 15] = vz[0];
            mpontos[ i, 16] = vz[1];
            mpontos[ i, 17] = vz[2];
            soma = soma + vz[2];
        }
        return soma/np;
    }

}
}

```

APÊNDICE III

Tutorial de utilização do AG utilizado neste trabalho

a) Acesse o link a seguir e baixe os arquivos:

https://drive.google.com/drive/folders/1IrLYHag_gscRhmVmb6U2KgBB3Pp6Hgm_?usp=sharing

b)

Os limites dos parâmetros do semivariograma experimental podem ser determinados da seguinte forma:

Parâmetros semivariograma experimental					
Limites	Direção	Largura da banda	Tolerância angular	nº pares de pontos	Tamanho <i>lag</i> (h)
Min.	0°	-	1°	30	Menor distância entre as amostras
Max.	180°	-	90° (> 90° = omnidirecional)	-	Menor ou igual à distância de corte

- c) Depois de inserir a variável com respectivas coordenadas planas e definir os parâmetros, de acordo o quadro acima, salvar o arquivo “*input*” e, em seguida, executar o arquivo **ag.exe** (passo 2). Aparecerá uma tela, como mostrado a seguir:

```

C:\Users\Samille\Dropbox\TESE 2018\AG auto grade 85\ag.exe
Quantidade de pontos: 4320
Largura na horizontal: 15385
Largura na vertical: 18615
Distância de corte: 9307,5
-----
INÍCIO DO AG
-----

```

- d) Quando o algoritmo for finalizado, na mesma tela irá apresentar os seguintes resultados:

- o menor valor da função objetivo em cada geração e;
- a duração acumulada do término cada geração.

```

Quantidade de pontos: 4320
Largura na horizontal: 15385
Largura na vertical: 18615
Distância de corte: 9307.5
-----
INÍCIO DO AG
-----
Geração:1 - Fobj: 25.1735/100 - Duração Acumulada: 2.21 h
Geração:2 - Fobj: 24.1306/100 - Duração Acumulada: 3.42 h
Geração:3 - Fobj: 24.1306/100 - Duração Acumulada: 4.54 h
Geração:4 - Fobj: 24.1306/100 - Duração Acumulada: 5.64 h
Geração:5 - Fobj: 24.1306/100 - Duração Acumulada: 6.67 h
Geração:6 - Fobj: 24.1306/100 - Duração Acumulada: 7.74 h
Geração:7 - Fobj: 24.1306/100 - Duração Acumulada: 8.85 h
Geração:8 - Fobj: 24.003/100 - Duração Acumulada: 9.88 h
Geração:9 - Fobj: 24.003/100 - Duração Acumulada: 10.94 h
Geração:10 - Fobj: 24.003/100 - Duração Acumulada: 12.02 h
Algoritmo FInalizado... Aperte o ENTER

```

- e) Ao finalizar a execução do AG, no mesmo arquivo “input” utilizado na etapa 2, conterà uma aba incluindo os parâmetros do semivariograma e o valor *fitness* de cada indivíduo gerado. Com isso, é possível escolher o indivíduo que obteve menor erro entre valores estimados e observados, na etapa de krigagem.
- f) Concluída estas etapas, o código fornece inúmeros semivariogramas com seus respectivos parâmetros para cálculo e ajuste do semivariograma experimental em um *software* geoestatístico da escolha do usuário.

APÊNDICE IV

Dispersão entre valores observados x estimados (krigagem)

APÊNDICE V

Distribuição de frequência dos erros da estimação

Viagens por automóvel (%)

Viagens por auto no domicílio (AG)

Viagens por auto no domicílio (ajuste manual)

Viagens por auto 'grade 85m' (AG)

Viagens por auto 'grade 85m' (ajuste manual)

APÊNDICE VI

Distribuição de frequência dos erros da estimação

Renda familiar (SM)