ANÁLISE ESPACIAL EXPLORATÓRIA E CONFIRMATÓRIA BASEADA EM PRECEITOS DE GEOESTATÍSTICA

Lucas Assirati
Cláudia Cristina Baptista Ramos Naizer
Cira Souza Pitombo
Universidade de São Paulo
Escola de Engenharia de São Carlos
Departamento de Engenharia de Transportes

RESUMO
Dinâmicas urbanas podem ser caracterizadas mais efetivamente ao se considerar aspectos espaciais nos estudos. Este trabalho, utilizando uma cidade fictícia que simula deslocamentos por meio de automóveis ou transporte coletivo, atesta a existência de autocorrelação espacial do conjunto de dados por meio de dois indicadores: Moran e SivarG (*Global Spatial Indicator Based on Variogram*). A corroboração da dependência espacial, apontada pelos indicadores globais, é confirmada por meio de dois modelos de escolha discreta. Um contendo apenas variáveis originais da base de dados. Outro, análogo ao primeiro, porém adicionado de covariáveis regionais obtidas por preceitos da geoestatística. Um aumento de 15% das taxas de acerto em validações cruzadas é alcançado quando inclui-se variáveis regionais.

ABSTRACT
Urban dynamics can be characterized more effectively by considering spatial aspects in studies. This work, using a fictitious city that simulates urban traffic by cars or transit, attests to the existence of spatial autocorrelation of the data set by two indicators: Moran and SivarG (*Global Spatial Indicator Based on Variogram*). Corroboration of spatial dependence, pointed by the global indicators, is confirmed by two discrete choice models. The first one includes just the original database variables. The second one, is analogous to the first, but added of regional covariates obtained by geostatistical concepts. A 15% increase in cross-validations hit rates is achieved when regional variables are included.

1. INTRODUÇÃO

Modelos de escolha discreta são uma consagrada ferramenta, amplamente utilizada em estudos da demanda por transportes (Hess, 2005; Bhat et al., 2008; Ahern e Tapley, 2008; Qin et al., 2017). Pode-se citar como exemplo de aplicação os modelos de escolha modal, pois nestes, deve-se estimar uma alternativa discreta de um conjunto de possibilidades através de atributos das alternativas modais e dos indivíduos.

No entanto, para o caso de demanda por transportes, vale ressaltar que, na prática, é muito difícil conhecer todos os fatores que afetam as decisões individuais, uma vez que existe uma heterogeneidade subjacente ao comportamento dos indivíduos. Além disso, as localizações domiciliares e das atividades de destino podem influenciar as escolhas modais, bem como os atributos individuais e das alternativas. Desta forma, muitos trabalhos (Pitombo et al., 2015; Zhou, 2012; Páez et al., 2013; Miyamoto et al., 2004) passaram a incorporar variáveis relacionadas à localização geográfica aos estudos de previsão de demanda por transportes, promovendo aprimoramento das estimativas realizadas por modelos de escolha discreta.

Contudo, o uso de variáveis regionalizadas para incremento de modelos para previsão de demanda por transportes, requer uma investigação inicial acerca da dependência espacial desses dados. Uma forma de avaliação da associação espacial se dá por meio de indicadores que se destinam a averiguar a existência de tendências espaciais das informações estudadas. Confirmando-se eventuais associações, tem-se a garantia que tais dados podem compor modelos espaciais confirmatórios para previsão de viagens (Anselin, 1995; Getis e Ord, 1992; Naizer, 2018).

Valendo-se dos indicadores espaciais Moran (Moran, 1950; Anselin, 1995) e SivarG (Naizer, 2018) este trabalho, aplicado a um conjunto de dados sintéticos, busca atestar de maneira sistemática as regiões geográficas do conjunto de dados cujas informações apresentem associação espacial – Análise espacial exploratória. A seguir, promove-se a comparação de dois modelos de escolha discreta. O primeiro apenas com covariáveis tradicionais e o segundo com a incorporação de covariáveis espaciais, propostas a partir de pressupostos da geoestatística – Análise espacial confirmatória.

Este artigo apresenta 5 seções, além desta introdução. A Seção 2 traz uma descrição sobre análise espacial exploratória, indicadores de associação espacial e análise espacial confirmatória. A Seção 3 apresenta uma sucinta definição de modelos de escolha discreta. A Seção 4 traz a descrição dos materiais e do procedimento metodológico adotado. A Seção 5 traz os resultados e discussões. Finalmente, a Seção 6 apresenta as principais conclusões.

2. ANÁLISE ESPACIAL EXPLORATÓRIA E CONFIRMATÓRIA

Análise espacial consiste em um conjunto de técnicas que se destinam a analisar a localização geográfica de determinadas observações, avaliando as possíveis relações entre os valores das variáveis e diferentes localizações. Diversas questões podem ser respondidas mediante análises espaciais. A análise espacial de um fenômeno compreende uma etapa preliminar, exploratória, na qual é visualizado ou mensurado o grau de dependência espacial, de forma local ou global, de um conjunto de dados. Corroborada a dependência espacial dos dados, é indicada a etapa de análise espacial confirmatória, na qual modelos que incluem a estrutura espacial dos dados são utilizados.

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

O indicador global SivarG (*Global Spatial Indicator Based on Variogram*) é baseado em pressupostos da geoestatística, mais especificamente na ferramenta semivariograma (teórico e experimental). O semivariograma permite representar a variação quantitativa das variáveis regionalizadas através do cálculo da semivariância segundo a distância entre pontos, medindo a dissimilaridade entre os dados. Quanto menor o valor da semivariância, mais parecidos são os dados entre os pontos analisados. Portanto, se há dependência espacial é natural que o valor da semivariância cresça com a distância e se estabilize em um patamar, a partir do qual não há mais dependência espacial (Burrough, 1986). O semivariograma é dado pela Equação 1.

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

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

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

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

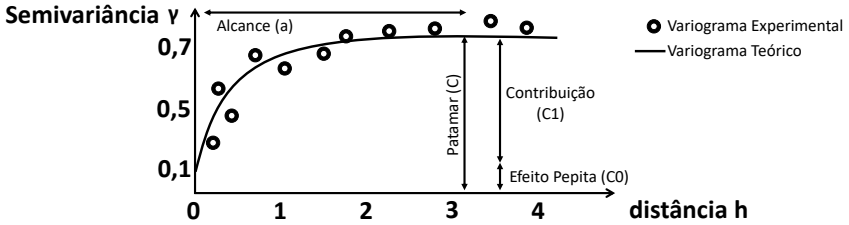

Figura 1: Esquema semivariograma teórico/experimental. Adaptado de Burrough (1986).

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

A Análise espacial confirmatória engloba o conjunto de modelos de estimação, além de procedimentos de validação. São realizadas análises multivariadas ou univariadas com componentes espaciais. A etapa de análise espacial confirmatória deste trabalho se deu com aplicação de modelos de escolha discreta, a partir da inclusão de covariáveis espaciais. A próxima seção descreve, sucintamente, essa família de modelos.

3. MODELOS DE ESCOLHA DISCRETA

Modelos de escolha discreta avaliam a probabilidade de um indivíduo escolher uma opção entre um conjunto de alternativas discretas. Tais modelos contrastam com modelos de consumo padrão nos quais a quantidade consumida por um indivíduo é assumida como uma variável contínua. No caso contínuo, métodos de cálculo de primeira ordem podem ser utilizados na determinação da quantidade ótima da variável dependente, sendo sempre possível realizar a modelagem empiricamente por meio de análises de regressão. Por outro lado, a análise de escolha discreta examina situações onde o resultado (variável dependente) é discreto, de modo que o ótimo não pode mais ser caracterizado por condições de primeira ordem. Assumindo que os modelos de escolha discreta relacionam estatisticamente a escolha (variável dependente discreta) aos atributos dos indivíduos e das alternativas disponíveis, deve-se recorrer a modelagens realizadas

através de métodos de máxima verossimilhança paramétricos. Novamente, o leitor que buscar maior formalismo matemático acerca da técnica deve consultar o trabalho de Ben-Akiva et al. (1985).

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

Sabe-se, no entanto, que a vizinhança pode influenciar os atributos envolvidos nos modelos de escolhas discreta. Portanto, variáveis regionalizadas passaram a ser incorporadas aos modelos de escolha discreta com o intuito de promover melhores estimativas. Esse trabalho, utilizando uma abordagem baseada em preceitos da geoestatísca, propõe-se a extrair covariáveis espaciais a fim de promover incrementos à modelagem paramétrica.

A etapa de análise espacial confirmatória, relativa a este trabalho, consiste na calibração de dois modelos de escolha discreta (*logit* multinomial): o primeiro utilizando somente as variáveis originais do banco de dados (características dos indivíduos e das alternativas); e um segundo com o acréscimo de covariáveis espaciais, cuja utilização justifica-se pela análise espacial exploratória, realizada previamente.

4. MATERIAIS E MÉTODOS

4.1. Cidade Fictícia

A cidade fictícia, utilizada nesse trabalho, foi criada no trabalho de Assirati et al. (2016) por meio de metodologia baseada em interpolação bi-linear de dados. Consiste de 580 indivíduos georreferenciados, localizados em um plano com 10 unidades de distância no eixo X e 5 unidades de distância no eixo Y. Além das coordenadas, cada um dos pontos da amostra é dotado de outras sete variáveis: Distância média de viagens para realização de atividades (em unidades de distância); Tempo de viagem para o modo de transporte ônibus (em min.); tempo de viagem para o modo de transporte automóvel (em min.), tarifa do ônibus (em reais), custos do uso do automóvel (em reais), indicador de renda (escala de 1 – menor renda, até 10 – maior renda) além de uma variável binária responsável por indicar a escolha modal do indivíduo: (0) uso de ônibus ou (1) uso do automóvel. A Tabela 1 apresenta medidas descritivas das variáveis e a Figura 2 apresenta a distribuição geográfica da variável binaria:

Tabela 1: Medidas descritivas do banco de dados – Cidade Fictícia

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

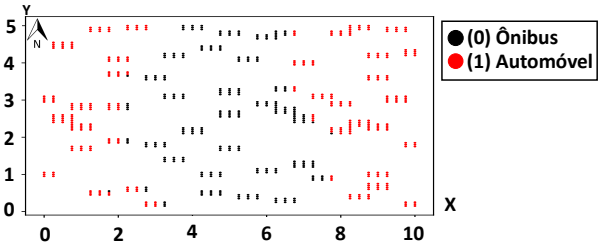

Figura 2: Representação espacial do binário de escolha modal na cidade sintética.

4.2. Procedimento metodológico

As etapas metodológicas, realizadas neste trabalho, são descritas nas próximas subseções.

4.2.1. Análise espacial exploratória

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

As dimensões do espaço amostral são de 10 unidades no eixo X e 5 unidades no eixo Y. Assim, a maior distância possível nesse espaço é a diagonal principal que vale aproximadamente 12 unidades, esse conceito é herdado da geoestatística e é denominado *cut-distance*. Trata-se de um limite superior para cálculos, sejam eles de probabilidade, de variância, de correlação, dentre outros. Constituiu-se quatro faixas de vizinhança através da divisão igualitária da diagonal em quatro parcelas: (I) 0-3 unidades; (II) 3,1-6 unidades; (III) 6,1-9 unidades; (IV) 9,1-12 unidades. A constituição das faixas também é outro conceito herdado da geoestatística e é denominado *lag-distance*. Refere-se aos segmentos, delimitados pela *cut-distance*, usados para auxiliar na aplicação dos cálculos (probabilidade, variância, correlação) por pontos. Todos as *lag-distances* comumente tem o mesmo tamanho e são equidistantes.

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

4.2.2. Proposta de covariáveis espaciais

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

Baseado nos preceitos geoestatísticos, que consistem da caracterização estatística de variáveis regionalizadas (Yamamoto e Landim, 2015), se estabeleceu um método de mensurar as probabilidades que um individuo tem de pertencer a uma determinada categoria (variável analisada) segundo os valores de seus vizinhos (caráter de regionalidade). O intuito da abordagem é o de complementar a análise paramétrica tradicional através da inclusão de covariáveis regionais, pois segundo o trabalho de Naizer (2018), o acréscimo dessas informações pode ser benéfico para modelagens de escolha discreta se for atestado o caráter de associação espacial da amostra.

As covariáveis espaciais deste trabalho são as probabilidades dos vizinhos (determinados a partir das quatro vizinhanças) escolherem as duas categorias de modo de transporte, adotadas neste estudo: ônibus ou automóvel.

Para o cálculo das probabilidades, o primeiro passo consiste do estabelecimento de vizinhanças de análise. Sugere-se aferir a distância máxima existente entre os dois pontos mais distantes entre si e dividir em porções iguais essa distância a fim de se determinar *n* vizinhanças de interesse, sendo que as mesmas não devem se sobrepor.

Estabelecidas as vizinhanças, efetua-se o cálculo, por indivíduos, das probabilidades de pertencimento às categorias, dadas as categorias dos seus vizinhos. Para cada indivíduo e para cada vizinhança, conta-se a quantidade de elementos compreendidos na região que pertençam a cada categoria existente no estudo. A probabilidade desse individuo pertencer a cada uma das categorias é a razão entre o número de elementos de uma dada categoria e o número total de elementos, encontrados na vizinhança considerada. A Figura 3 apresenta um exemplo dos procedimentos citados.

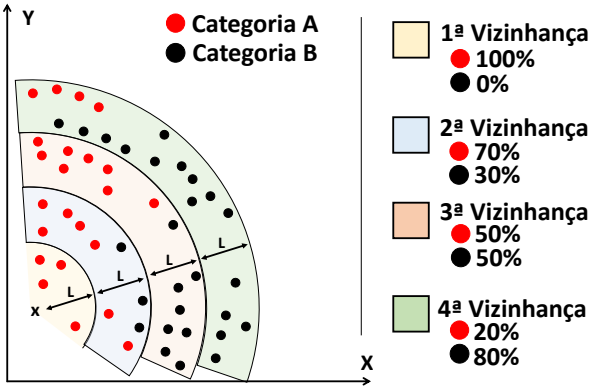

Figura 3: Exemplo de cálculo de probabilidades por categorias para o ponto X, segundo quatro vizinhanças de interesse.

Caso deseje-se caracterizar um indivíduo localizado no ponto X segundo as probabilidades de pertencimento a uma determinada categoria baseadas em vizinhanças, procede-se:
a) Dadas as dimensões máximas da distribuição de dados, divide-se o espaço em um número *n* de vizinhanças de interesse. Para o exemplo, os dados estão dispostos em um quarto de circunferência com raio de tamanho 4L. Assim, é possível delimitar n=4 faixas de tamanho L cada. Cada faixa será, portanto, uma vizinhança de análise.
b) Conta-se o número de elementos de todas as categorias presentes na primeira vizinhança. Para o exemplo, na primeira faixa há 4 elementos vermelhos da categoria A e nenhum elemento preto da categoria B. Logo um indivíduo localizado em X tem, segundo a primeira vizinhança (n=1):
- P1(A) = 1 (4 elementos de um total de 4) de ser da categoria A; e
- P1(B) = 0 (0 elementos de um total de 4) de ser da categoria B.
c) Atua-se de maneira análoga para as demais vizinhanças (n=2, 3 e 4). Assim, o indivíduo X é caracterizado:
- P2(A) = 0,7 (7 elementos de um total de 10) de ser da categoria A;
- P2(B) = 0,3 (3 elementos de um total de 10) de ser da categoria B;

- P3(A) = 0,5 (8 elementos de um total de 16) de ser da categoria A;
- P3(B) = 0,5 (8 elementos de um total de 16) de ser da categoria B;
- P4(A) = 0,2 (4 elementos de um total de 20) de ser da categoria A; e
- P4(B) = 0,8 (16 elementos de um total de 20) de ser da categoria B.

d) Repete-se os procedimentos anteriores para todos os pontos da amostra.

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

4.2.3. Calibração de funções utilidade apenas com variáveis originais

Esta etapa corresponde à calibração do modelo de escolha discreta não espacial. A determinação das funções utilidade, nesta etapa, é baseada nas variáveis que representam as alternativas modais (tempo do ônibus, tempo do automóvel, preço do ônibus e preço do automóvel), da viagem (distância de deslocamento), bem como variáveis individuais (renda). A variável resposta é o binário de escolha modal (0) ônibus e (1) automóvel. As Equações 2 e 3, na sua forma literal, representam as funções utilidade do modelo não espacial:

$$V_0 = \beta_{0\_dist} * dist + \beta_{0\_renda} * renda + \beta_{tempo} * tempo\_onibus + \beta_{preco} * preco\_onibus \quad (2)$$

$$V_1 = CTE + \beta_{1\_dist} * dist + \beta_{1\_renda} * renda + \beta_{tempo} * tempo\_auto + \beta_{preco} * preco\_auto \quad (3)$$

As variáveis *"renda"* e *"distância"* foram incluídas em ambas as funções utilidade V_0 e V_1 apesar de não variarem entre alternativas (apenas entre indivíduos). Sendo assim, para que seja garantido o pressuposto de distinguibilidade entre alternativas, foi necessário que houvesse parâmetros distintos β_{0_dist}, β_{0_renda} e β_{1_dist}, β_{1_renda}, para as funções V_0 e V_1, respectivamente. Já os parâmetros β_{tempo} e β_{preco} são genéricos pois o pressuposto de distinguibilidade está garantido.

4.2.4. Calibração de funções utilidade com variáveis originais e covariáveis espaciais propostas

Caso a análise espacial exploratória aponte que existe correlação espacial entre as variáveis, espera-se que a inclusão de informações espaciais aumente o potencial de acerto e qualidade da modelagem paramétrica. Aplica-se então o cálculo das probabilidades de pertencimento às categorias baseado nos valores de vizinhança.

O estabelecimento das probabilidades por categoria em cada uma das vizinhanças provê o acréscimo informacional de caráter regionalizado que se buscava para incrementar as análises paramétricas de escolha discreta. Antes se levava em conta quatro atributos modais (tempo do ônibus, tempo do carro, custo do ônibus e custo do carro), um atributo individual (renda) e um atributo de viagem (distância).

Agora, além destes, conta-se com novos oito atributos dos indivíduos: PI(0), PI(1), PII(0), PII(1), PIII(0), PIII(1), PIV(0), PIV(1), que são respectivamente a probabilidade P de um indivíduo escolher realizar deslocamentos por ônibus (0) a partir dos valores dos elementos na primeira vizinhança I, probabilidade P de um indivíduo escolher realizar deslocamentos por automóvel (1) a partir dos valores dos elementos na primeira vizinhança I, e assim por diante, alternando as probabilidade P de escolha modal até a vizinhança IV.

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

O acréscimo informacional locacional define agora novas funções utilidade, representadas pelas Equações 4 e 5:

$$V_0 = \beta_{0\_dist} * dist + \beta_{0\_renda} * renda + \beta_{tempo} * tempo\_onibus + \beta_{preco} * preco\_onibus + \\ + \beta_{0\_vI} * PI(0) + \beta_{0\_vII} * PII(0) + \beta_{0\_vIII} * PIII(0) + \beta_{0\_vIV} * PIV(0) \quad (4)$$

$$V_1 = CTE + \beta_{1\_dist} * dist + \beta_{1\_renda} * renda + \beta_{tempo} * tempo\_auto + \beta_{preco} * preco\_auto + \\ + \beta_{1\_vI} * PI(1) + \beta_{1\_vII} * PII(1) + \beta_{1\_vIII} * PIII(1) + \beta_{1\_vIV} * PIV(1) \quad (5)$$

4.2.5. Análise comparativa entre modelos de escolha discreta não espacial e espacial

Para a comparação de modelos de escolha discreta (espaciais e não espaciais) serão utilizados três métricas: Os valores de rho-quadrado ajustado; valor do critério Akaike; e taxa de acertos por validação cruzada e logaritmo da verossimilhança para amostra de validação.

A métrica rho-quarado é definida pela equação:

$$\rho^2 = 1 - \frac{L^*}{L_0}, \quad (6)$$

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

Caso os modelos contassem com mesmo número de parâmetros, a métrica rho-quadrado bastaria. Entretanto, pretende-se confrontar modelos com variáveis *"originais" versus* modelos que adicionam covariáveis espaciais às variáveis *"originais"*. Os modelos apresentam diferente número de parâmetros e portanto demandam avaliação pela métrica rho-quadrado ajustado, pois esta permite comparar modelos estimados a partir da mesma amostra de observações, mas com quantidade diferente de parâmetros.

A métrica rho-quadrado ajustado é definida pela equação:

$$\rho^2{}_* = 1 - \frac{L^* - K}{L_0}, \quad (7)$$

L_0 e L^* são análogos a Equação 6. K é o número de parâmetros estimados. Assim a interpretação da métrica ajustada é análoga a convencional, levando-se em consideração o desconto do número K de parâmetros do valor L^*.

O critério Akaike é definido por:

$$A = 2K - 2 \ln L^* \quad (8)$$

Os valores de K e L^* são análogos às Equações 6 e 7. Estabelecida por meio de uma subtração

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

Finalmente a validação cruzada consiste em segregar uma parcela da amostra para ser usada no processo de estimação de parâmetros e uma outra parcela usada no processo de validação dos parâmetros estimados. Após, aplica-se a modelagem junto dos parâmetros calibrados advindos do grupo de calibração, nos elementos do grupo de validação, obtendo-se assim valores estimados para esse grupo. Como os valores do grupo de validação são conhecidos, pode-se aferir a taxa de acertos ao se comparar os valores reais com os valores estimados. Bons modelos tendem a ter taxas de acerto altas e a qualidade deste valor pode ser mensurada pelo cálculo de verossimilhança:

$$L = p^y.(1-p)^{(n-y)}, \tag{9}$$

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

5. RESULTADOS E DISCUSSÕES

Inicialmente, foi realizada a análise espacial exploratória, com base no cálculo dos indicadores Moran e SivarG, apresentados na Tabela 2. Vale ressaltar aqui que, as vizinhanças escolhidas foram aquelas que os indicadores foram considerados estatisticamente significativos pelos testes de hipótese, associados a pelo menos um dos indicadores. A hipótese nula, para os testes de hipótese de ambos os indicadores, é de aleatoriedade espacial. Neste caso, as vizinhanças de interesse são aquelas onde a hipótese nula foi rejeitada em pelo menos um dos indicadores (Vizinhanças I, II e III).

Tabela 2: Resultado dos indicadores Moran e SivarG aplicados ao banco de dados sintético.

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

Verifica-se, a partir da Tabela 2, que os indicadores são grandezas quase que complementares. A proximidade do valor 1, para o índice de Moran, indica alta autocorrelação espacial, enquanto que o valor 1 indica muito baixa autocorrelação espacial para o indicador SivarG.

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

Posteriormente, foram calculadas as probabilidades dos vizinhos, que pertencem às vizinhanças

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

A próxima etapa, relativa à calibração das funções utilidades das variáveis originais, baseou-se nas Equações 2 e 3. Ao realizar a calibração pelo método da máxima verossimilhança, foram obtidos os parâmetros β estimados. Entretanto, nem todos foram significativos, sendo assim, foram retiradas, das funções utilidade, aquelas variáveis, associadas aos parâmetros não significativos (β_{renda} em ambas as Equações 2 e 3). Assim sendo, faz-se necessário recalibrar o modelo, desconsiderando as variáveis que não são pertinentes ao estudo, conforme Equações 9 e 10.

$$V_0 = \beta_{0\_dist} * dist + \beta_{tempo} * tempo\_onibus + \beta_{preco} * preco\_onibus \tag{10}$$

$$V_1 = CTE + \beta_{1\_dist} * dist + \beta_{tempo} * tempo\_auto + \beta_{preco} * preco\_auto \tag{11}$$

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

Tabela 3: Calibração pelo método da máxima verossimilhança considerando as variáveis originais significativas da base de dados

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

Em seguida, é realizada a calibração do modelo espacial com a incorporação das covariáveis espaciais relativas às vizinhanças I, II e III, por indivíduo. Como os testes de hipótese não apontaram significância para a região IV, as variáveis relativas a essa região e seus respectivos parâmetros foram retirados das Equações 4 e 5. Novamente, como na primeira modelagem, após a calibração, houve um coeficiente não-significativo (coeficiente relativo a terceira vizinhança na Equação 5) demandando a eliminação de sua variável da equação e uma nova calibração.

Desconsiderando os coeficientes não significantes e suas respectivas variáveis associadas, promove-se a nova calibração de acordo com as Equações 11 e 12:

$$V_0 = \beta_{0\_dist} * dist + \beta_{tempo} * tempo\_onibus + \beta_{preco} * preco\_onibus + \\ + \beta_{0\_vI} * PI(0) + \beta_{0\_vII} * PII(0) + \beta_{0\_vIII} * PIII(0) \tag{12}$$

$$V_1 = CTE + \beta_{1\_dist} * dist + \beta_{tempo} * tempo\_auto + \beta_{preco} * preco\_auto + \\ + \beta_{1\_vI} * PI(1) + \beta_{1\_vII} * PII(1) \tag{13}$$

Os resultados são expressos na Tabela 4. Novamente, segregando a amostra em uma proporção 7/3 para testes de validação cruzada com os coeficientes calibrados obteve-se uma taxa de acertos de 89% para a segunda modelagem, com valor de verossimilhança $L = 1,82 \times 10^{-39}$ e $\log(L) = -79,98$.

Tabela 4: Calibração pelo método da máxima verossimilhança considerando as covariáveis espaciais acrescidas significativas.

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

Ao se comparar a primeira modelagem que considerava as variáveis significativas originais da base de dados (Tabela 3) com a segunda modelagem que considerava as variáveis significativas originais e as covariáveis regionais (Tabela 4), corrobora-se a hipótese de que, se averiguada a existência de associação espacial das variáveis por meio de análise exploratória, pode-se incluir o aspecto espacial ao modelo pois este trará maior precisão aos resultados.

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

6. CONCLUSÕES

Este trabalho apresentou um método espacial sequencial para previsão de escolha modal, a partir de uma base fictícia, composta de dados espacialmente dependentes. Inicialmente, foram calculados indicadores de associação espacial global: Moran (Anselin, 1995) e SivarG (Naizer e Pitombo, 2017; Naizer, 2018). Com a observação dos testes de hipótese, associados aos indicadores, foram determinadas as vizinhanças a serem acrescidas à análise.

Em seguida, na etapa de análise espacial confirmatória, foram comparados modelos de escolha discreta não espaciais e espaciais. Os modelos espaciais foram aqueles com incorporação de covariáveis espaciais. Observou-se um aprimoramento na modelagem, quando consideradas as covariáveis espaciais (aumento de 15% de taxas de acertos a partir de amostra de validação).

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

Salienta-se que, a despeito do uso de uma base de dados sintéticos, a maior contribuição deste trabalho é metodológica. O método proposto baseia-se na Análise Espacial Exploratória, inicialmente, a qual utiliza um indicador de associação espacial global, baseado no variograma teórico (SivarG). Em seguida, são propostas covariáveis regionais, onde a delimitação das vizinhanças também é determinada a partir de conceitos variográficos. Em seguida, são calibradas funções utilidade com e sem covariáveis regionais. O método proposto, baseado em preceitos da geoestatística, pode ser replicado para qualquer estudo de caso com dados espacialmente dependentes.

Agradecimentos

Agradecimentos à Fundação de Amparo à Pesquisa do Estado de São Paulo - FAPESP (número do processo 13/25035-1), ao Conselho Nacional de Desenvolvimento Científico e Tecnológico - CNPq, e à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Capes, pelos recursos disponibilizados para condução desta pesquisa.

REFERÊNCIAS BIBLIOGRÁFICAS

Ahern, A. A. e N. Tapley (2008). The use of stated preference techniques to model modal choices on interurban trips in ireland. *Transportation Research Part A: Policy and Practice 42*(1), 15–27.

Anselin, L. (1995). Local indicators of spatial association—lisa. *Geographical analysis 27*(2), 93–115.

Assirati, L., S. S. Rocha, M. U. Caldas, e C. S. Pitombo (2016). Interpolação bi-linear para simulação de dados espacialmente correlacionados: Uma aplicação relativa à demanda por transportes. In *Anais do XXX Congresso da Associação Nacional de Ensino e Pesquisa em Transportes*. ANPET.

Ben-Akiva, M. E., S. R. Lerman, e S. R. Lerman (1985). *Discrete choice analysis: theory and application to travel demand*, Volume 9. MIT press.

Bhat, C. R., N. Eluru, e R. B. Copperman (2008). Flexible model structures for discrete choice analysis. *Handbook of Transport Modelling, 5*, 75–104.

Burrough, P. (1986). 1986. principles of geographical information systems for land resources assessment. *Monographs on Soil and Resources Survey* (12).

Getis, A. e J. K. Ord (1992). The analysis of spatial association by use of distance statistics. *Geographical analysis 24*(3), 189–206.

Hess, S. (2005). *Advanced discrete choice models with applications to transport demand*. Ph. D. thesis, University of London.

Miyamoto, K., V. Vichiensan, N. Shimomura, e A. Páez (2004). Discrete choice model with structuralized spatial effects for location analysis. *Transportation research record: journal of the transportation research board* (1898), 183–190.

Moran, P. A. (1950). Notes on continuous stochastic phenomena. *Biometrika 37*(1/2), 17–23.

Naizer, C. (2018). *Procedimento Metodológico para Proposta de Indicadores de Associação espacial Global e Local através de conceitos variográficos*. Ph. D. thesis, Universidade de São Paulo.

Naizer, C. e C. S. Pitombo (2017). Procedimento metodológico para proposta de indicadores de associação espacial global através de conceitos variográficos. In *Anais do XXV Congresso da Associação Nacional de Ensino e Pesquisa em Transportes*. ANPET.

Páez, A., F. A. López, M. Ruiz, e C. Morency (2013). Development of an indicator to assess the spatial fit of discrete choice models. *Transportation Research Part B: Methodological 56*, 217–233.

Pitombo, C. S., A. R. Salgueiro, A. S. G. da Costa, e C. A. Isler (2015). A two-step method for mode choice estimation with socioeconomic and spatial information. *Spatial Statistics 11*, 45–64.

Qin, H., J. Gao, H. Guan, e H. Chi (2017). Estimating heterogeneity of car travelers on mode shifting behavior based on discrete choice models. *Transportation Planning and Technology 40*(8), 914–927.

Tobler, W. R. (1970). A computer movie simulating urban growth in the detroit region. *Economic geography 46*(sup1), 234–240.

Yamamoto, J. K. e P. M. B. Landim (2015). *Geoestatística: conceitos e aplicações*. Oficina de textos.

Zhou, J. (2012). Sustainable commute in a car-dominant city: Factors affecting alternative mode choices among university students. *Transportation research part A: policy and practice 46*(7), 1013–1029.