ABORDAGEM BAYESIANA NA ESTIMAÇÃO DE MATRIZES ORIGEM-DESTINO SINTÉTICAS EM REDES DE TRANSPORTES

Anselmo Ramalho Pitombeira Neto
Bruno Vieira Bertoncini
Carlos Felipe Grangeiro Loureiro
Grupo de Pesquisa em Transporte, Trânsito e Meio Ambiente - GTTEMA
Universidade Federal do Ceará - UFC

RESUMO
A modelagem das origens e destinos das viagens em uma rede de transportes constitui-se em etapa fundamental da atividade de Planejamento de Transportes. Um dos principais produtos dessa modelagem é uma matriz de origens e destinos (matriz OD), a qual representa o padrão de fluxos das viagens entre zonas em uma área de estudo. Em geral, costuma-se empregar métodos estatísticos e de otimização matemática para auxiliar a obtenção de uma matriz OD. O objetivo deste artigo é apresentar a abordagem da Estatística Bayesiana na modelagem das origens e destinos de viagens em redes de transporte. Inicialmente, são apresentados os conceitos básicos da Estatística Bayesiana, ressaltando-se suas diferenças em relação à Estatística Frequentista. Em seguida, desenvolve-se um exemplo para ilustrar a aplicação da abordagem bayesiana em estudos em transportes, estabelecendo as bases conceituais para a discussão de alguns modelos bayesianos propostos na literatura para a obtenção de matrizes OD sintéticas. Por fim, são analisadas algumas questões de pesquisa relevantes à aplicação efetiva da abordagem bayesiana na estimação de matrizes OD sintéticas em redes reais de transportes.

ABSTRACT
Modeling trip origins and destinations in a transportation network constitutes an essential stage in the Transportation Planning process. A major product of this modeling effort is a matrix of origins and destinations (OD matrix), which represents the pattern of trip flows between zones in a study area. In general, it is common to employ statistical methods and mathematical optimization to help obtaining an OD matrix. The aim of this paper is to present the Bayesian approach to statistical modeling of trip origins and destinations in transport networks. Initially, the basic concepts of Bayesian statistics are presented, highlighting their differences to the frequentist statistics. Then, it is developed an example to illustrate the application of the Bayesian approach in transportation studies, establishing the conceptual basis to the discussion of some Bayesian models proposed in the literature to obtain synthetic OD matrices. Finally, some research questions are analyzed, relevant to the effective application of the Bayesian approach in the estimation of synthetic OD matrices in real transportation networks.

1. INTRODUÇÃO

A modelagem da relação entre a demanda e a oferta no sistema de transportes vem sendo realizada nos últimos 50 anos por meio de um processo sequencial no qual se busca representar o contexto decisório dos usuários ao tentarem satisfazer suas necessidades de deslocamento, definindo os horários de realização das viagens, seus destinos, assim como os modos e rotas utilizados na rede de transportes. Entende-se que o objetivo geral deste processo tradicional tem sido modelar, de forma agregada, o padrão dos fluxos das viagens produzidas e atraídas pelas zonas de origem (O) e destino (D) na área de estudo, tendo como produto final a simulação dos volumes de tráfego de pessoas ou cargas, nos nós e arcos de uma rede de transporte urbano ou regional.

Especificamente no contexto urbano, o método clássico de modelagem do padrão dos fluxos entre pares de zonas OD tem se baseado nas análises de correlação entre as viagens realizadas e as variáveis explicativas representando os atributos socioeconômicos da população, as características físicas e operacionais do sistema de transportes, assim como os aspectos quantificáveis do uso e ocupação do solo (Ortúzar e Willumsen, 1994). Nesta abordagem tradicional, os valores das variáveis que descrevem os diversos atributos das viagens, e as respectivas características socioeconômicas dos usuários, são normalmente obtidos por meio de entrevistas domiciliares e pesquisas complementares em vias, terminais e pontos de parada,

realizadas para caracterizar o padrão de deslocamento dos usuários ao longo de um dia útil típico. Como fruto principal da consolidação dos dados resultantes desses levantamentos, tem-se uma matriz amostral das viagens observadas entre pares OD, por motivo, horário ou modo, que se constitui na referência básica para o desenvolvimento dos demais modelos de geração, distribuição, divisão modal e alocação das viagens na rede analisada.

Um aspecto digno de ressalva é que esta matriz OD amostral representa apenas uma única observação (amostra com n = 1) da população dos deslocamentos que acontecem na área de estudo, sendo então expandida para representar o conjunto dos deslocamentos realizados em um dia útil típico. Vale destacar, portanto, que esta matriz OD expandida não pode ser entendida como um estimador da matriz populacional dos fluxos OD. Ortúzar e Willumsen (1994) descrevem as limitações dos métodos que são normalmente utilizados para se obter a matriz OD expandida, especialmente considerando a quantidade significativa de células vazias na matriz OD amostral, destacando esse esforço de expansão como uma importante fonte de erros nas etapas de calibração e validação dos modelos agregados de demanda.

Na tentativa de reduzir custos de coleta de dados em campo e superar as dificuldades inerentes à obtenção de uma matriz OD representativa do padrão de fluxos na área em estudo, o problema da estimação de matrizes OD sintéticas, isto é, geradas a partir dos carregamentos observados nos arcos da rede, vem sendo estudado desde a década de 1970, a partir de trabalhos pioneiros como os de Robillard (1975) e Nguyen (1977), que abordaram, respectivamente, redes viárias sem e com congestionamento. O principal desafio teórico nesta linha de pesquisa diz respeito ao fato do número de pares OD ser normalmente bastante superior ao número de arcos com volumes conhecidos, tornando o sistema de equações resultante subespecificado, isto é, com múltiplas soluções possíveis. Segundo resumido por Rakha *et al.* (2005), é teoricamente possível que existam múltiplas matrizes OD que repliquem exatamente os volumes observados em campo, levando à necessidade de avaliar qual delas seria a “mais provável” de ter gerado aquela configuração de carregamento da rede.

Willumsen (1981), Nguyen (1984), Cascetta e Nguyen (1988) e, mais recentemente, Abrahamsson (1998) e Timms (2001) realizaram revisões detalhadas da literatura relativa a esse problema, buscando classificar, sob óticas distintas, os esforços de pesquisa empreendidos até o final do século XX, de acordo com as diferentes formulações matemáticas e abordagens de solução propostas. No entanto, apesar dos avanços teóricos alcançados, os algoritmos computacionais já implementados continuam sendo pouco utilizados em contextos práticos, algumas vezes de forma inadequada, pela comunidade técnica internacional, em função não só da dificuldade advinda da complexidade matemática das abordagens propostas, como também da falta de compreensão das premissas e objetivos das diversas formulações do problema (Timms, 2001).

Conforme o tipo de abordagem de solução, os principais modelos sintéticos propostos na literatura podem ser classificados em dois grandes grupos: modelos de otimização e modelos estatísticos. Nos modelos de otimização, geralmente baseados nos princípios da maximização da entropia ou minimização da informação (Van Zuylen e Willumsen, 1980), o objetivo é determinar a matriz OD mais provável dentre as soluções possíveis, a partir da minimização de uma função dos erros na estimativa das demandas ou na replicação dos volumes observados, sujeita a um conjunto de restrições de fluxo. Ou, como colocado por Cascetta (1984), encontrar a matriz OD que minimiza uma medida de distância, “entrópica” ou euclidiana, em relação a uma matriz alvo ou semente (gerada a partir de um modelo comportamental da demanda ou estimada para um período anterior), respeitando a restrição

que, uma vez alocada na rede, reproduza os carregamentos observados em campo. Dessa forma, pode-se dizer que esses modelos tentam reconstruir – e não estimar, como se convencionou chamar – a matriz OD que gerou os volumes observados em campo (Hazelton, 2001). Vale ressaltar, entretanto, que essa matriz corresponde a apenas uma realização particular da população de matrizes que representam o fenômeno do padrão de deslocamento, de forma que uma nova observação de volumes produzirá uma matriz OD diferente.

Por outro lado, nos modelos estatísticos o objetivo é estimar os parâmetros da população de matrizes OD, dada uma amostra de volumes observados. Nesse sentido, entende-se por estimação da matriz OD sintética o processo de obtenção dos parâmetros do modelo subjacente ao padrão das origens e destinos das viagens. No conjunto das técnicas estatísticas, pode-se fazer distinção entre as que seguem os princípios da inferência frequentista e aquelas que incorporam a abordagem bayesiana. Enquanto na estatística frequentista os valores estimados dos parâmetros baseiam-se em geral somente em observações amostrais, na abordagem bayesiana é possível incorporar explicitamente informação prévia à coleta de dados, a qual pode ser obtida a partir da experiência do analista, de matrizes OD produzidas por estudos anteriores, ou quaisquer outras fontes que forneçam informação relevante para o processo de estimação. Acredita-se que essa característica torne a abordagem bayesiana uma alternativa promissora para a melhoria do processo de modelagem da demanda por transporte.

Face ao exposto, os objetivos e a estrutura deste artigo são: i) evidenciar as diferenças conceituais e metodológicas entre as abordagens frequentista e bayesiana da Inferência Estatística (item 2); ii) revisar as principais aplicações da Estatística Bayesiana na estimação de matrizes OD sintéticas em redes de transportes (item 3); e iii) propor questões que possibilitem avanços teórico-práticos nesta linha de pesquisa para uma melhor compreensão do padrão de deslocamento de pessoas e cargas, ou seja, para uma modelagem mais efetiva das relações entre demanda e oferta nos processos de planejamento estratégico e operacional de sistemas de transporte urbano ou regional (item 4).

2. ESTATÍSTICA BAYESIANA: CONCEITOS BÁSICOS E METODOLOGIA

Inferência é o ramo da Estatística cujo objetivo é generalizar para uma população conclusões obtidas a partir de uma amostra. Particularmente, chamam-se de parâmetros as características populacionais que se deseja estimar a partir de observações (informações) amostrais. Nas seções seguintes, são evidenciadas as diferenças entre as abordagens frequentista e bayesiana de estimação de parâmetros, com a apresentação de um exemplo ilustrativo de aplicação da Estatística Bayesiana no contexto de redes de transporte.

2.1. Diferenças conceituais e metodológicas entre as abordagens frequentista e bayesiana

Costuma-se referir como “abordagem frequentista” da Estatística a um conjunto de métodos estatísticos, e sua teoria subjacente, que partem da premissa de que a probabilidade associada à ocorrência de um evento corresponde aproximadamente à freqüência relativa da ocorrência desse evento em uma sequência de repetições de um experimento aleatório. Nessa interpretação, a probabilidade associada a um evento é uma propriedade objetiva deste, a qual pode ser atribuída somente por meio da observação empírica da ocorrência do evento. De acordo também com essa interpretação, os parâmetros de uma população são propriedades objetivas e fixas, as quais não podem ser diretamente observadas ou calculadas com precisão já que os dados disponíveis se referem a uma amostra, e não a toda a população. Nesse caso, estimar parâmetros corresponde a utilizar um método que proponha o melhor valor possível para o parâmetro avaliado, dadas as observações amostrais e um critério de escolha. Alguns dos métodos típicos da Estatística Frequentista são o método dos momentos, o método da

máxima verossimilhança e o método dos mínimos quadrados.

A abordagem bayesiana, por outro lado, considera a probabilidade de um evento ocorrer sob um ponto de vista subjetivo, interpretando-a como o nível de crença associado à sua ocorrência. Antes da observação do evento, o analista possui uma crença prévia sobre a probabilidade de esse ocorrer, chamada de probabilidade *a priori*. Após a observação de uma ou mais ocorrências do evento, o analista atualiza seu nível de crença, atribuindo uma probabilidade *a posteriori* ao evento. Nos métodos bayesianos de estimação de parâmetros, diferentemente dos métodos frequentistas, parâmetros desconhecidos não são tratados como constantes da população, mas sim como variáveis aleatórias que possuem distribuições de probabilidade *a priori* (antes das observações amostrais) e *a posteriori* (após as observações amostrais). Dessa forma, é preciso se atribuir uma distribuição de probabilidades *a priori* para os parâmetros, representando o nível de crença inicial quanto aos seus valores possíveis. Após a obtenção de observações amostrais, o nível de crença é então atualizado para distribuições de probabilidades *a posteriori*. A regra de atualização é dada pelo chamado Teorema de Bayes, apresentado em **(1)**, o qual relaciona matematicamente a distribuição *a priori*, a função de verossimilhança e a distribuição *a posteriori*, consistindo numa expressão das probabilidades condicionais entre um parâmetro θ e as observações y.

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} \qquad (1)$$

Na qual:
- $p(\theta)$: Distribuição de probabilidades *a priori* do parâmetro θ;
- $p(\theta|y)$: Distribuição de probabilidades *a posteriori* do parâmetro θ (dadas as observações amostrais y);
- $p(y|\theta)$: Função de verossimilhança, representando a probabilidade das observações y dado um valor específico do parâmetro θ;
- $p(y)$: Probabilidade de ocorrência das observações y.

Um estimador bayesiano para um parâmetro θ pode ser obtido por meio do valor esperado de θ com relação à distribuição *a posteriori*, como dado por (**2**).

$$\hat{\theta} = \sum_{i=1}^{m} \theta_i \, p(\theta_i|y) \qquad (2)$$

Em (**2**), um parâmetro θ assume valores discretos θ_1, θ_2 , ..., θ_m, e $p(\theta|y)$ é uma função massa de probabilidades. Caso θ assuma valores continuamente em um intervalo da reta dos números reais, o somatório em (**2**) deve ser substituído por uma integral e $p(\theta|y)$ se torna uma função densidade de probabilidades. Outros estimadores possíveis para θ são a moda da distribuição *a priori* (também chamado de máximo *a posteriori*) ou a mediana.

2.2. Exemplo ilustrativo da abordagem bayesiana no contexto de redes de transporte

As diferenças entre as duas abordagens de Inferência Estatística podem ser melhor explicitadas por meio de um exemplo simples de aplicação da abordagem bayesiana na análise de fluxos em redes de transporte. Seja uma interseção semaforizada para a qual se desenvolve um estudo para estimar a taxa média de chegada de veículos na aproximação principal, durante o período de pico do tráfego. Denotando-se por θ o parâmetro desconhecido "taxa média de chegadas em um ciclo semafórico", considere que o analista coletou uma amostra de tamanho n = 1 correspondente à observação de y = 6 veículos chegando na via principal em um determinado ciclo. Sob a abordagem frequentista, toda a informação disponível para a estimação do parâmetro θ encontra-se na amostra. Como a amostra

contempla uma única observação, a melhor estimativa que se pode produzir é $\hat{\theta} = y = 6$.

Já na abordagem bayesiana, o analista leva em conta seu conhecimento prévio sobre o comportamento do tráfego na rede viária analisada. Considere que, da experiência de estudos anteriores em interseções semelhantes, o analista acredita que a taxa θ possa estar em três patamares distintos: (i) valor mínimo θ_1 = 5 veículos/ciclo; (ii) valor mais provável θ_2 = 10 veículos/ciclo; e (iii) valor máximo θ_3 = 20 veículos/ciclo. O analista atribui a cada um desses valores as probabilidades 20%, 50% e 30%, respectivamente, que serão tomadas como probabilidades *a priori* do parâmetro θ. Considere a mesma amostra unitária de y = 6 veículos chegando em um ciclo semafórico. Como o valor y = 6 é próximo ao valor mínimo θ_1 = 5 para a taxa média de veículos, é razoável que o analista atualize sua crença em relação aos valores possíveis de θ atribuindo uma probabilidade *a posteriori* maior que 20% para essa estimativa. O cálculo dessa probabilidade *a posteriori* pode ser feito por meio de (1). Para isso, entretanto, é preciso determinar a função de verossimilhança, a qual indica a probabilidade da observação y dado um valor específico do parâmetro θ. Uma função típica utilizada no caso do processo de chegada de veículos em interseções é a distribuição de Poisson, dada por (**3**):

$$p(y|\theta) = \frac{\theta^y e^{-\theta}}{y!} \tag{3}$$

Na qual e é o número de Euler ($e \approx 2{,}718$). Dessa forma, a probabilidade *a posteriori* de θ é calculada conforme (**4**), que é obtida pela substituição de (**3**) em (1).

$$p(\theta|y) = \frac{\frac{\theta^y e^{-\theta}}{y!} p(\theta)}{p(y)} \tag{4}$$

A probabilidade *a priori* de θ ser igual ao valor otimista θ_1 = 5 é $p(\theta = 5) = 0{,}2$. Para realizar o cálculo, é preciso ainda obter o valor de $p(y = 6)$, ou seja, a probabilidade de 6 carros chegarem na aproximação principal em um ciclo semafórico. Essa probabilidade pode ser obtida somando-se as probabilidades de chegarem 6 carros dado cada um dos valores possíveis para o parâmetro θ, de acordo com (**5**):

$$p(y) = \sum_{i=1}^{m} p(y|\theta_i) \cdot p(\theta_i) \tag{5}$$

Na qual m = 3 é o número de valores possíveis para θ, $p(y|\theta_i)$ é calculada por (**3**) e $p(\theta_i)$ é a probabilidade *a priori* de $\theta = \theta_i$. Substituindo-se os valores em (**5**), o valor $p(y = 6)$ é calculado em (**6**):

$$\begin{aligned} p(y=6) &= p(y=6|\theta=5)p(\theta=5) + p(y=6|\theta=10)p(\theta=10) \\ &\quad + p(y=6|\theta=20)p(\theta=20) \\ &= 0{,}146 \times 0{,}2 + 0{,}0631 \times 0{,}5 + 0{,}000183 \times 0{,}3 \\ &= 0{,}0608 \end{aligned} \tag{6}$$

Finalmente, a probabilidade *a posteriori* $p(\theta = 5|y = 6)$ é obtida substituindo-se os valores em (**4**), obtendo-se (**7**):

$$p(\theta = 5|y = 6) = \frac{\frac{5^6 \times 2{,}718^{-5}}{6!} \times 0{,}2}{0{,}0608} = \frac{0{,}146 \times 0{,}2}{0{,}0608} = 0{,}481 \tag{7}$$

Logo, a observação de que 6 carros chegaram em um ciclo modificou a probabilidade *a priori*

de θ = 5 de 20% para uma probabilidade *a posteriori* de 48%. De forma semelhante, podem ser calculadas as probabilidades *a posteriori* de θ = 10 e θ = 20, obtendo-se $p(\theta = 10|y = 6)$ = 52% e $p(\theta = 20|y = 6) \sim 0\%$. Nota-se que a probabilidade da taxa de chegada ser igual a 10 aumentou levemente, enquanto a probabilidade de ser igual a 20 caiu consideravelmente. Com a nova distribuição de probabilidades, um estimador bayesiano do parâmetro θ é o valor esperado de θ com relação à distribuição de probabilidades *a posteriori*, como dado por (**2**), obtendo-se assim o valor dado em (**8**):

$$\hat{\theta} = 5 \times 0{,}48 + 10 \times 0{,}52 + 20 \times 0 = 7{,}6 \qquad (8)$$

Portanto, uma estimativa da taxa média de chegadas, utilizando informação prévia e informação da observação amostral é de 7,6 carros por ciclo semafórico. Caso o tamanho da amostra seja $n > 1$, o que normalmente ocorre na prática, do ponto de vista frequentista pode-se utilizar como estimador de θ a média amostral das observações, $\bar{y} = \frac{1}{n}\sum_{j=1}^{n} y_j$. Sob a abordagem bayesiana, o procedimento é essencialmente o mesmo como descrito acima, com as devidas modificações na expressão do Teorema de Bayes, que passa agora a refletir todas as observações y_j da amostra.

2.3. Distribuições conjugadas na Estatística Bayesiana

No exemplo apresentado acima, a distribuição de probabilidades *a priori* do parâmetro θ foi definida por meio de uma tabela para um número finito de valores discretos do parâmetro. Na prática, é comum se empregar distribuições conhecidas de probabilidades contínuas, tais como as distribuições Normal, Gama e Beta. Algumas dessas distribuições exibem uma propriedade peculiar: quando especificadas como distribuição *a priori*, a distribuição *a posteriori* obtida será do mesmo tipo. No entanto, isso só ocorre quando se escolhem determinados tipos de distribuição para a função de verossimilhança. Por exemplo, se a função de verossimilhança for uma distribuição de Poisson, especificando-se uma distribuição *a priori* do tipo Gama, a distribuição *a posteriori* também será do tipo Gama. Dessa forma, diz-se que a distribuição Gama é a conjugada da distribuição de Poisson. Em (9) tem-se a distribuição de probabilidades Gama com parâmetros α e β, em que $\Gamma(\alpha)$ denota a função Gama, $\Gamma(\alpha) = \int_0^\infty t^{\alpha-1}e^{-t}dt$.

$$p(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}e^{-\beta\theta} \qquad (9)$$

A utilidade das funções conjugadas reside no fato de que não é necessário aplicar o Teorema de Bayes diretamente para calcular as probabilidades *a posteriori*. A partir da especificação dos parâmetros da distribuição *a priori* (também chamados de hiperparâmetros *a priori*), e das observações amostrais, pode-se calcular diretamente os parâmetros da distribuição *a posteriori* (também chamados de hiperparâmetros *a posteriori*). A Tabela 1 exibe outros casos de distribuições conjugadas (Gelman *et al.*, 2003). Como exemplo ilustrativo, considere que na especificação da distribuição *a priori* no item 2.2 fosse admitida uma distribuição Gama, com valor esperado *a priori* igual a 10 veículos/ciclo e um desvio-padrão *a priori* igual a 5 veículos/ciclo. Uma vez que, para a distribuição Gama, o valor médio é dado por α / β e o desvio-padrão por α / β^2, substituindo-se os valores, obtém-se os hiperparâmetros *a priori* α = 20 e β = 2. Dada uma amostra de n observações, y_1, y_2, ..., y_n, pode-se mostrar que os hiperparâmetros *a posteriori* serão dados por $\alpha' = \alpha + \sum_{j=1}^{n} y_j$ e $\beta' = \beta + n$ (Bolstad, 2007). Logo, dada uma amostra de n = 5, y = (12, 15, 9, 13, 14), os valores dos hiperparâmetros *a posteriori* são $\alpha' = 20 + 63 = 83$ e $\beta' = 2 + 5 = 7$. A Figura 1 exibe as distribuições Gama

a priori e *a posteriori* do parâmetro θ. É possível observar o efeito da informação amostral, reduzindo a incerteza com relação aos valores possíveis de θ, expressa pela menor dispersão na distribuição *a posteriori*. Finalmente, um estimador bayesiano para θ é dado pela média da distribuição Gama *a posteriori*, $\alpha' / \beta' = 83/7 = 11{,}9$. A média amostral, um estimador frequentista, é igual a 12,6. A diferença entre os dois valores se dá pela informação contida na distribuição *a priori*.

Tabela 1: Funções de verossimilhança e sua distribuição conjugada

Função de verossimilhança	Parâmetro	Distribuição conjugada
Binomial	p : Probabilidade de "sucesso"	Beta
Poisson	λ : Taxa média de ocorrência de eventos	Gama
Exponencial	λ : Taxa média de ocorrência de eventos	Gama
Normal	μ : Média	Normal
Multinomial	$p_1, p_2, ..., p_k$: Probabilidade da classe k ocorrer	Dirichlet

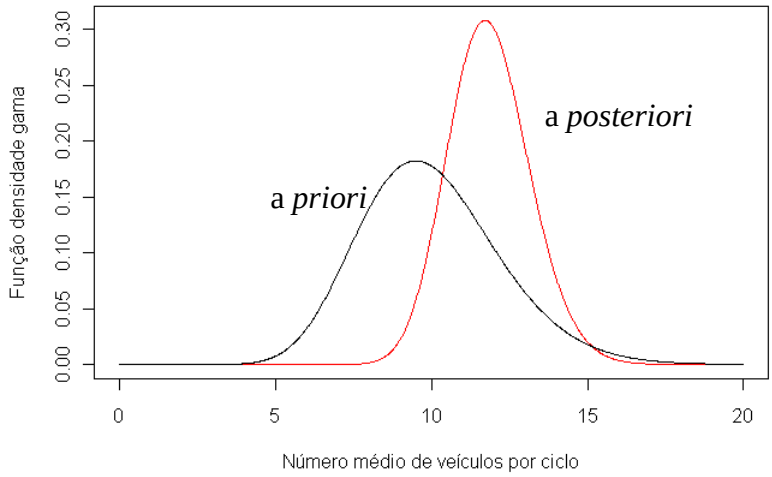

Figura 1: Distribuições Gama *a priori* e *a posteriori*

Quando não são utilizadas distribuições conjugadas, o Teorema de Bayes deve ser empregado para se calcular a distribuição *a posteriori*. No entanto, esse cálculo pode ser computacionalmente intensivo, devido principalmente ao denominador do Teorema de Bayes, ou seja, o termo $p(y)$ em (1). Esse termo funciona como uma constante de normalização, garantindo que a função *a posteriori* $p(\theta|y)$ seja de fato uma distribuição de probabilidades. Para o caso de distribuições contínuas, o termo $p(y)$ torna-se uma integral, que pode ser de difícil solução. Devido a essa dificuldade computacional, é crescente o uso de métodos denominados *Markov Chain Monte Carlo* (MCMC) em Estatística Bayesiana (Diaconis, 2009). Esses métodos possibilitam a geração de amostras da distribuição *a posteriori* sem a necessidade do cálculo da constante de normalização. A ideia básica desses métodos é gerar uma cadeia de Markov de forma que sua distribuição de probabilidades no estado de equilíbrio corresponda à distribuição de probabilidades da qual se deseja produzir amostras. Em Estatística Bayesiana, a distribuição que se deseja amostrar é a distribuição *a posteriori*, de forma que essa deve ser a distribuição de equilíbrio da cadeia de Markov gerada.

3. ESTATÍSTICA BAYESIANA NA ESTIMAÇÃO DE MATRIZES OD SINTÉTICAS

Para uma melhor compreensão do escopo de aplicação da abordagem bayesiana na estimação de matrizes OD sintéticas, considere a rede de transportes da Figura 2, a qual é constituída por 6 nós, dentre os quais os nós 1 e 2 são zonas produtoras de viagens, os nós 3 e 4 são zonas atratoras, e os nós 5 e 6 são pontos intermediários (interseções) da rede.

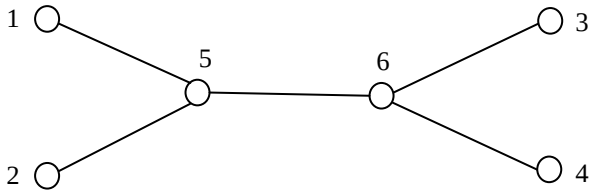

Figura 2: Exemplo de uma rede simples de transportes

Uma matriz OD para essa rede consiste no arranjo bidimensional $\mathbf{T} = (T_{13}, T_{14}, T_{23}, T_{24})$, cujas componentes correspondem ao número de viagens entre os pares de zonas, em um intervalo de tempo considerado. Sob a perspectiva da Estatística Bayesiana, o processo de estimação da matriz OD consiste em: dada uma distribuição *a priori* para os parâmetros **θ** do modelo que rege **T,** e os valores observados de volumes nos arcos, propor uma distribuição de probabilidades *a posteriori* a partir da qual é possível obter estimativas para os parâmetros **θ**. Encontram-se na literatura algumas tentativas de modelagem bayesiana de matrizes OD sintéticas para aplicações em redes de transportes, iniciando com o trabalho de Maher (1983). Tebaldi e West (1999) deram uma nova direção à abordagem, seguindo o trabalho de Vardi (1996), por meio do uso de métodos *Markov Chain Monte Carlo*. Mais recentemente, tem-se destacado os esforços de pesquisa desenvolvidos por Hazelton (2001, 2008, 2010).

3.1. Modelo de Maher (1983)

Maher (1983) formulou o problema da obtenção de uma matriz OD a partir da contagem volumétrica de tráfego em *m* arcos de uma rede viária. T_{ij} é uma variável aleatória correspondente ao número de viagens que se originam em uma zona *i* e se destinam a uma zona *j* de uma região geográfica, em um certo intervalo de tempo. As variáveis T_{ij} são não-observáveis diretamente, de forma que Maher as considera como parâmetros θ_l para o par *i*-*j* (reindexado como par *l*). Cada volume y_k em um arco *k* é uma combinação linear dos parâmetros θ_l , com constantes de proporcionalidade definidas pelas proporções p_{kl} de viagens no par *l* cujas rotas incluem o arco *k*, conforme (**10**). Ademais, um termo ε_l com valor esperado igual a zero é adicionado para representar o erro aleatório devido à imprecisão dos dados de volumes, coletados por métodos manuais ou automatizados.

$$y_k = \sum_{l=1}^{n} p_{kl}\theta_l + \varepsilon_l \qquad (10)$$

Neste modelo, admite-se que os parâmetros $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_n)$ seguem uma distribuição *a priori* Normal multivariada, com média $\boldsymbol{\mu}_0$ e matriz de covariância $\mathbf{V}_0$. $\boldsymbol{\mu}_0$ pode ser obtido a partir de uma matriz prévia, com $\mathbf{V}_0$ incorporando o grau de incerteza quanto a **θ**. Maher argumenta que, embora as variáveis sejam discretas e sua distribuição deva ser uma forma multivariada da Poisson, a Normal multivariada é uma boa aproximação para contextos de redes reais. Uma dificuldade adicional reside no fato de que as proporções p_{kl} dependem da alocação do tráfego na rede, isto é, das rotas escolhidas pelos usuários, que normalmente não são conhecidas. Como simplificação, Maher admite uma rede não congestionada, aplicando uma alocação do tipo proporcional "tudo-ou-nada", independente dos valores de θ_l, com apenas uma rota carregada entre cada par *l*; desse modo, p_{kl} = 1 se o arco *k* faz parte da rota, e

p_{kl} = 0 caso contrário.

Na formulação bayesiana, representa-se o vetor $\mathbf{y} = (y_1, y_2, ..., y_m)$ como $\mathbf{y} = \mathbf{A\theta} + \boldsymbol{\varepsilon}$, em que **A** é a matriz *m* por *n* de proporções p_{kl}, e $\boldsymbol{\varepsilon}$ é o vetor de erros, o qual admite-se seguir uma distribuição Normal multivariada com média **0** e matriz de covariância **Σ.** Logo, a função verossimilhança de **y**, ou seja, $p(\mathbf{y}|\ \boldsymbol{\theta})$, é uma Normal multivariada com média **Aθ** e matriz de covariância **Σ**. Maher prova que a distribuição *a posteriori* de **θ**, ou seja, $p(\boldsymbol{\theta}|\mathbf{y})$, é também uma distribuição Normal multivariada, com média $\boldsymbol{\mu}_1$ e matriz de covariância $\mathbf{V}_1$. As Equações (**11**) e (**12**), desenvolvidas no artigo, indicam a relação entre os hiperparâmetros *a priori* $\boldsymbol{\mu}_0$ e $\mathbf{V}_0$, e *a posteriori* $\boldsymbol{\mu}_1$ e $\mathbf{V}_1$.

$$\boldsymbol{\mu}_1 = \boldsymbol{\mu}_0 + \mathbf{V}_0\mathbf{A}^{\mathbf{t}}(\boldsymbol{\Sigma} + \mathbf{A}\mathbf{V}_0\mathbf{A}^{\mathbf{t}})^{-1}(\mathbf{y} - \mathbf{A}\boldsymbol{\mu}_0) \quad (11)$$

$$\mathbf{V}_1 = \mathbf{V}_0 - \mathbf{V}_0\mathbf{A}^{\mathbf{t}}(\boldsymbol{\Sigma} + \mathbf{A}\mathbf{V}_0\mathbf{A})^{-1}\mathbf{A}\mathbf{V}_0 \quad (12)$$

3.2 Modelo de Tebaldi e West (1999)

O trabalho de Tebaldi e West (1999) segue uma direção diferente daquela traçada por Maher (1983), apoiando-se na abordagem estatística de Vardi (1996). Os autores consideram que, para cada par OD na rede, indexado por *l*, o número de viagens é dado por uma variável aleatória não observável θ_l, que segue uma distribuição de Poisson com taxa média λ_l. Dadas então as observações de volumes $\mathbf{y} = (y_1, y_2, ..., y_m)$ em *m* links da rede (a partir de medidas de volumes isentas de erro), a Equação (13), uma versão matricial e sem a componente de erro da Equação (**10**), relaciona os volumes observados **y** ao número de viagens $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_n)$.

$$\mathbf{y} = \mathbf{A\theta} \quad (13)$$

A matriz **A** identifica as proporções p_{kl} de viagens no par *l* cujas rotas incluem o arco *k*. No artigo, os autores admitem que cada par OD utiliza apenas uma rota; portanto, a matriz **A** é binária pois, para um dado arco *k* e par OD *l*, ou todas as viagens passam pelo arco, caso em que p_{kl} = 1, ou o arco não recebe viagens, caso em que p_{kl} = 0. Como os autores admitem que o número de pares OD é maior que o número de arcos com volumes observados, a Equação (**13**) implica que há componentes de **θ** que são dependentes. Identificando por $\boldsymbol{\theta}_1$ o vetor das *m* componentes dependentes e $\mathbf{A}_1$ a matriz correspondente às colunas de $\boldsymbol{\theta}_1$ em **A**, e por $\boldsymbol{\theta}_2$ o vetor das (*n*-*m*) componentes independentes e $\mathbf{A}_2$ a matriz correspondente às colunas de $\boldsymbol{\theta}_2$ em **A**, a Equação (**14**) expressa $\boldsymbol{\theta}_1$ em função de $\boldsymbol{\theta}_2$.

$$\boldsymbol{\theta}_1 = \mathbf{A}_1^{-1}(\mathbf{y} - \mathbf{A}_2\boldsymbol{\theta}_2) \quad (14)$$

Portanto, os autores consideram a estimação bayesiana tanto do número de viagens **θ,** como dos parâmetros **λ** das distribuições de Poisson, ou seja, tentam estimar a distribuição *a posteriori* $p(\boldsymbol{\theta},\boldsymbol{\lambda}|\mathbf{y})$. Eles adotam um esquema de simulação MCMC conhecido como amostragem de Gibbs, o qual realiza amostragens recursivamente para os valores de **θ** e **λ** a partir das distribuições *a posteriori* condicionais. O método proposto consiste nos seguintes passos:

1. Fixe valores iniciais para o vetor $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_n)$ (possivelmente valores aleatórios);
2. Amostre valores para as taxas médias $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, ..., \lambda_n)$ a partir da distribuição *a posteriori* $p(\lambda_j|\theta_j)$. A distribuição a posteriori $p(\lambda_j|\theta_j)$ é proporcional a $p(\theta_j|\lambda_j)p(\lambda_j)$, em que $p(\theta_j|\lambda_j)$ é a função de verossimilhança de θ_j, a qual possui uma forma de Poisson, e $p(\lambda_j)$ é a distribuição *a priori* de λ_j. Caso se admita uma distribuição Gama para $p(\lambda_j)$, por exemplo, a distribuição $p(\lambda_j|\theta_j)$ também será Gama, tal como ilustrado no item 2.3. Caso se utilize

uma distribuição *a priori* cuja *a posteriori* seja difícil de amostrar, pode-se utilizar o algoritmo de Metropolis-Hastings.

3. Amostre novos valores para $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_n)$ condicionais nos valores $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, ..., \lambda_n)$ e $\mathbf{y} = (y_1, y_2, ..., y_m)$ por meio da Equação (**15**).

$$p(\theta_j|\boldsymbol{\theta}_{2,-j}, \boldsymbol{\lambda}, \mathbf{y}) \propto \frac{\lambda_j^{\theta_j}}{\theta_j!} \prod_{a=1}^{m} \frac{\lambda_a^{\theta_a}}{\theta_a!} \quad (15)$$

   Na qual $\boldsymbol{\theta}_{2,-j}$ denota o vetor de componentes independentes de $\boldsymbol{\theta}$ excluindo-se a componente θ_j, e o índice j assume valores no conjunto de índices das componentes independentes, ou seja, j = m+1, m+2, ..., n (os índices j = 1, 2, ..., m identificam as componentes dependentes). Para gerar a amostra, deve-se empregar o algoritmo de Metropolis-Hastings, o qual possibilita a obtenção de amostras sem a necessidade do cálculo da constante de normalização (por esta razão (**15**) é escrita como uma relação de proporcionalidade), e o vetor $\boldsymbol{\theta}_{2,-j}$ contém as amostras mais atuais de suas componentes.
4. Repita os passos a partir de 2 para gerar uma nova amostra de $\boldsymbol{\theta}$ e $\boldsymbol{\lambda}$.

Os autores realizaram experimentos em redes artificiais e redes reais de pequeno porte, chegando a resultados melhores que os obtidos por Vardi (1996), atribuindo a diferença ao uso da informação nas distribuições *a priori*.

3.3. Desenvolvimentos recentes

Entre as contribuições recentes, destacam-se a sequência de trabalhos de Hazelton (2001, 2008, 2010) e o de Castillo *et al*. (2008). Enquanto Tebaldi e West (1999) admitem rotas únicas entre pares OD, a modelagem de Hazelton vai além e considera todas as rotas possíveis em um par OD. Ademais, o autor incorpora o processo de escolha das rotas por meio de um modelo Logit, obtendo distribuições *a posteriori* para as proporções de escolha das rotas. Hazelton ilustra suas aplicações em uma rede de pequeno porte na cidade de Leicester, Inglaterra. No entanto, em redes de grande porte, a enumeração das rotas pode ser um procedimento computacionalmente intensivo, requerendo técnicas para a redução do número de rotas consideradas. Castillo *et al*. (2008) seguem uma abordagem diferente, propondo um modelo em dois níveis para a estimação simultânea da matriz OD e a alocação dos fluxos na rede. Os autores propõem uma representação do problema por meio de uma rede bayesiana, empregando o conceito de equilíbrio estocástico do usuário no modelo de alocação, produzindo distribuições *a posteriori* para os parâmetros.

4. QUESTÕES A SEREM INVESTIGADAS NESTA LINHA DE PESQUISA

Como apresentado na introdução deste artigo, a principal dificuldade na determinação dos padrões de fluxo entre origens e destinos em uma rede de transportes, a partir de contagens volumétricas, reside no fato de que o número de pares OD em uma rede é normalmente muito superior ao número de arcos para os quais há observações de volume de tráfego. Quando se admite a abordagem de reconstrução da matriz OD sintética, essa dificuldade se manifesta por meio da multiplicidade de matrizes que possivelmente geraram os fluxos observados, levando o analista a adotar um critério para a determinação de uma matriz mais provável. Entre os critérios mais utilizados estão o critério da máxima entropia, ou, equivalentemente, da mínima informação, e o critério de minimização de erros quadráticos. Por outro lado, na abordagem estatística, o objetivo é estimar os parâmetros da distribuição de probabilidade das viagens, a partir de uma amostra de volumes. Com base no modelo estimado, podem-se obter medidas como o número médio de viagens entre pares OD e suas variâncias.

A primeira questão que vem à tona na estimação de matrizes OD sintéticas, tanto sob a

perspectiva frequentista como sob a bayesiana, é a especificação dos modelos estatísticos que descrevem as variáveis aleatórias e seus pressupostos. As aplicações descritas na literatura admitem modelos paramétricos matematicamente convenientes – Normal no caso de Maher (1983) e Poisson no caso de Tebaldi e West (1998) – mas que podem não corresponder à realidade. Nesse ponto, faz-se necessário o esforço de melhor compreensão do fenômeno da variabilidade das viagens, tanto sob a dimensão espacial como sob a temporal, para que se possam especificar modelos válidos e úteis, ou seja, capazes de representar a realidade em um nível de precisão aceitável, e de serem utilizados no auxílio à tomada de decisão em sistemas de transportes. Essa questão também se apresenta na abordagem bayesiana, sob forma diversa, na *determinação das distribuições a priori*. A dificuldade consiste em como traduzir em distribuições de probabilidade o conhecimento intuitivo dos pesquisadores e profissionais acerca dos fenômenos de transporte. Neste artigo, foram dados exemplos em que o analista atribui probabilidades diretamente aos valores dos parâmetros, ou especifica valores como a média e o desvio-padrão, possibilitando a determinação de uma distribuição *a priori* específica. No caso da matriz sintética, uma matriz-semente pode estar disponível, mas não está claro qual a melhor forma de transformá-la em uma distribuição *a priori*.

Outra questão de pesquisa ainda pouco explorada diz respeito à *aplicação da estimação bayesiana em redes congestionadas*, para as quais as rotas escolhidas em cada par OD dependem dos volumes alocados nos arcos da rede. Em outras palavras, as rotas efetivamente utilizadas em redes congestionadas resultam do comportamento dinâmico da interação dos usuários com o sistema de transportes, e, portanto, não se pode predeterminar as rotas, como ocorre em redes não congestionadas. Nos modelos de otimização, o fenômeno da interdependência entre as rotas e os volumes (carregamento da rede) tem sido modelado por meio de princípios de equilíbrio, como o primeiro princípio de Wardrop (1952). No caso dos modelos estatísticos bayesianos, ainda são incipientes as tentativas de incorporar princípios de equilíbrio ou escolha de rotas.

Ainda com relação à estimação dos fluxos OD, a *tendência à superestimação da quantidade de viagens em pares OD de baixo fluxo que compartilham arcos com pares OD de grande fluxo*, questão esta apontada por Tebaldi e West (1998), demanda investigação subseqüente em virtude de suas implicações práticas, as quais podem conduzir planejadores de transporte a recomendarem investimento desnecessário em infraestrutura. Nesse ponto, a abordagem bayesiana pode contribuir a partir do uso do conhecimento *a priori* de analistas sobre o comportamento do fenômeno das trocas de viagens entre zonas produtoras e atratoras na região em estudo. Vale também destacar que a inferência bayesiana pode facilitar a incorporação de outros tipos de informação adicional na estimação da matriz sintética, além dos volumes observados nos arcos e da matriz semente, tais como: a) matrizes parciais obtidas, por exemplo, a partir de pesquisas de placas em cordões internos (especialmente com o advento de sistemas de identificação automática de veículos), permitindo uma melhor modelagem do processo de escolha de rotas pelos usuários; b) contagens diretas de fluxos entre pares de zonas, assim como contagens dos totais de viagens entrando e saindo de cada zona; c) modelos desagregados ou de demanda direta que expliquem o comportamento na distribuição e na escolha modal das viagens.

5. CONSIDERAÇÕES FINAIS

Neste artigo, foram expostos conceitos básicos da Estatística Bayesiana e aplicações iniciais na modelagem das origens e destinos de viagens em redes de transportes. Ressaltaram-se as diferenças fundamentais existentes entre as abordagens frequentista e bayesiana da Estatística, e um exemplo ilustrativo foi desenvolvido com o objetivo de explicitar as implicações dessas

diferenças na estimação de parâmetros em estudos estatísticos em transportes. Apresentaram-se modelos bayesianos propostos na literatura para a estimação de matrizes OD sintéticas, e por fim, discutiram-se questões de pesquisa que representam desafios para o uso efetivo de modelos estatísticos bayesianos na modelagem das origens e destinos das viagens. Por fim, cabe ressaltar as possibilidades, ainda pouco exploradas, de aplicação da abordagem bayesiana em outras etapas do esforço de modelagem das relações entre oferta e demanda no processo de planejamento estratégico de sistemas de transportes urbanos e regionais, diretamente na calibração de modelos de geração, distribuição e escolha modal das viagens.

REFERÊNCIAS BIBLIOGRÁFICAS

Abrahamsson, T. (1998) Estimation of origin-destination matrices using traffic counts – a literature survey. *International Institute for Applied Systems Analysis*. Technical Report.

Bolstad, W. M. (2007) *Introduction to Bayesian Statistics*. 2nd Ed. John Wiley & Sons.

Cascetta, E.; Nguyen, S. (1988) A unified framework for estimating or updating origin/destination matrices from traffic counts. *Transportation Research B*, v.22, p.437-455.

Castillo, E.; Menéndez, J.M.; Sánchez-Cambronero, S. (2008). Predicting traffic flow using Bayesian networks. *Transportation Research B*. v.42, p.482-509.

Diaconis, P. (2009). The Markov Chain Monte Carlo revolution. *Bulletin of the American Mathematical Society*, v.46, n.2, p.179-205.

Gelman, A.; Carlin, J.B.; Stern, H.S.; Rubin, D.B. (2003) *Bayesian Data Analysis*. 2nd Ed. Chapman-Hall.

Hazelton, M.L. (2001) Inference for origin-destination matrices: estimation, prediction and reconstruction. *Transportation Research B*. v.35, p.667-676.

Hazelton, M.L. (2008) Statistical inference for time varying origin-destination matrices. *Transportation Research B*. v.42, p.542-525.

Hazelton, M.L. (2010) Bayesian inference for network-based models with a linear inverse structure. *Transportation Research B*. v.44, p.674-685.

Maher, M.J. (1983) Inferences on trip matrices from observations on link volumes: a Bayesian statistical approach. *Transportation Research B*. vol.17, n.6, p.435-447.

Nguyen, S. (1977) Estimating an OD matrix from network data: A network equilibrium approach. *Publication 87. Centre de Recherche sur les Transports, Université de Montreal*.

Nguyen, S. (1984) Estimating origin-destination matrices from observed flows. *Transportation Planning Models*, v.1, p.363-380.

Ortúzar, J.D.; Willumsen, L.G. *Modelling Transport*. 2nd Ed. John Wiley & Sons, 1994.

Rakha, H.; Paramahamsan, H.; Van Aerde M. (2005), Comparison of static maximum likelihood origin-destination formulations. *Transportation and Traffic Theory: Flow, Dynamics and Human Interaction* , Proceedings of the 16th International Symposium on Transportation and Traffic Theory (ISTTT16), p.693-716.

Robillard, P. (1975) Estimating the O-D matrix from observed link volumes. *Transportation Research*, vol.9, p.123-128.

Tebaldi, C.; West, M. (1998) Bayesian inference on network traffic using link count data. *Journal of the American Statistical Association*, v.93, n.442.

Timms, P. (2001). A philosophical context for methods to estimate origin-destination trip matrices using link counts. *Transport Reviews*. v.21, n.3, p.269-301.

Van Zuylen, H.J.; Willumsen, L.G. (1980) The most likely trip matrix estimated from traffic counts. *Transportation Research B*, vol.14, p.281-29.

Vardi, Y. (1996) Network tomography: estimating source-destination traffic intensities from link data. *Journal of the American Statistical Association*. v.91, n.433, p.365-377.

Wardrop, J. G. (1952) *Some Theoretical Aspects of road Traffic Research*. Proc. Inst, Civil Engineers, Part 2, p. 325-378.

Willumsen, L.G. (1981) Simplified transport models based on traffic counts. *Tranportation*, v.10, p.257-278.