Instituto de MATEMÁTICA E ESTATÍSTICA UFRGS

DEPARTAMENTO DE ESTATÍSTICA
E-mail: dest@mat.ufrgs.br

Trabalho de Conclusão de Curso

Viés de amostragem em análises filogeográficas

Vitória Silva Garcia

30 de agosto de 2024

Vitória Silva Garcia

Viés de amostragem em análises filogeográficas

Trabalho de Conclusão apresentado à comissão de Graduação do Departamento de Estatística da Universidade Federal do Rio Grande do Sul, como parte dos requisitos para obtenção do título de Bacharel em Estatística.

Orientadora: Profa. Dra. Gabriela Bettella Cybis

Porto Alegre
Agosto de 2024

Vitória Silva Garcia

Viés de amostragem em análises filogeográficas

Este Trabalho foi julgado adequado para obtenção dos créditos da disciplina Trabalho de Conclusão de Curso em Estatística e aprovado em sua forma final pela Orientadora e pela Banca Examinadora.

Orientadora:_____
Profa. Dra. Gabriela Bettella Cybis, UFRGS
Doutora pela Universidade Federal do Rio Grande do Sul, Porto Alegre, RS

Banca Examinadora:

Profa. Dr. Taiane Schaedler Prass, UFRGS
Doutora em Matemática pela Universidade Federal do Rio Grande do Sul

Mestre Felipe Grillo Pinheiro, Data Scientist na National Institute of Agricultural Botany (NIAB) Cambridge, UK
Mestre em Estatística pela Universidade Federal do Rio Grande do Sul

Porto Alegre
Agosto de 2024

Agradecimentos

Aos meus pais, pelo cuidado, incentivo e amor incondicional ao longo de toda a minha vida.

À minha irmã, por todo o apoio, conselhos e suporte em todos os momentos, especialmente durante o período da graduação.

Agradecimento especial à professora Dra. Gabriela Cybis, pelos valiosos ensinamentos, pela compreensão em momentos difíceis e pelo constante incentivo ao longo deste trabalho.

À professora Dra. Adriana Neumann, pelas inúmeras conversas, por me apresentar ao curso de estatística e por me guiar em diversos momentos da minha trajetória acadêmica.

Ao Mestre Felipe Vargas pelo apoio e ensinamentos compartilhados.

Aos meus amigos, que sempre estiveram ao meu lado nos momentos de angústia acadêmica e que tornaram os meus dias mais leves.

E aos professores do departamento de estatística, por todo o conhecimento compartilhado, pelas experiências enriquecedoras e pelo apoio ao longo dessa jornada.

Resumo

Análises filogeográficas combinam dados de sequências genéticas, locais geográficos e datas de amostragem. Por meio dessas análises, é possível estimar a migração do vírus entre diferentes locais geográficos e reconstruir o processo espacial de dispersão do vírus. Com a filogeografia bayesiana, é possível realizar estudos sobre a dispersão viral e, através de um modelo estocástico, mapear a propagação dos vírus analisados. É importante destacar que esses modelos filogeográficos bayesianos são muito afetados por vieses de amostragem decorrentes de fatores que não estão relacionados ao processo de dispersão, como heterogeneidades espaciais e temporais devido à escassez de recursos nas localidades para o sequenciamento do vírus. Um exemplo de viés amostral está relacionado a amostras não coletadas de maneira fiel à verdadeira prevalência das localizações. Esta pesquisa visa estudar as análises filogeográficas, o que ocorre com a estimação dos parâmetros quando a amostra aleatória possui um viés amostral e a mensuração dos efeitos desse viés amostral. Para isso, utilizaremos um modelo de cadeias de Markov na árvore filogenética, simulando cenários com e sem viés amostral. Utilizaremos o método *Bias-correcting Subsampling Trait Model* (BSTM) com o objetivo de corrigir o viés amostral, ponderando as subamostras com base em informações externas relacionadas às verdadeiras frequências populacionais e sem o descarte de dados. Compararemos os métodos de estimação por máxima verossimilhança, inferência bayesiana através do algoritmo MCMC, considerando o modelo simples de cadeia de Markov e inferência bayesiana através do método com *Bayesian Stochastic Search Variable Selection* (BSSVS).

Observou-se para os diferentes cenários simulados indícios que o método BSTM é vantajoso para corrigir vieses amostrais na estimação dos parâmetros da matriz de taxas para muitos dos cenários de simulação considerados, principalmente nos métodos Bayesianos.

Palavras-Chave: Filogeografia, Inferência Filogeográfica Bayesiana, Viés de amostragem, Monte Carlo por cadeias de Markov.

Abstract

Phylogeographic analyses combine data from genetic sequences, geographical locations, and sampling dates. Through these analyses, it is possible to estimate the migration of the virus between different geographical locations and reconstruct the spatial process of virus dispersal. With Bayesian phylogeography, it is possible to conduct studies on viral dispersions and, through a stochastic model, map the propagation of the viruses. It is important to highlight that these Bayesian phylogeographic models are significantly affected by sampling biases arising from factors unrelated to the dispersal process, such as spatial and temporal heterogeneities due to the scarcity of resources in localities for virus sequencing. An example of sampling bias is related to samples not being collected in a manner faithful to the true prevalence of the locations. This research aims to study phylogeographic analyses, what happens to parameter estimation when the random sample has sampling bias, and measure of the effects of this sampling bias. For this purpose, we use a Markov chain model on the phylogenetic tree, simulating scenarios with and without sampling bias. We use the Bias-correcting Subsampling Trait Model (BSTM) to correct the sampling bias by subsampling and weighting samples based on external information related to true population frequencies, without discarding any data. We compare the methods of maximum likelihood estimation, Bayesian inference through the MCMC algorithm, considering the simple Markov chain model and the model with Bayesian Stochastic Search Variable Selection (BSSVS).

Showd evidence of being advantageous for correcting sampling biases in the estimation of the rate matrix parameters for many of the simulation scenarios considered, especially for the Bayesian methods.

Keywords: Phylogeography, Bayesian phylogeographic inference, Sampling bias, Markov chain Monte Carlo.

Sumário

Lista de Figuras

Lista de Tabelas

1 Introdução

Estudos filogeográficos especificamente relacionados a inferência filogeográfica bayesiana são de grande importância para pesquisas epidemiológicas, como por exemplo na descoberta da origem de propagação de um vírus (Lemey et al., 2009). Além de compreendermos o processo filogenético que representa as relações evolutivas entre diferentes espécies, organismos ou genes, mostrando como eles divergem ou se aproximam de um ancestral comum, queremos estudar como os vírus migram entre diferentes regiões geográficas e como essas migrações ocorrem ao longo do tempo.

Para realizarmos essa pesquisa, utilizaremos modelos filogeográficos que fazem o uso de dados de sequências genéticas, locais geográficos e datas de amostragem. Com esses modelos é possível, por exemplo, reconstruir a história da disseminação geográfica de um vírus, usando as sequências genéticas do mesmo coletadas em diferentes momentos e locais (Kalkauskas et al., 2021; Bloomquist et al., 2010).

Das abordagens utilizadas para os modelos de filogeografia, iremos estudar os modelos de Cadeias de Markov para filogeografia discreta. A abordagem discretizada é utilizada quando as amostras são agrupadas com base em a sua localização geográfica, sendo que esses dados geográficos compõem uma unidade geográfica, como um país ou uma cidade (Kalkauskas et al., 2021).

Nesse trabalho utilizaremos a abordagem discreta, em que os estados serão constituídos de informações discretizadas referentes às localizações das amostras. Para tanto, com as informações das sequências genéticas e localizações da amostra, temos que a cadeia de Markov evolui ao longo da árvore filogenética relacionando os indivíduos da amostra e com isto originando as localizações geográficas dos indivíduos.

Apesar das muitas aplicações realizadas com essa abordagem, como, por exemplo, SARS-CoV-2 (Lemey et al., 2020) e no caso Ebola (De Maio et al., 2015) a coleta de amostras com ausência de viés é muito difícil, pois para isso seria necessário o acesso de todo espaço geográfico. Assim, um grande problema encontrado no processo de inferência para esses modelos são os vieses de amostragem, que ocorrem pela falta de uma coleta amostral nas localizações geográficas de maneira proporcional à sua prevalência (De Maio et al., 2015). Com o intuito de sanar esse problema, Vargas (2023) propôs uma nova abordagem chamada *Bias-correcting Subsampling Trait Model* (BSTM) que utiliza informações externas que representam a verdadeira frequência amostral, como o número de casos em cada estado. Nessa nova abordagem para a correção do viés, é considerada a média sobre todas as possíveis subamostras da árvore, sendo equivalente a considerar um modelo que combina as diferentes árvores subamostradas.

Assim sendo, os objetivos desse projeto são estudar as os métodos de análises filogrográficas com o propósito de mensurar os efeitos que os vieses de amostragem possuem sobre a reconstrução filogeográfica e seus parâmetros.

Para atingirmos o nosso objetivo, analisaremos as simulações com o método de correção de viés de amostragem e com o método padrão, ou seja, sem a correção do viés, para os diferentes cenários propostos. Utilizaremos métodos de estimação por máxima verossimilhança, inferência bayesiana com a cadeias de Markov padrão e inferência bayesiana com a cadeias de Markov e *Bayesian Stochastic Search Variable Selection* (BSSVS). Em suma, queremos concluir para quais casos propostos o método de correção de viés de amostragem foi satisfatório corrigindo o viés de amostragem.

2 Metodologia

2.1 Árvore filogenética

Para representar as relações evolutivas entre os organismos que surgem de um ancestral comum é possível utilizar árvores filogenéticas. Uma árvore é composta pelos seguintes elementos: os nós externos que representam as unidades das quais a árvore foi inferida, ou seja, a amostra; os ramos ou arestas, esses ligam os nós da árvore; os nós internos ou nós ancestrais que correspondem ao último ancestral comum dos nós e ramos que decorrem dele; e a raiz da árvore que representa o ancestral comum mais recente de toda a árvore (Moreira, 2015).

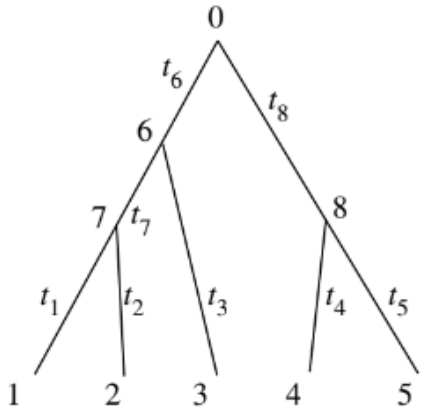

Figura 2.1: Exemplo de Árvore Filogenética

A Figura 2.1 ilustra como pode ser estruturada uma árvore filogenética. O exemplo de árvore contém cinco nós externos: 1, 2, 3, 4, 5, respectivamente. Nesta árvore, os nós ancestrais são, respectivamente, 0 (raíz), 6,7,8. Os ramos que levam aos nós descendentes $i \in \{1, 2, ..8\}$ são chamados de t_i, e estes, por sua vez, podem ser interpretados como a distância temporal ou genética de um nó para o outro (Yang, 2006).

2.2 Cadeias de Markov a tempo contínuo na filogeografia

O processo estocástico $\{X_t\}_{t\geq 0}$ é uma cadeia de Markov a tempo contínuo se, para quaisquer tempos $t, g_0 < g_1 < g_2 \cdots < g_n < g$ positivos e quaisquer $i_0, ..., i_n, i, j$ no espaço de estados discreto $\{G\}$, temos

$$P(X_{t+g} = j | X_g = i, X_{g_n} = i_n, \cdots, X_{g_0} = i_0) = P(X_{t+g} = j | X_g = i).$$

Ou seja, nos processos de Markov as probabilidades que estão relacionadas com os valores futuros do processo dependem somente do estado atual e não de todos os estados anteriores. Portanto, os eventos passados são condicionalmente independentes dos eventos futuros dado o estado atual. Os nossos processos também serão assumidos como estacionários, portanto

$$P(X_{t+g} = j | X_g = i) = P(X_t = j | X_0 = i).$$

Para compreender o processo espacial que viabiliza o deslocamento desde o ancestral comum da árvore até a mais recente observação dos dados, é interessante entender os momentos e locais discretos das migrações. Nesse sentido, a compreensão das cadeias de Markov em tempo contínuo ao longo da árvore se torna indispensável. Na filogeografia, é razoável assumir que as probabilidades de transição para um novo local dependem apenas da localização atual e não de todas as localizações ocorridas no passado. Nesse contexto a geografia pode ser dividida em um número finito de locais discretos, como por exemplo cidades ou países, esse espaço de estados pode ser definido como $\{G_k\}_{k=1}^K$. Assim, nos N nós externos de uma filogenia $\boldsymbol{F}$, podemos registrar através da variável aleatória X, de maneira discretizada, as N localizações da amostra em que $\boldsymbol{X} = (X_1, ..., X_N)$, com $X_i \in \{G_k\}_{k=1}^K$, para todo $i \in \{1, \cdots, N\}$.

A cadeia de Markov a tempo contínuo que rege essa dispersão espacial pode ser caracterizada pela matriz de taxas infinitesimais $\boldsymbol{\Lambda}$, $K \times K$, que satisfaz

$$\boldsymbol{\Lambda} = \begin{bmatrix} \lambda_{11} & \cdots & \lambda_{1K} \\ \vdots & \ddots & \vdots \\ \lambda_{K1} & \cdots & \lambda_{KK} \end{bmatrix}, \quad \text{com} \quad \lambda_{ij} \geq 0, \text{ se } i \neq j, \quad \text{e} \quad \sum_{j=1}^{K} \lambda_{ij} = 1. \tag{2.1}$$

Ou seja, $\boldsymbol{\Lambda}$ possui valores maiores ou iguais a zero fora da diagonal e as suas linhas somam zero.

Utilizando a propriedade de Chapman-Kolmogorov é possível obter as probabilidades de transição no tempo finito. Seja $P(t)$ a matriz cuja jk - ésima entrada é dada por

$$p_{jk}(t) = P(\boldsymbol{X}(t) = G_k | \boldsymbol{X}_0 = G_j), \quad \text{para } 1 \leq j, k \leq K, \tag{2.2}$$

então, a matriz $\boldsymbol{\Lambda}$ e a matriz $P(t)$ satisfazem a seguinte relação

$$\left\{p_{jk}(t)\right\} = \boldsymbol{P}(t) = e^{\boldsymbol{\Lambda} t}, \quad \forall t \geq 0. \tag{2.3}$$

Para realizar o cálculo das probabilidades de transição no tempo finito é utilizada a exponenciação de matrizes, que nesse caso é realizada por uma decomposição

espectral de $\mathbf{\Lambda}$ que é restrito às matrizes de taxas infinitesimais que produzem autovalores e autovetores reais. Com isso, consideremos a seguinte parametrização de $\mathbf{\Lambda}$

$$\mathbf{\Lambda} = \boldsymbol{S\Pi}, \tag{2.4}$$

em que $\boldsymbol{S} = \{s_{jk}\}$ é uma matriz simétrica $K \times K$ de taxas e $\mathbf{\Pi}$ é a matriz diagonal das frequências de equilíbrio do modelo com elementos não nulos $\pi = (\pi_1, ..., \pi_K)$. Sendo π a distribuição estacionária do processo, com probabilidades que somam 1.

Definidas as cadeias de Markov a tempo contínuo na filogeografia, avançamos para o próximo passo, que é inferência nesse modelo de Markov.

2.3 Função de Verossimilhança na árvore

A função de verossimilhança é muito utilizada na estatística para realizar inferência a partir de um conjunto de dados. Essa função de verossimilhança pode ser calculada como a probabilidade de observar a base de dados, dado determinados parâmetros (Yang, 2006). No contexto da filogeografia, vamos assumir que a migração de uma linhagem para uma localidade é independente das migrações de outras linhagens, dada a informação do mais recente ancestral comum. Para compreender o cálculo da função de verossimilhança, segue o exemplo abaixo.

Exemplo: Árvore com 5 nós externos

Com o objetivo de mostrar o cálculo da verossimilhança, neste exemplo, será utilizado como base uma árvore que contém observações em quatro localidades $\{G_k\}_{k=1}^{K} = \{A, B, C, D\}$ apresentada na Figura 2.2. As localidades representadas na figura são, respectivamente, A, B, C, D, D, representando os estados observados nos 5 nós externos (amostra). Nesta árvore, os nós ancestrais são, respectivamente, 0 (raíz), 6, 7, 8 e os comprimentos de ramos começam pelo comprimento mais próximo da raíz, respectivamente, $t_6, t_8, t_7, t_1, t_2, t_3, t_4, t_5$.

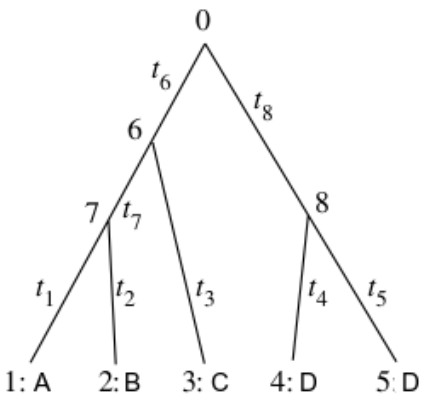

Figura 2.2: Exemplo Árvore na Filogeografia

Nesse caso, os estados observados nos nós externos são $\boldsymbol{X} = (A, B, C, D, D)$. Para esse exemplo, os parâmetros estudados são o comprimento dos ramos que são

os t_i e os parâmetros da matriz de taxas da cadeia de Markov apresentados em (2.4). O vetor de parâmetros será denotado por

$$\boldsymbol{\theta} = (t_1, t_2, t_3, t_4, t_5, t_6, t_7, t_8, \boldsymbol{S}, \boldsymbol{\Pi}). \tag{2.5}$$

Como assumimos a independência condicional de migrações nos diferentes ramos da árvore, se conhecêssemos o estado (localidade) dos nós ancestrais 0, 6, 7, e 8, a probabilidade conjunta seria o produto das probabilidades de cada observação individual de migração em cada ramo da árvore (Yang, 2006). Entretanto essa informação não é conhecida.

Note que os dados observados $\boldsymbol{X}$ podem resultar de qualquer combinação de estados ancestrais da cadeia. Essa incerteza sobre os nós ancestrais deve ser levada em consideração no cálculo, assim, a função de verossimilhança dada por

$$\begin{aligned} f(\boldsymbol{X}|\boldsymbol{\theta}) = \sum_{x_0\in\{G\}} \sum_{x_6\in\{G\}} \sum_{x_7\in\{G\}} \sum_{x_8\in\{G\}} & \pi_{x_0} p_{x_0x_6}(t_6) p_{x_6x_7}(t_7) p_{x_{7A}}(t_1) \\ & p_{x_{7B}}(t_2) p_{x_{6C}}(t_3) p_{x_0x_8}(t_8) p_{x_{8D}}(t_4) p_{x_{8D}}(t_5). \end{aligned} \tag{2.6}$$

deste modo, a probabilidade de observar $\boldsymbol{X}$ resulta na soma das probabilidades de todas as possibilidades de combinações de estados ancestrais. Temos que π_{x_0} é a probabilidade do estado x_0 da raiz na distribuição estacionária e, por exemplo, $p_{x_0x_6}(t_6)$ é a probabilidade de um estado x_0 se tornar x_6 em um tempo t_6.

É importante mencionar que o cálculo da função de verossimilhança na árvore não é trivial, pois a localização da raiz na árvore em estudo é desconhecida e, portanto, tratada como tal. Além disso, o processo em questão ocorre de forma condicionalmente independente nos ramos da árvore. Neste estudo, apenas os estados finais dos ramos são observados, enquanto os estados intermediários (nós internos) permanecem desconhecidos. Por esse motivo, no cálculo acima, é realizada a soma de todas as possibilidades de estados dos ramos internos (Yang, 2006).

Desse modo, o cálculo da função de verossimilhança apresenta um custo computacional muito alto. Com esses problemas, Felsenstein (Felsenstein, 1981) desenvolveu um algoritmo chamado "pruning algorithm" para calcular a verossimilhança nos estados discretos da árvore. O algoritmo é relevante, pois identifica fatores em comum e os calcula apenas uma vez. Desse modo, os cálculos realizados pela técnica de Felsenstein aumentam linearmente de acordo com o número de nós ou estados.

2.4 Inferência Bayesiana para o Modelo Filogeográfico

A inferência bayesiana é baseada na probabilidade à posteriori $p(\boldsymbol{\theta}|\boldsymbol{X})$, ou seja, a probabilidade (ou densidade) do parâmetro θ dado as localizações $\boldsymbol{X}$. O cálculo para encontrar o valor da posteriori se dá pela fórmula de Bayes,

$$p(\boldsymbol{\theta}|\boldsymbol{X}) = \frac{p(\boldsymbol{X}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\boldsymbol{X})}. \tag{2.7}$$

Nosso conhecimento de θ é dado pela distribuição à priori $p(\boldsymbol{\theta})$ e a verossimilhança marginal dos dados é definida por $p(\boldsymbol{X})$. Para o processo espacial geográfico a verossimilhança $p(\boldsymbol{X}|\boldsymbol{\theta})$ pode ser calculada pelo algoritmo de (Felsenstein, 1981).

A inferência bayesiana tem sido muito usada no contexto de filogenias e filogeografia (Layan et al., 2023). Uma das vantagens dessa abordagem é que ela permite a integração da incerteza filogenética e a incerteza dos parâmetros do modelo de cadeias de Markov. É importante mencionar que existem modelos de substituição de nucleotídeos que utilizam a mesma estrutura de cadeias de Markov, mas em que os estados correspondem a nucleotídeos como ACGT (Yang, 2006). Estes se tornam relevantes para considerar a incerteza filogenética.

Desse modo, podemos considerar duas diferentes estruturas para o modelo. Na primeira consideramos a árvore filogenética $\boldsymbol{F}$ fixa e utilizamos o modelo apenas para estimar os parâmetros do modelo de migração ou reconstruir o processo de dispersão geográfica sobre a árvore. É essa alternativa que utilizaremos nas simulações desse trabalho, pois constituem um modelo mais simples que permite isolar o processo geográfico.

Na segunda estrutura consideramos a árvore $\boldsymbol{F}$ como parâmetro desconhecido, e utilizamos dados de sequências genéticas e dados geográficos para inferir a própria árvore, além dos parâmetros do modelo de migração e sua reconstrução. Essa alternativa se adapta melhor a situações reais, em que a verdadeira $\boldsymbol{F}$ não é conhecida com precisão.

Realizar inferência com resultados analíticos para filogeografia é virtualmente impossível devido à complexidade do modelo, assim a inferência é feita por métodos computacionais de simulação. Devido ao custo computacional para $p(\boldsymbol{X})$ do cálculo da posteriori (2.7), é necessário utilizar Monte Carlo via Cadeias de Markov (MCMC). No MCMC as cadeias utilizadas são ergódicas e possuem distribuição de equilíbrio $p(\boldsymbol{\theta}|\boldsymbol{X})$, e portanto eventualmente convergem para a posteriori (Metropolis et al., 1953; Hastings, 1970). Por trabalharmos no contexto bayesiano, ao longo do trabalho, é necessário especificar as prioris para os parâmetros do modelo, são eles: $\boldsymbol{S}$, $\boldsymbol{\Pi}$ definidos na expressão (2.4).

2.4.1 Priori

Nesse trabalho consideramos uma priori uniforme para as frequências de equilíbrio da matriz diagonal $\boldsymbol{\Pi}$ cujas entradas não nulas são dadas por $\pi = (\pi_1, \cdots, \pi_K)$. Assim, temos

$$(\pi_1, \cdots, \pi_K) \sim Dirichlet(1, \cdots, 1). \tag{2.8}$$

Desse modo, à priori, todos os estados tem a mesma probabilidade de serem encontrados na distribuição de equilíbrio. E para a escolha da priori para a entrada S_{ij} matriz de taxas S tem-se

$$S_{ij} \sim Exp(1), \quad \text{para} \quad i, j \in 1, \cdots, K, \quad i \neq j. \tag{2.9}$$

2.4.2 Posteriori

Lemey et al. (2009) propôs um modelo que possibilita a integração conjunta das localizações $\boldsymbol{X}$ e os dados da sequência molecular $\boldsymbol{Y} = (Y_1, \cdots, Y_N)$, com Y_i representando a sequência genética (DNA ou RNA) de comprimento L do i-ésimo indivíduo da amostra. Esse modelo mostra uma maneira de incorporar a incerteza da filogenia $\boldsymbol{F}$, se a mesma não é observada, e o processo de substituição de bases de $\boldsymbol{Y}$.

Aqui tem-se um modelo filogenético em que assume-se que uma cadeia de Markov contínua no tempo, com parâmetros $\boldsymbol{\phi}$, gera $\boldsymbol{Y}$. Também assume-se que as cadeias para evolução molecular e dispersão geográfica são independentes dada a filogenia $\boldsymbol{F}$, e por consequência nos permite escrever a distribuição a posteriori como o produto

$$\begin{aligned} p(\boldsymbol{F}, \boldsymbol{\Pi}, \boldsymbol{S}, \boldsymbol{\phi} | \boldsymbol{X}, \boldsymbol{Y}) &\propto p(\boldsymbol{X}, \boldsymbol{Y} | \boldsymbol{F}, \boldsymbol{\Pi}, \boldsymbol{S}, \boldsymbol{\phi}) p(\boldsymbol{F}, \boldsymbol{\Pi}, \boldsymbol{S}, \boldsymbol{\phi}) \\ &\propto p(\boldsymbol{X} | \boldsymbol{F}, \boldsymbol{\Pi}, \boldsymbol{S}) p(\boldsymbol{Y} | \boldsymbol{F}, \boldsymbol{\phi}) p(\boldsymbol{F}) p(\boldsymbol{\Pi}) p(\boldsymbol{S}) p(\boldsymbol{\phi}). \end{aligned} \tag{2.10}$$

Note que as verossimilhanças do processo geográfico $p(\boldsymbol{X}|\boldsymbol{F}, \boldsymbol{\Pi}, \boldsymbol{S})$ e do processo de evolução molecular $p(\boldsymbol{Y}|\boldsymbol{F}, \boldsymbol{\phi})$ são calculadas pelo Pruning Algorithm (Felsenstein, 1981). Também são utilizadas prioris filogenéticas padrão na árvore $\boldsymbol{F}$ e nos parâmetros do modelo de substituição de bases $\boldsymbol{\phi}$ (Drummond, 2012).

Para o caso dessa pesquisa em que o $\boldsymbol{F}$ é fixado, a distribuição é reduzida a

$$p(\boldsymbol{\Pi}, \boldsymbol{S} | \boldsymbol{X}, \boldsymbol{F}) \propto p(\boldsymbol{X} | \boldsymbol{\Pi}, \boldsymbol{S}, \boldsymbol{F}) p(\boldsymbol{\Pi}) p(\boldsymbol{S}). \tag{2.11}$$

2.5 Bayesian Stochastic Search Variable Selection

O método Bayesian Stochastic Search Variable Selection (BSSVS) permite a descrição mais parcimoniosa do processo de difusão filogeográfica Lemey et al. (2009). Com base nesse contexto Lemey et al. (2009) descreve que para estudos de localizações geográficas pode haver muitos locais (estados) e cada táxon está em apenas uma localização, esperamos que a maioria das transições na cadeia de Markov ocorram raramente. Caso contrário, podem ocorrer estimativas equivocadas relacionadas a inferência das localizações ancestrais não observadas e na raiz da árvore. Esse método possibilita a determinação de quais taxas infinitesimais podem ser consideradas zero, dependendo das evidências nos dados, resultando em inferência mais eficiente das localizações ancestrais.

Para o BSSVS o vetor de parâmetros $\boldsymbol{\theta}$ especificado em 2.5 possui um novo parâmetro definido como uma matriz de variáveis binárias indicadoras $\boldsymbol{\delta}$ tal que

$$\boldsymbol{\delta} = \begin{bmatrix} \delta_{11} & \cdots & \delta_{1K} \\ \vdots & \ddots & \vdots \\ \delta_{K1} & \cdots & \delta_{KK} \end{bmatrix}. \tag{2.12}$$

As indicadoras dessa matriz multiplicam as entradas da matriz de taxas $\boldsymbol{\Lambda}$, de modo que no BSSVS temos

$$\boldsymbol{\Lambda} = \begin{bmatrix} S_{11}\pi_1\delta_{11} & \cdots & S_{1K}\pi_K\delta_{1K} \\ \vdots & \ddots & \vdots \\ S_{K1}\pi_1\delta_{K1} & \cdots & S_{KK}\pi_K\delta_{KK} \end{bmatrix}, \tag{2.13}$$

Portanto, o parâmetro $\boldsymbol{\theta}$ para o método BSSVS é definido por

$$\boldsymbol{\theta} = (t_1, t_2, t_3, t_4, t_5, t_6, t_7, t_8, \boldsymbol{S}, \boldsymbol{\Pi}, \boldsymbol{\delta}). \tag{2.14}$$

Para o método BSSVS também são definidas as prioris para os parâmetros do modelo. Além das prioris especificadas na seção 2.4.1, também é definida uma priori para a indicadora $\boldsymbol{\delta}$ definida como

$$\boldsymbol{\delta} \sim Bernoulli(0.4). \tag{2.15}$$

2.6 Correção de viés por subamostragem em modelo evolutivo

Nas análises de filogeográficas, a coleta de amostras não viesadas pode se tornar difícil de ser realizada devido à extensão geográfica do mapa e a necessidade de grandes esforços de amostragem. A coleta dos dados muitas vezes não é realizada de maneira proporcional às frequências populacionais e, com isso, a inferência estatística é afetada (Kalkauskas et al., 2021).

Uma forma de corrigir esse viés nas amostras filogeográficas, quando além dos dados há informações externas sobre as frequências populacionais, é subamostrar de acordo com essas frequências. Entretanto, esse processo nos obriga a descartar parte dos dados, e, se estamos em um cenários que as observações são escassas, descartar não se torna uma opção.

Com o intuito de resolver esse problema Vargas (2023) propôs um novo modelo que tem como objetivo unir todos os possíveis conjuntos de dados subamostrados, em que cada observação tem uma probabilidade de ser incluída na amostra. O método utiliza uma probabilidade conectada às verdadeiras frequências populacionais sem fazer o descarte dos dados. Além disso, também foi proposto um cálculo de verossimilhança filogenética que considera para cada nó externo da árvore um peso de inclusão que se baseia no uso de informações externas sobre a distribuição dos K estados na população.

As informações externas de como os K estados estão distribuídos na população são definidas como f_k, para $k \in \{1, \cdots, K\}$.

Como em algumas situações há uma dificuldade da obtenção de um número ideal de observações para cada um dos K estados, Vargas (2023) considera em seu método todas as possíveis subamostras dos dados originais. Ou seja, para cada subamostra corresponde uma árvore filogenética que é uma subárvore da filogenia da amostra completa F, em que os nós externos correspondem às observações de uma subamostra de todas as possíveis combinações da árvore.

O modelo considera que cada nó externo i com estado observado k terá uma probabilidade de inclusão igual à γ_i, gerada a partir da função

$$\gamma_i = \frac{\min_\ell(\frac{n_\ell}{f_\ell}) \times f_k}{n_k}, \tag{2.16}$$

em que n_k representa o número de nós externos com estado k na amostra.

O método considera um modelo de mistura das de todas as possíveis subamostras dos dados, de modo que cada nó externo $i \in \{1, ..., N\}$ aparece com probabilidade γ_i nas subamostras. Para fazer inferência sobre esse modelo, gostaríamos de considerar a média ponderada sob todas as possíveis subamostras. Tal média poderia ser obtida por MCMC, o que implicaria em um custo computacional considerável. Vargas (2023) apresenta um resultado que permite o cálculo analítico dessa verossimilhança média sob as subamostras.

Os valores calculados pela equação (2.16) são utilizados para encontrar pesos α_{ik} para cada nó externo $i \in \{1, ..., N\}$ e estado $k \in \{1, ..., K\}$. Seja d_{x_i} o estado do nó ancestral ao nó i. Assim, a verossimilhança desse modelo de mistura pode ser calculada pela seguinte equação

$$p(\boldsymbol{X}^*|\boldsymbol{\Lambda}, \boldsymbol{F}) = \sum_{x_{2N-1}} \cdots \sum_{x_1} \pi_{x_{2N-1}} \left(\prod_{i=1}^{2N-1} p_{d_{x_i}, x_i} \right) \left(\prod_{i=1}^{N} \alpha_{ik} \right), \tag{2.17}$$

em que

$$\begin{cases} \alpha_{ik} = (1 - \gamma_i), & X_i \neq k \\ \alpha_{ik} = 1, & X_i = k \end{cases}. \tag{2.18}$$

Aqui, $p(\boldsymbol{X}^* | \boldsymbol{\Lambda}, \boldsymbol{F})$ representa a verossimilhança média sob essa mistura ponderada de subamostras. Note que essa verossimilhança pode ser computada de modo eficiente por meio do algoritmo de Pruning.

3 Simulações

3.1 Implementação das simulações

A implementação computacional deste trabalho foi feita em linguagem de programação R, no software RStudio, versão 4.2.1 (R Core Team, 2022), utilizando os pacotes ape (Paradis and Schliep, 2019), dplyr (Wickham et al., 2023a), stringr (Wickham, 2023), phytools (Revell, 2012), XML (Temple Lang, 2023), xml2 (Wickham et al., 2023b), gsubfn (Grothendieck, 2018), LaplacesDemon (Statisticat and LLC., 2021), lava (Holst and Budtz-Joergensen, 2013, 2020), coda (Plummer et al., 2006), tictoc (Izrailev, 2023), matrixStats (Bengtsson, 2023), caret (Kuhn and Max, 2008), Metrics (Hamner and Frasco, 2018), gdata (Warnes et al., 2023), tidyverse (Wickham et al., 2019), reshape2 (Wickham, 2007), scales (Wickham et al., 2023c), ggplot2 (Wickham, 2016), e hrbrthemes (Rudis, 2020).

Os códigos utilizados foram baseados na implementação computacional para a dissertação do Mestre Felipe Vargas, com algumas alterações específicas para este trabalho. As principais modificações foram feitas para possibilitar a implementação do método BSSVS.

3.2 Processo de simulação

O processo de simulação foi projetado para reproduzir a geração de amostras de uma população conhecida. Nesse contexto, primeiro é gerada uma árvore populacional, e sob essa árvore simulamos o processo de dispersão espacial para toda a população com a raiz no estado A. Em seguida, geramos uma amostra da população sintética, de acordo com um de 9 diferentes esquemas de amostragem que representam diferentes situações de viés.

Para as simulações foi considerado uma árvore populacional de tamanho 10.000 com 6 estados ou localizações geográficas: A, B, C, D, E e F.

Em relação aos parâmetros do modelo de migração, a matriz $\boldsymbol{S}$ (2.4) foi definida como

$$\boldsymbol{S} = \begin{array}{c} \begin{matrix} A & B & C & D & E & F \end{matrix} \\ \begin{bmatrix} -4 & 0.8 & 0.8 & 0.8 & 0.8 & 0.8 \\ 0.8 & -4 & 0.8 & 0.8 & 0.8 & 0.8 \\ 0.8 & 0.8 & -4 & 0.8 & 0.8 & 0.8 \\ 0.8 & 0.8 & 0.8 & -4 & 0.8 & 0.8 \\ 0.8 & 0.8 & 0.8 & 0.8 & -4 & 0.8 \\ 0.8 & 0.8 & 0.8 & 0.8 & 0.8 & -4 \end{bmatrix} \end{array} \begin{matrix} A \\ B \\ C \\ D \\ E \\ F \end{matrix} , \tag{3.1}$$

e as frequências $\pi = (\pi_A, ..., \pi_F)$ utilizadas foram

$$\pi = (0.40, 0.12, 0.12, 0.12, 0.12, 0.12). \tag{3.2}$$

Com as informações de 3.1 e 3.2, conseguimos obter a matriz de migração $\boldsymbol{\Lambda}$

$$\boldsymbol{\Lambda} = \begin{bmatrix} -0.48 & 0.096 & 0.096 & 0.096 & 0.096 & 0.096 \\ 0.32 & -0.704 & 0.096 & 0.096 & 0.096 & 0.096 \\ 0.32 & 0.096 & -0.704 & 0.096 & 0.096 & 0.096 \\ 0.32 & 0.096 & 0.096 & -0.704 & 0.096 & 0.096 \\ 0.32 & 0.096 & 0.096 & 0.096 & -0.704 & 0.096 \\ 0.32 & 0.096 & 0.096 & 0.096 & 0.096 & -0.704 \end{bmatrix}. \tag{3.3}$$

Tabela 3.1: Frequência dos estados em subamostragem hipotética

	A	B	C	D	E	F
π	0.4	0.12	0.12	0.12	0.12	0.12
n_k	80	24	24	24	24	24

Da árvore que foi simulada, amostramos $N = 200$ nós externos. Para efeitos de ilustração, a tabela 3.1 mostra a distribuição dos estados em uma árvore subamostrada de acordo com as frequências de π, em que o estado que possui a maior frequência é o estado A com 40% das observações e os demais estados possuem uma frequência de 12%. Com isso, o número de elementos do estado k na amostra n_k é de 80 elementos para A e os demais estados tem 24 elementos.

Nas simulações, essa subamostragem seguiu um processo de 9 diferentes esquemas de amostragem, representando os diferentes vieses. Os vieses utilizados para as simulações estão apresentados na tabela 3.2.

Tabela 3.2: Tabela de vieses utilizados nas simulações para máxima verossimilhança

Vieses	A	B	C	D	E	F
S1	0.80	0.04	0.04	0.04	0.04	0.04
S2	0.60	0.08	0.08	0.08	0.08	0.08
S3	0.40	0.12	0.12	0.12	0.12	0.12
S4	0.20	0.16	0.16	0.16	0.16	0.16
S5	0.16	0.16	0.16	0.16	0.16	0.20
S6	0.10	0.18	0.18	0.18	0.18	0.18
S7	0.12	0.12	0.12	0.12	0.12	0.40
S8	0.08	0.08	0.08	0.08	0.08	0.60
S9	0.04	0.04	0.04	0.04	0.04	0.80

Os vieses escolhidos para este trabalho foram construídos de maneira a observar diferentes comportamentos nos estados.

Os vieses $S1$ e $S2$ são vieses mais desequilibrados para o estado A, quando comparado aos demais estados da árvore. O viés $S3$ representa a distribuição de equilíbrio π representada na tabela 3.1, enquanto os demais vieses apresentam amostragem reduzida de A em relação à π. Os vieses $S4$, $S5$ e $S6$ são vieses com valores mais equilibrados entre todos os estados. E $S7$, $S8$ e $S9$ são os vieses desequilibrados para o estado F, quando comparado com os demais estados.

Para todos os resultados foram comparados dois diferentes métodos, o método BSTM que procura corrigir o problema de viés amostral utilizando os pesos e o método filogeográfico original, aqui referido como Standard, que não utiliza os pesos para corrigir o viés amostral.

Foram considerados três estruturas de análise distintas: inferência por máxima verossimilhança, inferência Bayesiana com a cadeia de Markov original, e inferência Bayesiana com o BSSVS.

O número de replicações utilizado para cada cenário foi 100.

3.3 Máxima Verossimilhança

Para cada caso analisado nesta seção, a matriz de taxas foi multiplicada por diferentes escalares η com o objetivo de observar os distintos comportamentos de migração. Especificamente, para valores menores do escalar, observa-se uma taxa de migração mais lenta entre os estados, enquanto valores maiores do escalar resultam em uma taxa de migração mais rápida. Os escalares selecionados foram $\eta = (0.1, 1, 10, 100)$. É importante notar que para $\eta = 0.1$ encontramos alta proporção de A na amostra, para $\eta \geq 1$ encontramos uma distribuição de nós externos já próxima a π. Para cada um desses valores, foram realizadas análises nos casos em que foi aplicada a correção do viés de amostragem com pesos (BSTM) e nos casos em que não houve correção do viés de amostragem, ou seja, sem pesos (Standard).

Deve-se salientar que para os diferentes valores de η alguns casos apresentaram erros computacionais, nos quais para $\eta = 0.1$ no caso BSTM 13 das replicações apresentaram erros computacionais, para o caso de $\eta = 1$ e $\eta = 100$ apresentaram 3 e 2 erros computacionais, respectivamente, para o caso BSTM. Esses erros não geraram resultados válidos.

A tabela 3.3 representa a média sob todas as replicações dos estados nas árvores populacionais. Note que para $\eta \geq 1$ estas são muito próximas da distribuição estacionária.

Tabela 3.3: Frequências populacionais médias para as simulações que avaliam os métodos para Máxima Verossimilhança.

	A	B	C	D	E	F
$\eta = 0.1$	0.455	0.121	0.102	0.117	0.100	0.103
$\eta = 1$	0.399	0.120	0.121	0.121	0.120	0.119
$\eta = 10$	0.40	0.120	0.120	0.120	0.120	0.120
$\eta = 100$	0.401	0.120	0.120	0.120	0.120	0.120

Para a inferência por máxima verossimilhança, os resultados gerados para as

árvores foram analisados com base nos seguintes indicadores: Erro Quadrático Médio (EQM) para a matriz de migração $\mathbf{\Lambda}$, acurácia na reconstrução dos estados para os nós internos e a acurácia na reconstrução da raiz da árvore.

Na acurácia da reconstrução da árvore, consideraremos a localização dos nós internos dentro da árvore. Assim, para cada iteração, primeiro se realizou a reconstrução dos nós internos por máxima verossimilhança Felsenstein (1981). Na sequência calculou-se a acurácia dessa reconstrução quando comparada à verdadeira árvore simulada. Finalmente, calculamos a acurácia para as 100 replicações em cada cenário de simulação.

Para o cálculo da acurácia na raiz, também realizou-se a reconstrução da raiz por máxima verossimilhança, calculou-se a acurácia da reconstrução comparada à verdadeira árvore simulada que possui a maioria dos casos no estado A. E por fim, calculou-se a acurácia para as 100 replicações.

O cálculo do erro quadrático médio (EQM) para a matriz de taxas $\mathbf{\Lambda}$ foi realizado utilizando a matriz de taxas estimadas para cada uma das 100 replicações e a verdadeira matriz de taxas da árvore simulada.

Acurácia na reconstrução dos nós internos

Na tabela 3.4 e nas figuras 3.1 a 3.4 compararemos a acurácia entre dois casos: sem a correção do viés amostral (caso Standard) e com a correção do viés amostral (caso BSTM).

Idealmente gostaríamos de encontrar valores de acurácia no caso BSTM que sejam maiores que os valores do caso Standard, indicando a eficácia da correção do viés amostral.

Tabela 3.4: Acurácia nos nós internos da árvore estimados por máxima verossimilhança

	Método	S1	S2	S3	S4	S5	S6	S7	S8	S9
$\eta = 0.1$	Standard	0.73	0.70	0.70	0.71	0.72	0.73	0.71	0.71	0.73
	BSTM	0.51	0.49	0.45	0.52	0.46	0.50	0.48	0.50	0.49
$\eta = 1$	Standard	0.34	0.39	0.39	0.29	0.28	0.30	0.31	0.29	0.29
	BSTM	0.34	0.40	0.44	0.27	0.24	0.28	0.24	0.24	0.26
$\eta = 10$	Standard	0.20	0.22	0.29	0.15	0.16	0.17	0.15	0.17	0.16
	BSTM	0.21	0.19	0.24	0.17	0.19	0.24	0.19	0.21	0.22
$\eta = 100$	Standard	0.19	0.22	0.29	0.14	0.15	0.17	0.15	0.16	0.16
	BSTM	0.20	0.19	0.24	0.16	0.19	0.23	0.18	0.21	0.22

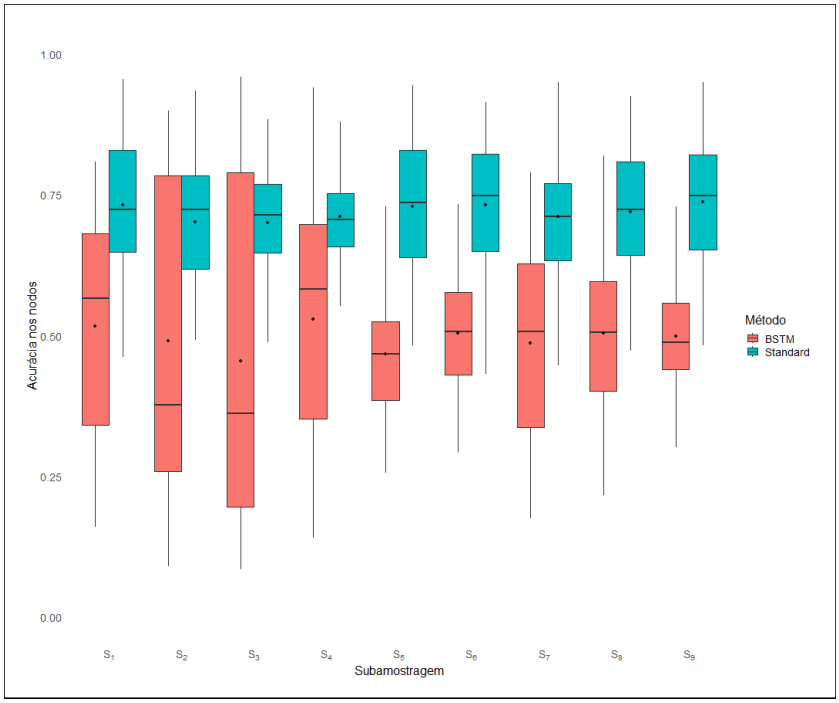

Figura 3.1: Acurácia dos nós internos estimados por máxima verossimilhança para $\eta = 0.1$ nos 9 vieses.

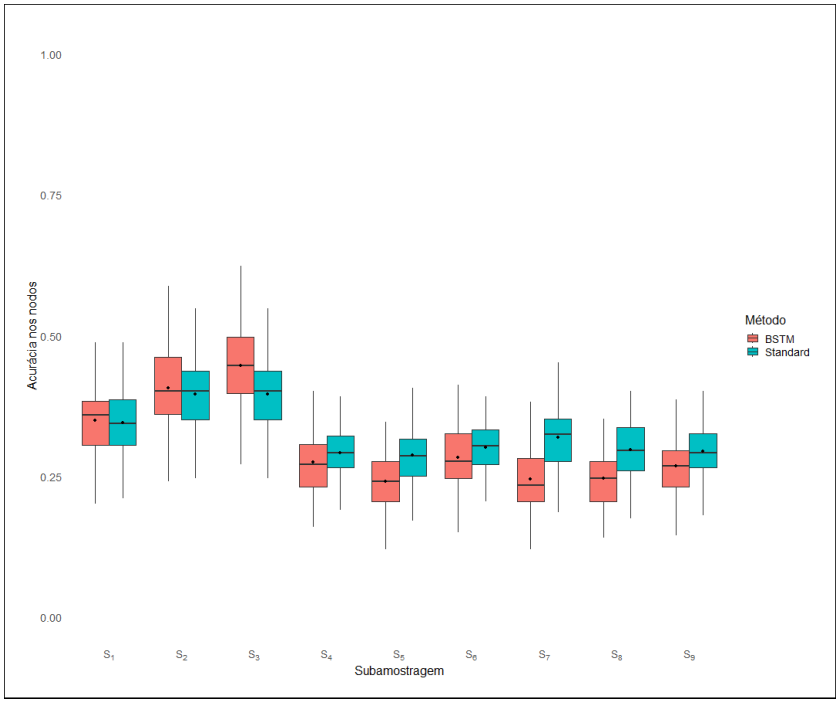

Figura 3.2: Acurácia dos nós internos estimados por máxima verossimilhança para $\eta = 1$ nos 9 vieses.

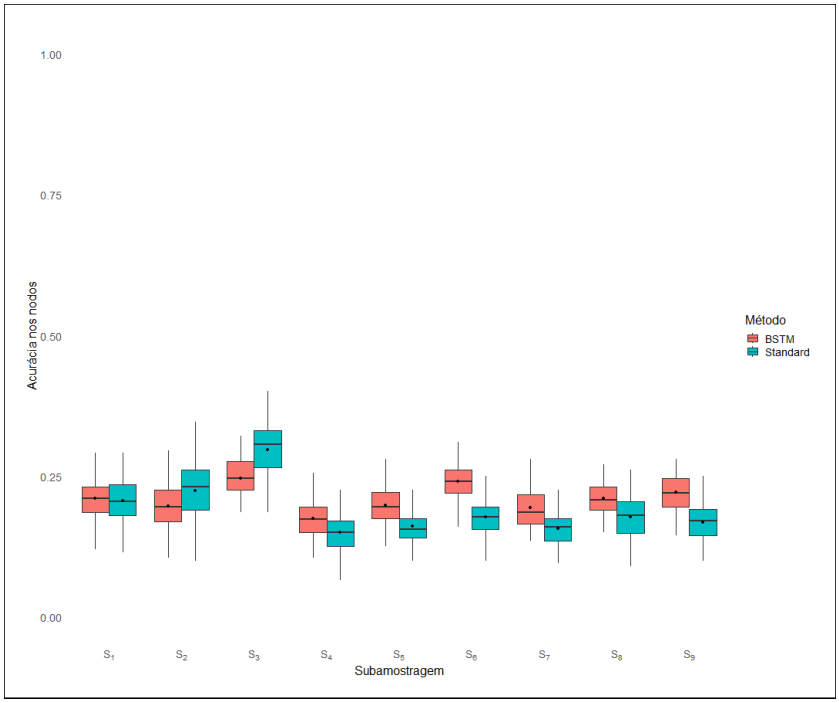

Figura 3.3: Acurácia dos nós internos estimados por máxima verossimilhança para $\eta = 10$ nos 9 vieses.

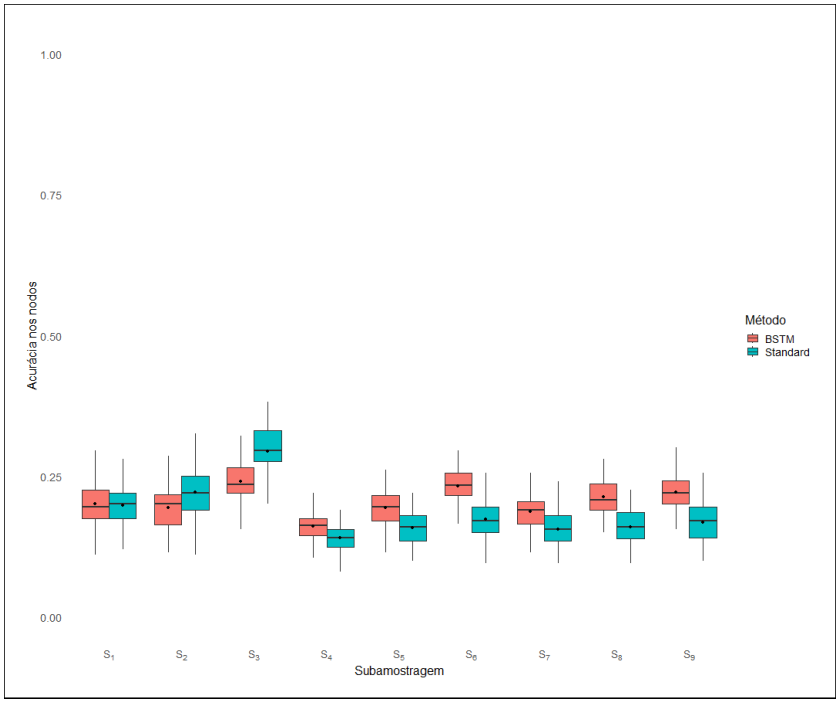

Figura 3.4: Acurácia dos nós internos estimados por máxima verossimilhança para $\eta = 100$ nos 9 vieses.

Os valores de acurácia aqui apresentados refletem o desempenho dos métodos Standard e BSTM nos diferentes vieses e para os valores de η apresentados anteriormente.

Para o valor de $\eta = 0.1$, a acurácia no método Standard é alta, variando entre

os valores 0.70 e 0.73. Já para o método BSTM, a acurácia é mais baixa quando comparada com o outro método, variando entre 0.45 e 0.52. Isso sugere que, para esse cenário, o método Standard se mostra melhor que o método BSTM.

Para $\eta = 1$, a acurácia do método Standard diminui drasticamente se comparado ao caso anterior ($\eta = 0.1$), variando os valores entre 0.28 e 0.39. Para o método BSTM, a acurácia é comparável com o método Standard, variando entre 0.24 e 0.44. Apenas para o viés $S3$, que representa a distribuição estacionária, o BSTM teve um desempenho superior.

Observando os valores de $\eta = 10$, a acurácia para o método Standard continua baixa, variando entre 0.15 e 0.29. Já para o método BSTM, a acurácia varia entre 0.17 e 0.24. Nos vieses $S5$ a $S9$ o BSTM teve um desempenho superior ao método Standard, vieses esses que representam situações em que o estado com mais frequências populacionais (A) encontra-se em proporção comparativa mais baixa na amostra.

Para $\eta = 100$, temos uma acurácia que varia de 0.14 a 0.29. E para o método BSTM que varia entre 0.16 e 0.24. Para os vieses de $S5$ a $S9$, a situação é parecida com o caso $\eta = 10$, o método BSTM mostrou um desempenho melhor que o método Standard.

Note que, como esperado, para $\eta \in \{1, 10, 100\}$, a acurácia da reconstrução é sempre maior em S3, ou seja, quando não há viés de amostragem

De maneira geral, os dados indicam que o desempenho dos métodos Standard e BSTM varia significativamente com o valor de η e com o viés amostral considerado. Para $\eta = 0.1$, cenário que não ocorre muitas migrações, o método padrão (Standard) apresenta uma acurácia alta, superando o BSTM. À medida que η aumenta para 1, o desempenho do método Standard diminui, enquanto o BSTM tem um desempenho ligeiramente superior. Com $\eta = 10$ e $\eta = 100$, o BSTM tem um desempenho superior ao Standard em vários dos cenários de vieses.

Esse panorama sugere que, enquanto o método Standard é mais eficaz em cenários com valores baixos de η, o BSTM tende a ser mais robusto em situações de maior complexidade, especialmente quando os vieses amostrais são mais pronunciados.

Acurácia da raiz

Na tabela 3.5 e nas figuras 3.5 a 3.8 comparamos a acurácia da raiz entre os dois casos.

Idealmente, também, gostaríamos de encontrar valores de acurácia na raiz no caso BSTM que sejam maiores que os valores do caso Standard, indicando a eficácia da correção do viés amostral.

Tabela 3.5: Acurácia da raiz estimada por máxima verossimilhança

	Método	S1	S2	S3	S4	S5	S6	S7	S8	S9
$\eta = 0.1$	Standard	0.71	0.59	0.52	0.38	0.75	0.66	0.49	0.68	0.82
	BSTM	0.56	0.42	0.11	0.42	0.42	0.53	0.40	0.51	0.55
$\eta = 1$	Standard	0.22	0.25	0.25	0.24	0.17	0.14	0.18	0.16	0.23
	BSTM	0.40	0.42	0.45	0.09	0.16	0.16	0.14	0.09	0.18
$\eta = 10$	Standard	0.21	0.16	0.20	0.09	0.25	0.16	0.19	0.19	0.15
	BSTM	0.23	0.21	0.36	0.22	0.17	0.20	0.24	0.31	0.25
$\eta = 100$	Standard	0.23	0.14	0.21	0.15	0.25	0.12	0.15	0.16	0.16
	BSTM	0.20	0.23	0.44	0.22	0.20	0.20	0.19	0.27	0.22

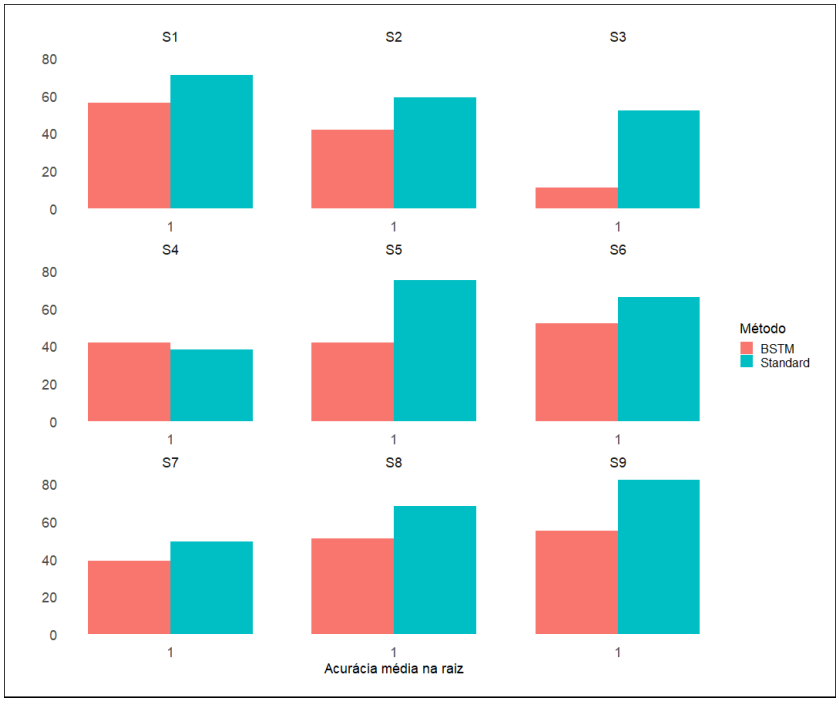

Figura 3.5: Acurácia da raiz estimada por máxima verossimilhança para $\eta = 0.1$ nos 9 vieses.

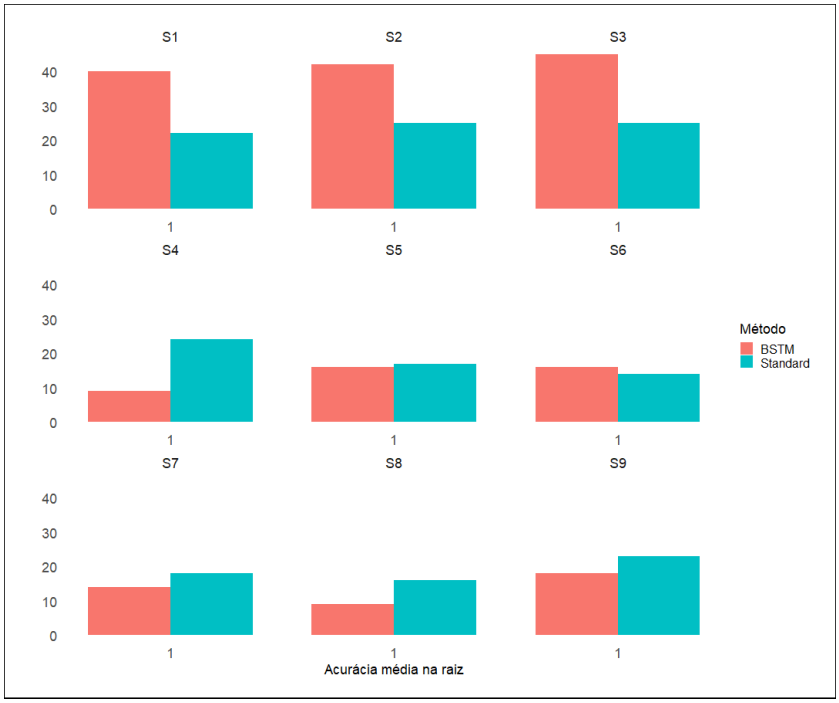

Figura 3.6: Acurácia da raiz estimada por máxima verossimilhança para $\eta = 1$ nos 9 vieses.

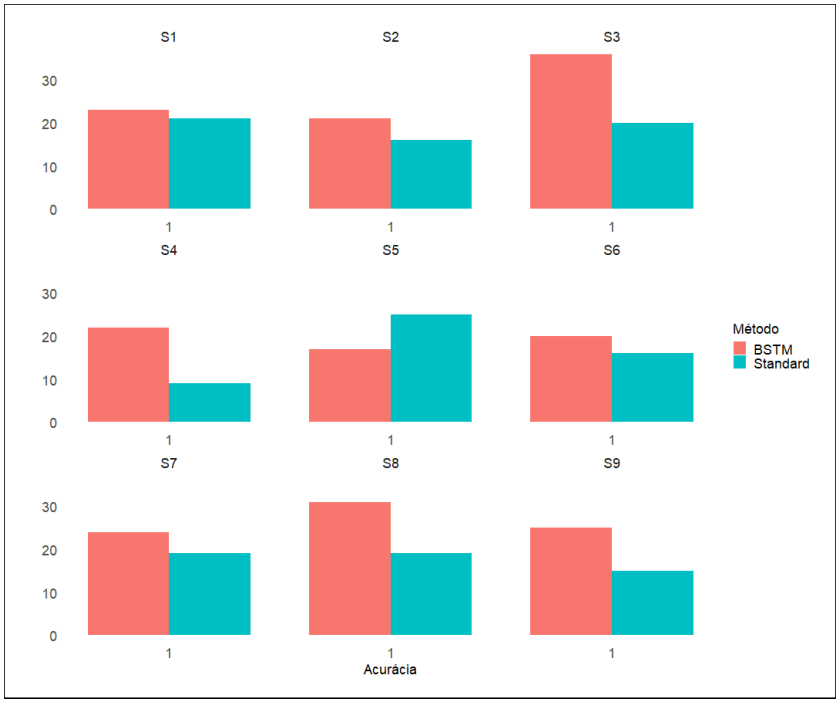

Figura 3.7: Acurácia da raiz estimada por máxima verossimilhança para $\eta = 10$ nos 9 vieses.

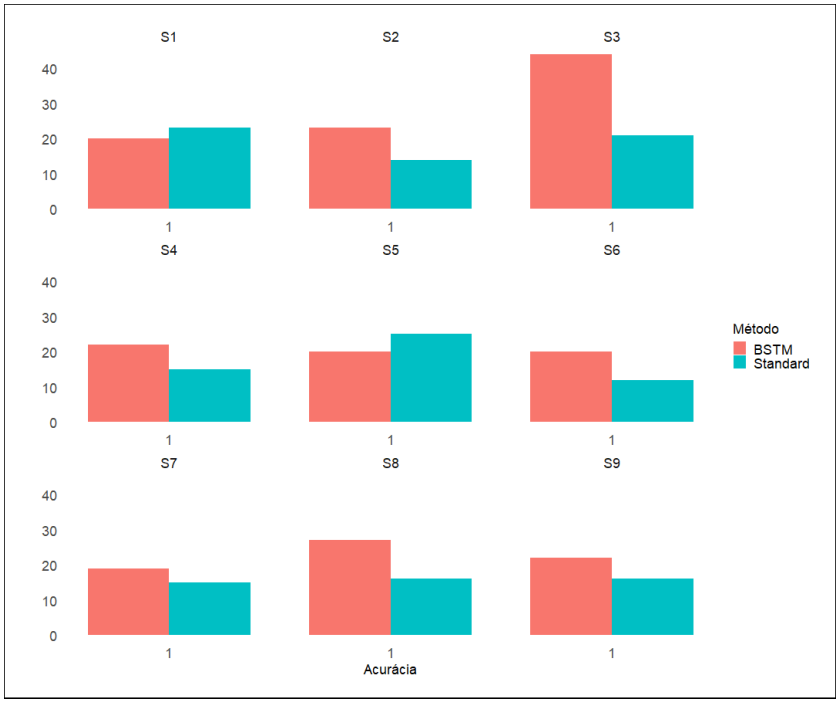

Figura 3.8: Acurácia da raiz estimada por máxima verossimilhança para $\eta = 100$ nos 9 vieses.

Os valores de acurácia aqui apresentados refletem o desempenho dos métodos Standard e BSTM nos diferentes vieses e para os valores de η apresentados anteriormente.

Para o valor de $\eta = 0.1$, no método Standard os valores se encontram entre 0.38 e 0.82. Já para o método BSTM os valores se encontram entre 0.11 e 0.56. A acurácia do método BSTM é mais baixa que do método Standard, exceto para o viés $S4$. Isso sugere que, para esse cenário, o método Standard possui uma performance melhor que o método BSTM.

Pra $\eta = 1$, a acurácia na raiz para o método Standard diminui drasticamente se comparado ao caso anterior ($\eta = 0.1$), variando entre 0.14 e 0.25. Para o método BSTM, a acurácia está variando entre 0.09 e 0.45. É possível observar que o método BSTM se mostrou superior ao método Standard para os vieses $S1$, $S2$ e $S3$.

Para o $\eta = 10$, temos que para o método Standard os valores se encontram entre 0.14 a 0.25. E para o método BSTM os valores estão entre 0.09 a 0.36. Os valores de acurácia da raiz indicam que o método BSTM teve uma performance melhor que o método Standard para os vieses $S1$ a $S4$ e $S6$ a $S9$.

Para $\eta = 100$, temos valores para o método Standard que variam entre 0.12 a 0.25. Para o método BSTM valores que variam entre 0.19 e 0.44. Os resultados de acurácia da raiz indicam que o método BSTM teve um desempenho superior ao método Standard para os vieses $S2$ a $S4$ e $S6$ a $S9$.

No geral, para a acurácia da raiz os dados indicam que o desempenho do método Standard e do método BSTM variam de acordo com o valor de η escolhido e com o viés amostral considerado. Para $\eta = 0.1$, novamente, o caso que não possui muitas migrações o método padrão teve um desempenho melhor quando comparado ao método BSTM. À medida que o valor de η aumenta vemos que o desempenho do método BSTM é melhor que o método Standard.

Erro Quadrático Médio da matriz de taxas $\mathbf{\Lambda}$

O erro quadrático médio (EQM) avalia a precisão do modelo, pois quantifica a diferença entre os valores previstos e os valores observados.

Gostaríamos que os valores de EQM para o método BSTM fossem próximos de zero, indicando que o modelo é mais preciso.

A tabela 3.6 apresenta o EQM das estimativas pontuais de máxima verossimilhança para a matriz de taxas $\mathbf{\Lambda}$.

Tabela 3.6: EQM de $\mathbf{\Lambda}$ estimado por máxima verossimilhança

	Método	S1	S2	S3	S4	S5	S6	S7	S8	S9
$\eta = 0.1$	Standard	54.40	13.00	84.50	4.56	44.40	71.40	0.93	0.83	67.70
	BSTM	163.00	38.80	253.00	13.30	133.00	214.00	2.69	2.42	203.00
$\eta = 1$	Standard	0.46	0.50	0.51	0.46	0.47	0.46	0.96	0.38	0.46
	BSTM	1.06	1.12	1.15	1.03	1.06	1.02	0.98	0.77	1.02
$\eta = 10$	Standard	13.50	13.50	13.50	33.00	13.40	13.70	15.20	47.20	13.40
	BSTM	13.50	13.60	13.90	72.30	13.10	13.90	18.60	115.00	13.30
$\eta = 100$	Standard	1367	1367	1367	1367	1367	1367	1370	1367	1375
	BSTM	1367	1367	1367	1367	1367	1366	1374	1366	1391

Para $\eta = 0.1$, o EQM no método Standard varia entre 0.93 a 84.50. Para o método BSTM os valores se encontram entre 2.42 a 253.00. Podemos observar que o método Standard possui valores inferiores ao método BSTM, indicando que o modelo padrão é melhor na estimação da matriz de taxas. Note que a falta de tendência clara e disparidade nesses resultados indicam possível dificuldade numérica dos métodos (principalmente BSTM) nesse contexto de tão pouca variabilidade.

Para $\eta = 1$, os valores de EQM no método Standard variam entre 0.38 e 0.96. Para o método BSTM os valores estão entre 0.77 a 1.15. Vemos que embora os valores do método BSTM sejam baixos, se compararmos com o método Standard eles são maiores, indicando que o modelo é um pouco inferior na estimação da matriz de taxas.

Para $\eta = 10$, os valores do método Standard estão entre 13.40 a 47.20 e para o método BSTM os valores variam entre 13.10 a 115.00. Podemos ver que para o método Standard a maioria dos valores de EQM estão constantes, exceto para o viés $S8$ que temos um EQM mais elevado. Para o método BSTM, há uma variabilidade mais alta entre os valores. Os resultados indicam que o método Standard é melhor, para esse cenário na estimação da matriz de taxas.

Para $\eta = 100$, temos valores grandes e parecidos para os dois métodos, isso indica um desempenho ruim em todos os cenários para ambos os métodos.

De modo geral, o método Standard tende a ser melhor que o método BSTM em termos de EQM na estimação da matriz de taxas para diferentes valores de η e os diferentes vieses apresentados. O método BSTM apresenta maior variabilidade, indicando dificuldades numéricas ou falta de robustez em determinados contextos. E ambos falham em obter bons resultados de EQM para $\eta = 100$, cenário esse que possui muitas migrações entre os estados.

3.4 Modelo Bayesiano original

Nessa seção foi considerado o modelo bayesiano como apresentado na seção 2.4, com inferência por MCMC.

Para cada caso analisado nesta seção, a matriz de taxas foi multiplicada por diferentes escalares η com o objetivo de observar os distintos comportamentos de migração. Como na seção 3.3 os valores de η menores mostram uma taxa de migração mais lenta entre os estados, e para valores maiores o resultado são taxas de migrações mais rápidas. Para o modelo bayesiano os escalares selecionados foram $\eta = (0.1, 1, 5)$.

Os vieses utilizados nesta seção são apresentados na tabela 3.7

Tabela 3.7: Tabela de vieses utilizados nas simulações para o modelo bayesiano original

Vieses	A	B	C	D	E	F
S1	0.80	0.04	0.04	0.04	0.04	0.04
S3	0.40	0.12	0.12	0.12	0.12	0.12
S5	0.16	0.16	0.16	0.16	0.16	0.20
S8	0.08	0.08	0.08	0.08	0.08	0.60

Os vieses escolhidos para este modelo foram construídos a observar diferentes comportamentos nos estados. Portanto, o viés $S1$ é mais desequilibrado para o estado A (estado esse que possui maior probabilidade no modelo) quando comparado aos outros estados. O viés $S3$ representa distribuição de equilíbrio π representada na tabela 3.1. E o viés $S8$ é mais desequilibrado para o estado F quando comparado aos outros estados.

Para cada um desses valores, foram realizadas análises nos casos em que foi aplicada a correção do viés de amostragem com pesos (BSTM) e nos casos em que não houve a correção do viés de amostragem, ou seja, sem pesos (Standard).

Foram utilizadas cadeias de tamanho 200.000 com burn-in de 20.000 resultando em uma taxa de aceitação no Metropolis-Hastings que varia entre 0.23 e 0.40. As distribuições à priori utilizadas para a matriz diagonal das frequências de equilíbrio $\mathbf{\Pi}$ (2.8) e para a matriz de taxas S (2.9) estão definidas na seção 2.4.1.

A tabela 3.8 representa as frequências populacionais médias sobre todas as populações dos estados Note que para $\eta \geq 1$ estas são muito próximas da distribuição estacionária.

Tabela 3.8: Frequências populacionais médias para as simulações que avaliam os métodos para Modelo Bayesiano Original.

	A	B	C	D	E	F
$\eta = 0.1$	0.455	0.121	0.102	0.117	0.100	0.103
$\eta = 1$	0.399	0.120	0.121	0.121	0.120	0.119
$\eta = 5$	0.40	0.120	0.120	0.120	0.120	0.120

Nesta seção, os resultados gerados para as árvores foram analisados com base nos seguintes indicadores: Erro Quadrático médio (EQM) para a matriz de migração $\mathbf{\Lambda}$

e para as frequências de equilíbrio $\mathbf{\Pi}$, acurácia na reconstrução dos estados para os nós internos e a acurácia na reconstrução da raiz da árvore.

Para analisar a acurácia da reconstrução da árvore, consideramos a localização dos nós internos dentro da árvore. Assim, para cada replicação, ou seja, para cada uma das 200 árvores simuladas no processo realizou-se a reconstrução dos nós internos pelo método MCMC original considerando a moda a posteriori. Na sequência calculou-se a acurácia para cada uma dessas replicações, para então obter-se a acurácia média para as 200 replicações.

Para o cálculo da acurácia da raiz, também realizou-se a reconstrução da raiz pelo método MCMC original considerando a moda a posteriori, para então calcular a acurácia sob todas as replicações.

Por fim, para o cálculo do erro quadrático médio (EQM) para a matriz de taxas $\mathbf{\Lambda}$ e a para as frequências de equilíbrio $\mathbf{\Pi}$ calculou-se a média das replicações do MCMC (média a posteriori de cada replicação), para então calcular-se o EQM sob essas estimativas.

Acurácia na reconstrução dos nós internos

Na tabela 3.9 e nas figuras 3.9 a 3.11 compararemos a acurácia na reconstrução dos nós internos entre o método padrão (Standard) e o método de correção de viés de amostragem (BSTM).

Idealmente, gostaríamos que que os valores de acurácia na reconstrução dos nós internos para o método BSTM sejam maiores que os valores do método padrão.

Tabela 3.9: Acurácia nos nós internos da árvore estimados pelo modelo bayesiano original

	Método	S1	S3	S5	S8
$\eta = 0.1$	Standard	0.61	0.67	0.73	0.71
	BSTM	0.42	0.55	0.54	0.45
$\eta = 1$	Standard	0.29	0.27	0.31	0.30
	BSTM	0.41	0.25	0.27	0.27
$\eta = 5$	Standard	0.20	0.16	0.17	0.17
	BSTM	0.32	0.16	0.17	0.15

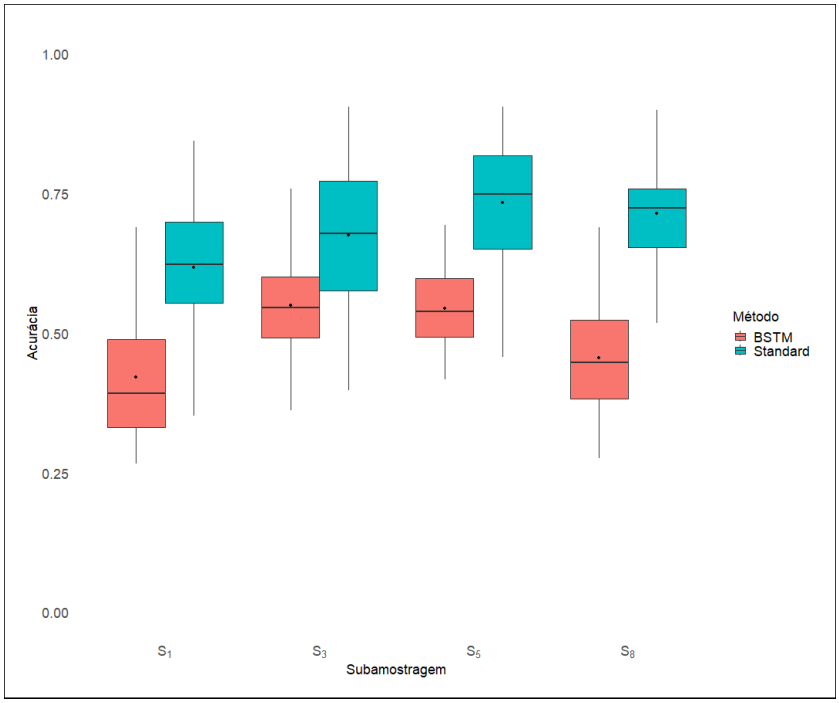

Figura 3.9: Acurácia nos nós internos estimada pelo modelo bayesiano original para $\eta = 0.1$ nos 4 vieses.

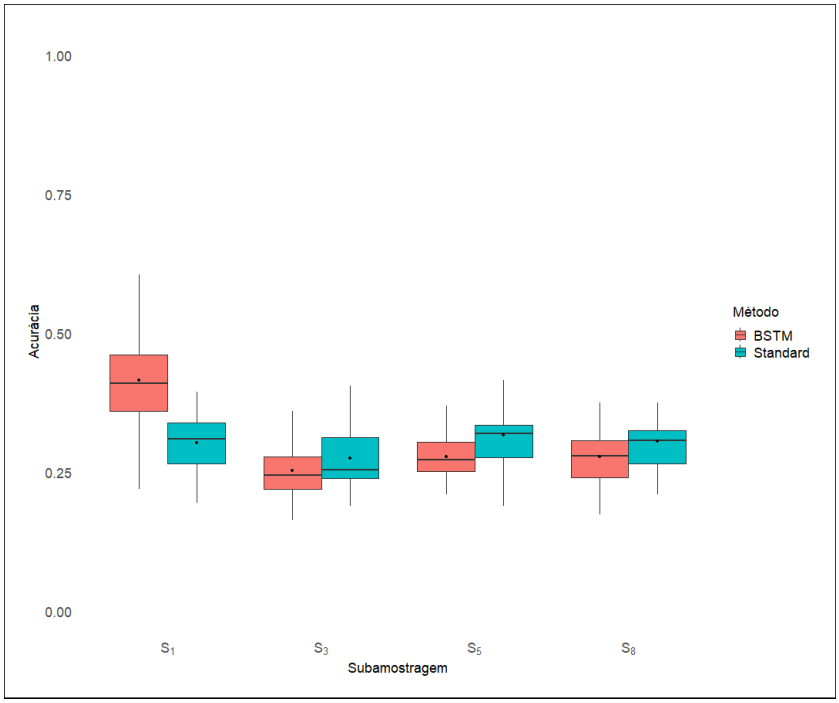

Figura 3.10: Acurácia nos nós internos estimada pelo modelo bayesiano original para $\eta = 1$ nos 4 vieses.

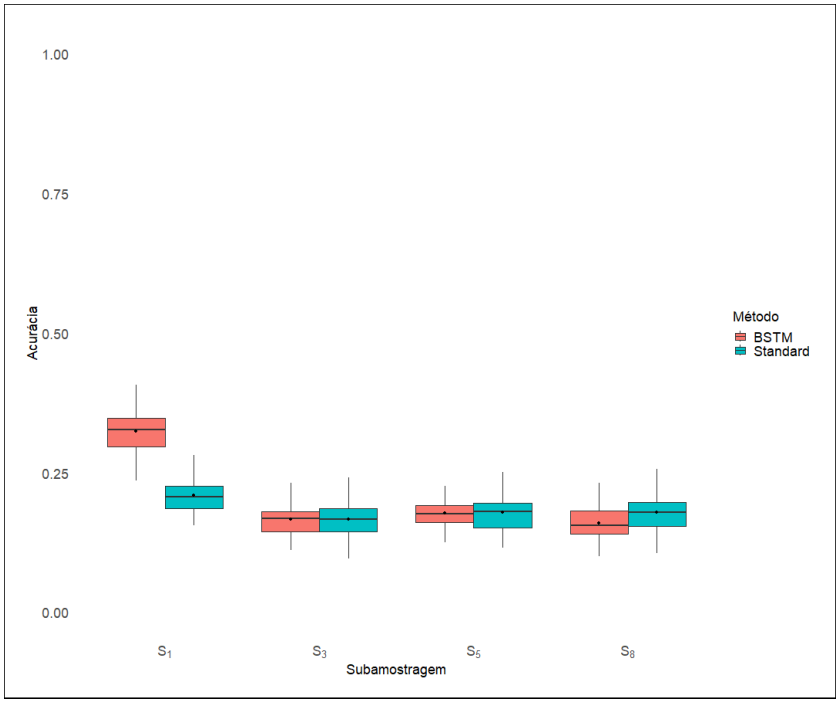

Figura 3.11: Acurácia nos nós internos estimada pelo modelo bayesiano original para $\eta = 5$ nos 4 vieses.

Para $\eta = 0.1$, no método Standard, a acurácia tem valores variando entre 0.61 e 0.73. Para o método BSTM temos valores variando entre 0.42 e 0.55, embora sejam inferiores aos do método Standard. Isso sugere que, nesse cenário, o método Standard teve um desempenho superior ao método BSTM.

Para $\eta = 1$, no método Standard, a acurácia varia entre 0.27 e 0.31, enquanto, para o método BSTM, os valores situam-se entre 0.25 e 0.41. Nota-se que os métodos possuem valores de acurácia próximos, com exceção do viés $S1$, onde o método BSTM apresentou um desempenho superior ao método Standard. O viés $S1$ representa a situação em que o estado com maior frequência populacional (A) encontra-se em proporção comparativamente mais alta na amostra.

Para $\eta = 5$, no método Standard, a acurácia está entre 0.16 e 0.20, enquanto no método BSTM os valores variam entre 0.15 e 0.32. Apesar dos valores de acurácia relativamente baixos, os métodos apresentam resultados próximos para os vieses $S3$ a $S8$. No entanto, para o viés $S1$, o método BSTM mostrou um desempenho superior ao método Standard.

Em termos gerais, os dados indicam que o desempenho dos métodos Standard e BSTM varia significativamente em função do valor de η e dos vieses selecionados. Observa-se que, para valores menores de η, o método Standard demonstra melhor desempenho na acurácia dos nós internos. Contudo, à medida que η aumenta, a acurácia do método Standard diminui, enquanto a do método BSTM se aproxima ou até supera a acurácia do método Standard em determinados casos.

Acurácia da raiz

Na tabela 3.10 e as figuras 3.12 e 3.13 comparamos a acurácia da raiz entre o método padrão (Standard) e o método de correção de viés de amostragem (BSTM).

Idealmente, também, gostaríamos de encontrar valores de acurácia da raiz no caso BSTM que sejam maiores que os valores do caso Standard, indicando a eficácia da correção do viés amostral.

Tabela 3.10: Acurácia da raiz estimada pelo modelo bayesiano original

	Método	S1	S3	S5	S8
$\eta = 0.1$	Standard	0.26	0.58	0.84	0.60
	BSTM	0	0	0	0
$\eta = 1$	Standard	0.04	0.10	0.16	0.22
	BSTM	0.52	0.04	0.08	0.10
$\eta = 5$	Standard	0.13	0.11	0.20	0.13
	BSTM	0.52	0.05	0.05	0.14

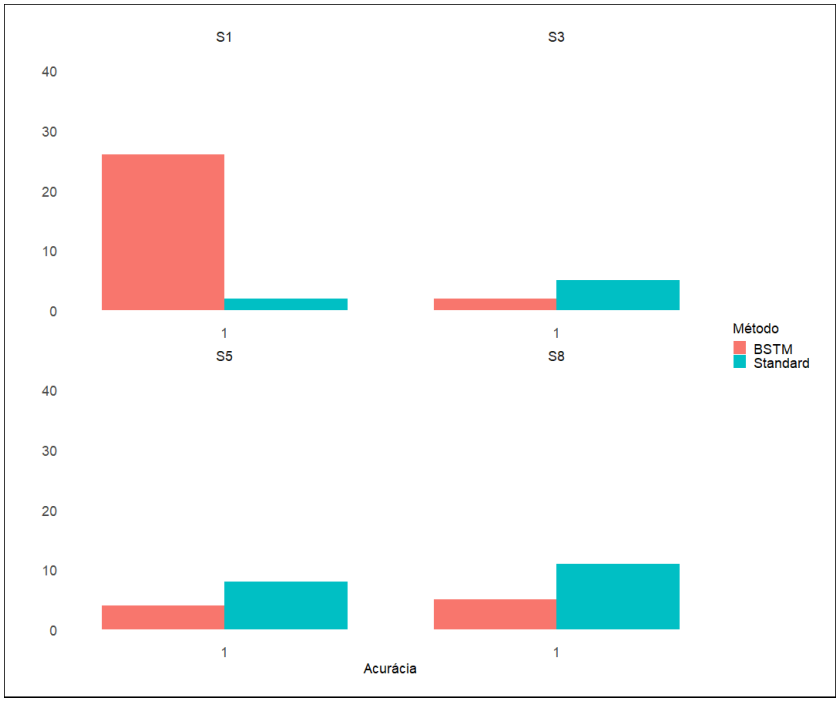

Figura 3.12: Acurácia da raiz estimada pelo modelo bayesiano original para $\eta = 1$ nos 4 vieses.

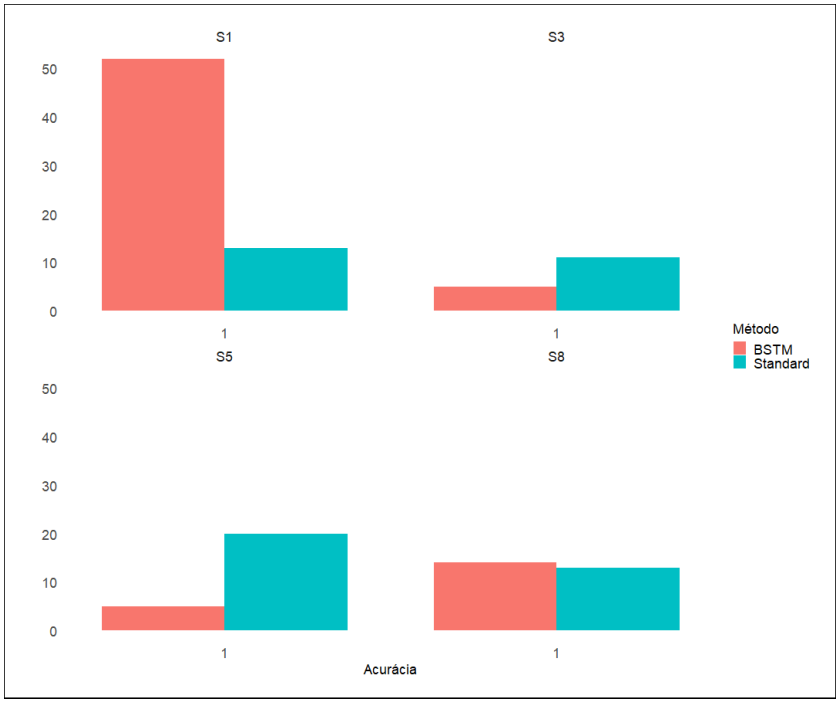

Figura 3.13: Acurácia da raiz estimada pelo modelo bayesiano original para $\eta = 5$ nos 4 vieses.

Para $\eta = 0.1$, no método Standard, os valores variam entre 0.26 a 0.60, evidenciando um desempenho satisfatório. Em contrapartida, o método BSTM exibiu uma acurácia igual a zero em todos os vieses analisados, indicando que o método não foi capaz de acertar a reconstrução da raiz para nenhuma das replicações nesse cenário. Por essa razão, não foi apresentado gráfico para este cenário.

Para $\eta = 1$, no método Standard, os valores estão entre 0.04 a 0.22. No método BSTM, os valores variam entre 0.04 a 0.52. Observa-se que, para a maioria dos vieses, o método Standard possui um desempenho superior ao método BSTM, exceto para o viés $S1$, no qual o método BSTM demonstrou desempenho claramente superior em comparação com o método Standard.

Para $\eta = 5$, no método Standard, os valores variam entre 0.11 e 0.20, enquanto no método BSTM os valores estão entre 0.05 e 0.52. Nota-se que, para o viés $S1$, o método BSTM apresentou uma performance bastante superior ao método Standard, enquanto apresentou performace um pouco inferior nos demais casos.

Em síntese, a análise da acurácia da raiz indica que o desempenho dos métodos variam de acordo com os valores de η e os vieses escolhidos. Para $\eta = 0.1$, o método Standard apresentou um desempenho superior ao do método BSTM. À medida que o valor de η aumenta, observa-se uma diminuição significativa no desempenho do método Standard para alguns vieses, enquanto o método BSTM exibe um desempenho melhor para casos em que estado com maior frequência populacional (A) encontra-se em proporção comparativamente mais alta na amostra (viés S1).

Erro Quadrático Médio

Erro Quadrático Médio para o parâmetro $\mathbf{\Lambda}$

A tabela 3.11 e as figuras 3.14 a 3.16 apresentam o EQM para $\mathbf{\Lambda}$ estimado pelo modelo bayesiano original para o método padrão (Standard) e o método de correção de viés de amostragem (BSTM).

Idealmente, gostaríamos de encontrar valores de EQM para o parâmetro $\mathbf{\Lambda}$ no caso BSTM que sejam menores que os valores do caso Standard, indicando que o modelo de correção de viés é mais preciso.

Tabela 3.11: EQM para o $\mathbf{\Lambda}$ estimado pelo modelo bayesiano original

	Método	S1	S3	S5	S8
$\eta = 0.1$	Standard	0.90	0.65	0.63	0.79
	BSTM	1.00	0.96	0.96	1.01
$\eta = 1$	Standard	1.29	1.51	1.55	1.41
	BSTM	1.17	1.28	1.00	1.02
$\eta = 5$	Standard	1.34	1.62	1.69	1.50
	BSTM	1.21	1.48	1.01	1.05

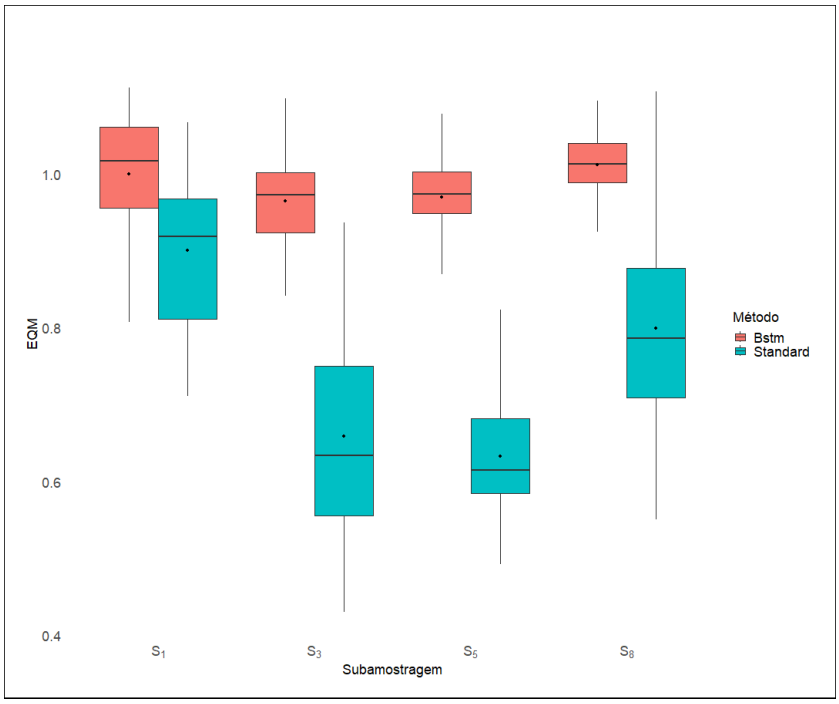

Figura 3.14: EQM de $\mathbf{\Lambda}$ estimado pelo modelo bayesiano original para $\eta = 0.1$ nos 4 vieses.

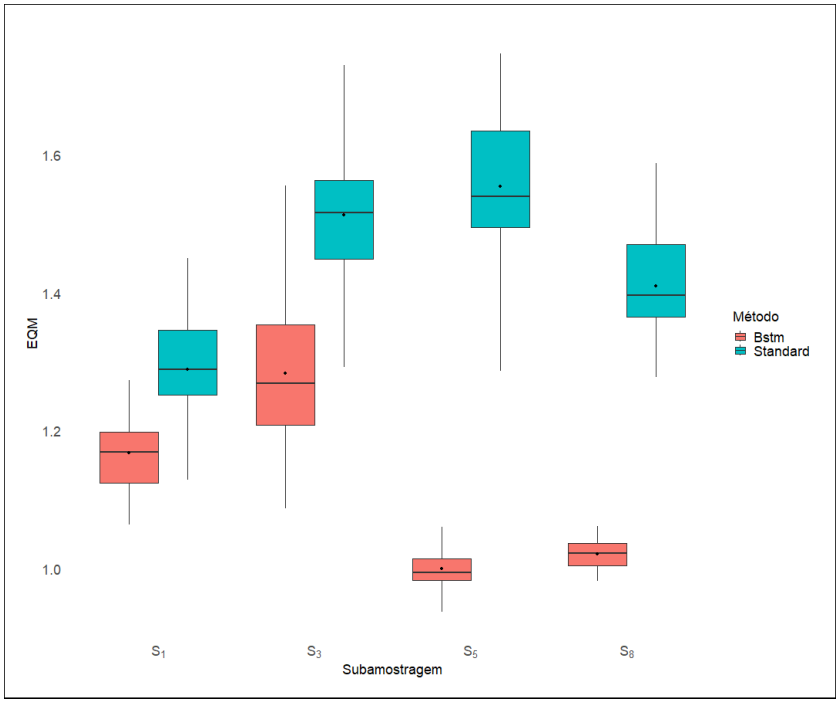

Figura 3.15: EQM de $\mathbf{\Lambda}$ estimado pelo modelo bayesiano original para $\eta = 1$ nos 4 vieses.

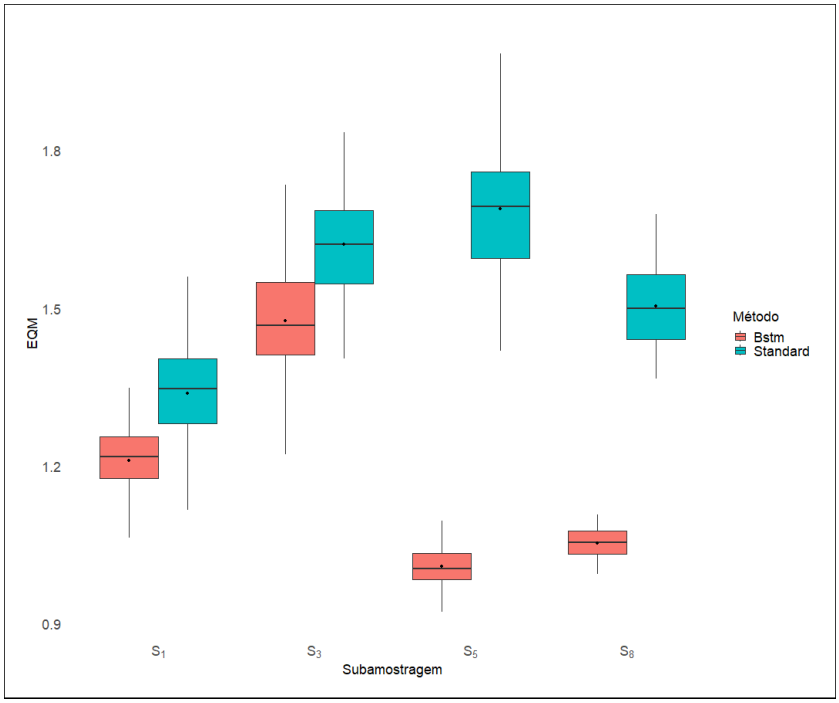

Figura 3.16: EQM de $\mathbf{\Lambda}$ estimado pelo modelo bayesiano original para $\eta = 5$ nos 4 vieses.

Para $\eta = 0.1$, no método Standard, os valores se encontram entre 0.63 a 0.90. Já para o método BSTM, os valores variam entre 0.96 e 1.01. Podemos observar que o método Standard apresenta valores inferiores aos do método BSTM, indicando que o modelo padrão é superior neste cenário para a estimação da matriz de taxas.

Em $\eta = 1$, os valores do EQM para o método Standard se encontram entre 1.29 a 1.55. Para o método BSTM, variam entre entre 1.00 a 1.28. Nota-se que, nesse cenário, os valores do método Standard são maiores, o que indica que, para esse cenário, o método BSTM teve um desempenho melhor que o método padrão para todos os vieses.

E para $\eta = 5$, o método Standard apresenta valores que variam entre 1.34 a 1.69, enquanto para o método BSTM os valores estão entre 1.01 a 1.48. Novamente, temos que os valores do método Standard são maiores que os valores do método BSTM, indicando que método BSTM teve um desempenho superior quando comparado com o método padrão para todos os vieses.

Em geral, o desempenho dos métodos também variam de acordo com a escolha dos valores de η e dos vieses. Nos casos analisados, observa-se que conforme os valores de η aumentam, indicando que o método BSTM se mostrou mais preciso que o método padrão.

Erro Quadrático Médio para as frequências de equilíbrio $\mathbf{\Pi}$

A tabela 3.12 e as figuras 3.17 a 3.19 apresentam o EQM para $\mathbf{\Pi}$ estimado pelo modelo bayesiano original para o método padrão (Standard) e o método de correção de viés de amostragem (BSTM).

Idealmente, também gostaríamos de encontrar valores de EQM para o parâmetro $\mathbf{\Pi}$ no caso BSTM que sejam menores que os valores do caso Standard, indicando que o método de correção de viés é mais preciso.

Tabela 3.12: EQM para $\mathbf{\Pi}$ estimado pelo modelo bayesiano original

	Método	S1	S3	S5	S8
$\eta = 0.1$	Standard	0.08	0.07	0.13	0.18
	BSTM	0.16	0.17	0.15	0.14
$\eta = 1$	Standard	0.14	0.01	0.10	0.21
	BSTM	0.15	0.04	0.08	0.07
$\eta = 5$	Standard	0.14	0.01	0.10	0.22
	BSTM	0.15	0.02	0.10	0.08

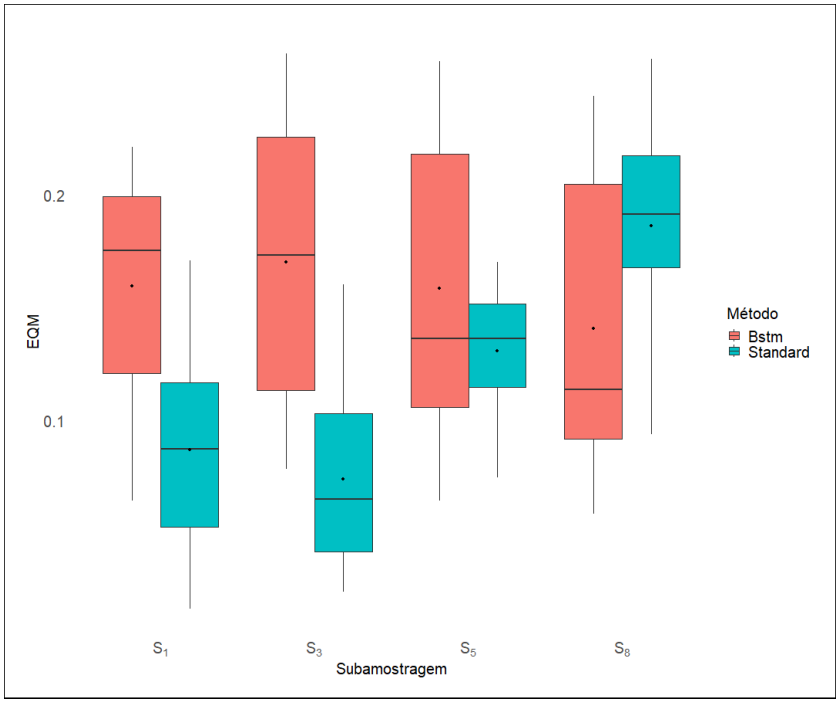

Figura 3.17: EQM de $\mathbf{\Pi}$ estimado pelo modelo bayesiano original para $\eta = 0.1$ nos 4 vieses.

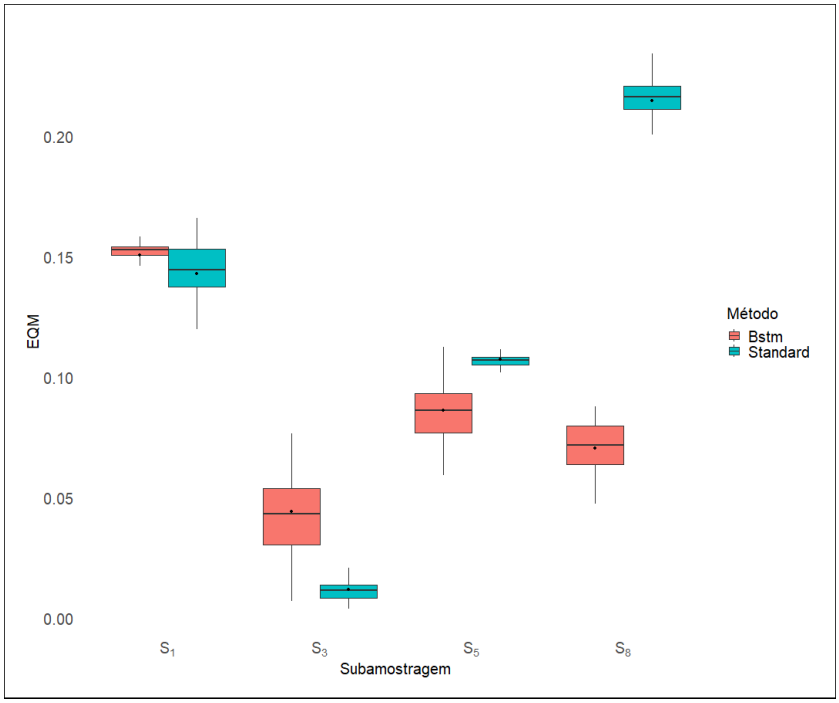

Figura 3.18: EQM de $\mathbf{\Pi}$ estimado pelo modelo bayesiano original para $\eta = 1$ nos 4 vieses.

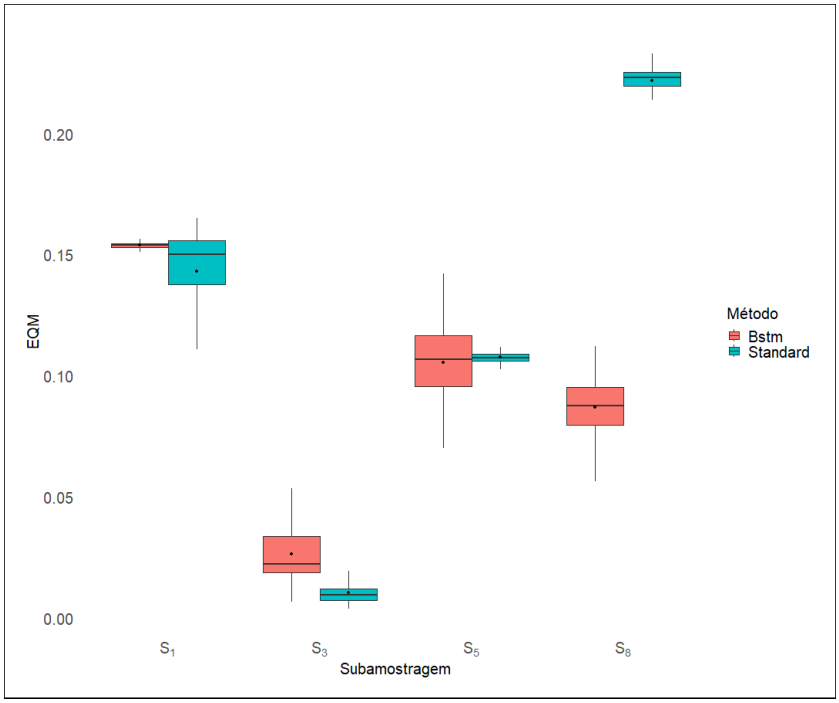

Figura 3.19: EQM de $\mathbf{\Pi}$ estimado pelo modelo bayesiano original para $\eta = 5$ nos 4 vieses.

Para o caso $\eta = 0.1$ o EQM para $\mathbf{\Pi}$ no método Standard encontra-se entre 0.07 a 0.18. Ao comparar com o método BSTM, cujos valores variam entre 0.14 a 0.17, observa-se que, apesar de o BSTM apresentar valores baixos, o método Standard demonstra um desempenho superior.

Para $\eta = 1$ método Standard apresenta valores que variam entre 0.01 a 0.21. Para o método BSTM os valores variam entre 0.04 a 0.15. O método BSTM, nesse cenário, teve um desempenho equivalente ao método Standard para os vieses $S1$ a $S5$ e um desempenho ótimo para o viés $S8$ quando comparado ao método Standard.

Para $\eta = 5$, o método Standard apresenta valores entre 0.01 a 0.22. E para o método BSTM, os valores variam entre 0.02 a 0.15. Ao comparar os métodos, é possível observar que para os vieses $S1$ a $S5$ os métodos apresentam valores próximos, mas para o viés $S8$ o método BSTM se mostrou mais preciso.

Em geral, os dados mostram que o desempenho dos métodos Standard e BSTM varia em função dos valores de η e dos vieses escolhidos. Observa-se que, para valores muito pequenos de η, o método Standard mostra melhor precisão em comparação com o método BSTM. Com valores maiores de η, há casos em que os resultados de ambos os métodos são equivalentes, como no caso do viés que representa a situação em que o estado com maior frequência populacional (A) está em proporção comprativa mais alta na amostra ($S1$), e também no caso que não há viés de amostragem ($S3$) ou a amostragem é aproximadamente homogênea ($S5$). Por fim, para os casos em que os vieses representam situações em que o estado com mais frequência populacional está em proporção comparativa mais baixa e um outro estado está super representado ($S8$) o método BSTM apresentou melhor precisão.

3.5 Modelo Bayesiano com BSSVS

Nessa seção foi considerado o modelo bayesiano com o BSSVS como apresentado na seção 2.5.

O processo de simulação foi projetado para reproduzir a geração de amostras de uma população conhecida. Nesse contexto, primeiro é gerada uma árvore populacional, e sob essa árvore simulamos o processo de dispersão espacial para toda a população com a raiz no estado A. Em seguida, geramos uma amostra da população sintética, de acordo com um de 2 diferentes esquemas de amostragem que representam diferentes situações de viés.

Para as simulações foi considerado uma árvore populacional de tamanho 10.000 com 6 estados ou localizações geográficas: A, B, C, D, E e F.

A matriz de indicadores $\boldsymbol{\delta}$ (3.4) utilizada foi

$$\boldsymbol{\delta} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix}. \tag{3.4}$$

A figura 3.20 apresenta o grafo de comunicação entre os estados (localizações geográficas).

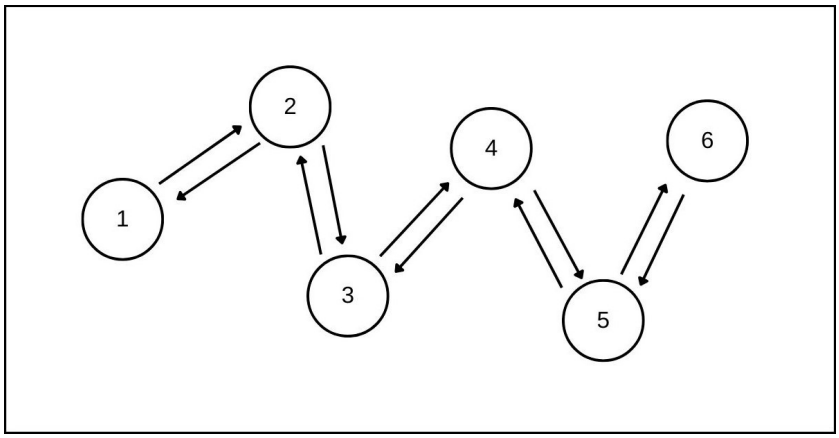

Figura 3.20: Grafo de comunicação entre os estados.

E a matriz $\boldsymbol{\Lambda}$ utilizada foi (2.13)

$$\boldsymbol{\Lambda} = \begin{array}{c} \begin{matrix} A & B & C & D & E & F \end{matrix} \\ \begin{bmatrix} -0.8 & 0.8 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.8 & -1.6 & 0.8 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.8 & -1.6 & 0.8 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.8 & -1.6 & 0.8 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.8 & -1.6 & 0.8 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.8 & -0.8 \end{bmatrix} \end{array} \begin{matrix} A \\ B \\ C \\ D \\ E \\ F \end{matrix}, \tag{3.5}$$

Os vieses utilizados nesta seção são apresentados na tabela 3.13

Tabela 3.13: Tabela de vieses utilizados nas simulações para o modelo bayesiano com BSSVS

Vieses	A	B	C	D	E	F
S1	0.80	0.04	0.04	0.04	0.04	0.04
S3	0.40	0.12	0.12	0.12	0.12	0.12
S4	0.16	0.16	0.16	0.16	0.16	0.16
S8	0.04	0.04	0.04	0.04	0.04	0.80

Os vieses escolhidos para este modelo foram construídos a observar diferentes comportamentos nos estados. Portanto, os vieses $S1$ e $S3$ são desequilibrados para o estado A (estado da raiz) quando comparado aos outros estados.[1]

Foram utilizadas cadeias MCMC de tamanho 100.000 com burn-in de 10.000

As distribuições à priori utilizadas para a matriz de taxas S e a para matriz de variáveis indicadoras $\boldsymbol{\delta}$ estão definidas, respectivamente, em (2.9) e (2.15).

A tabela 3.14 representa as frequências populacionais médias sobre todas as populações dos estados

Tabela 3.14: Frequências populacionais médias para as simulações que avaliam os métodos para o Modelo Bayesiano por BSSVS.

	A	B	C	D	E	F
$\eta = 1$	0.167	0.167	0.166	0.166	0.166	0.166

Nesta seção, os resultados gerados para as árvores foram analisados com base nos seguintes indicadores: Erro Quadrático médio (EQM) para a matriz de migração $\boldsymbol{\Lambda}$, acurácia na reconstrução dos estados para os nós internos, a acurácia na reconstrução da raiz da árvore e acurácia das indicadoras $\boldsymbol{\delta}$.

Para acurácia das indicadoras $\boldsymbol{\delta}$ primeiro, para cada replicação, foram estimadas as probabilidades a posteriori de inclusão de cada indicadora pelo MCMC. Essas estimativas finais das indicadoras foram obtidas calculando as proporções das respostas ao longo das iterações. Por fim, calculou-se a acurácia para cada uma dessas replicações, para então obter-se a acurácia média para as 200 replicações.

Acurácia na reconstrução dos nós internos

Na tabela 3.15 e na figura 3.21 compararemos a acurácia entre dois casos: sem a correção do viés amostral (caso Standard) e com a correção do viés amostral (caso BSTM).

Idealmente gostaríamos de encontrar valores de acurácia no caso BSTM que sejam maiores que os valores do caso Standard, indicando a eficácia da correção do viés amostral.

[1] Ao projetar essa seção, foram escolhidos também 2 vieses mais equilibrados, entretanto não foi possível concluir as simulações em tempo. Se possível, esses resultados serão incluídos na apresentação e versão final do TCC

Tabela 3.15: Acurácia nos nós internos da árvore estimados pelo modelo bayesiano com BSSVS

	Método	S1	S3	S4	S8
$\eta = 1$	Standard	0.17	0.17	0.16	0.16
	BSTM	0.17	0.17	0.16	0.16

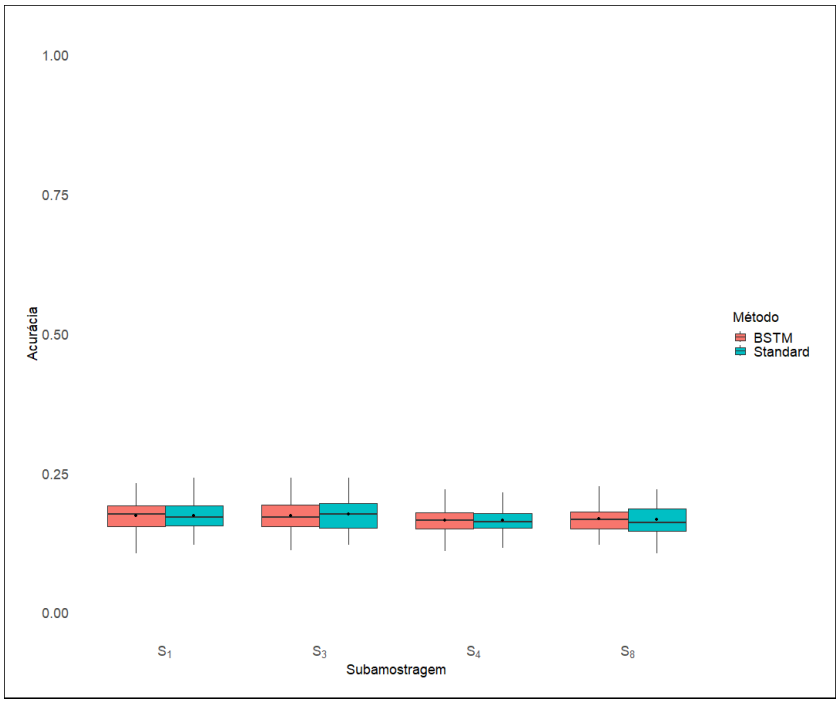

Figura 3.21: Acurácia nos nós internos estimada pelo modelo bayesiano BSSVS nos 4 vieses.

Com os resultados apresentados podemos observar que para ambos os métodos os valores médios são iguais, e suas distribuições muito semelhantes. Indicando que ambos possuem o mesmo desempenho.

Acurácia da raiz

Na tabela 3.16 e a figura 3.22 comparamos a acurácia da raiz entre o método padrão (Standard) e o método de correção de viés de amostragem (BSTM).

Idealmente, também, gostaríamos de encontrar valores de acurácia da raiz no caso BSTM que sejam maiores que os valores do caso Standard, indicando a eficácia da correção do viés amostral.

Tabela 3.16: Acurácia da raiz estimado pelo modelos bayesiano com BSSVS

	Método	S1	S3	S4	S8
$\eta = 1$	Standard	0.15	0.19	0.16	0.20
	BSTM	0.16	0.16	0.18	0.09

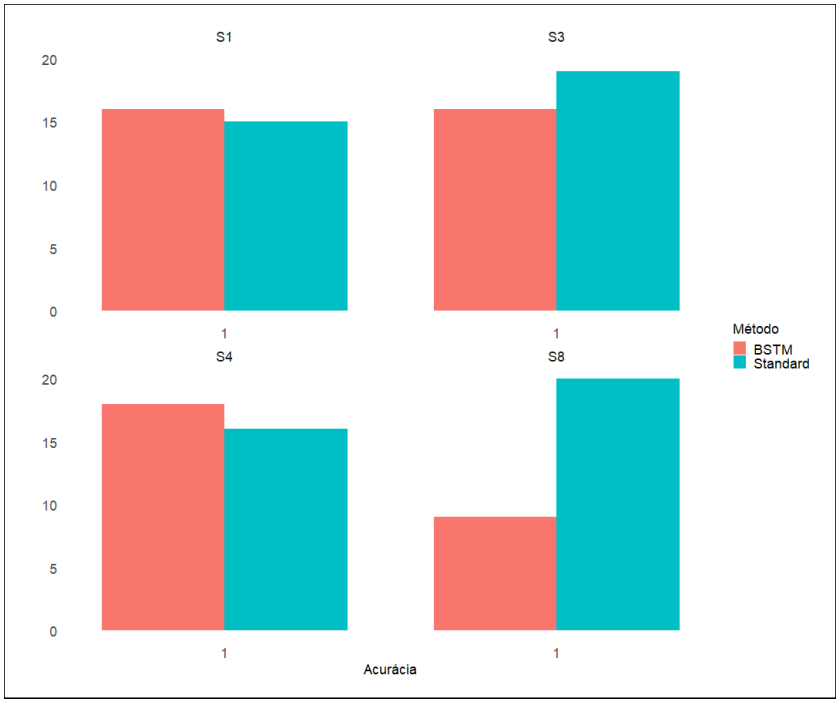

Figura 3.22: Acurácia na raiz estimada pelo modelo bayesiano BSSVS nos 4 vieses.

Observa-se que para ambos os métodos os valores são muito parecidos, e suas distribuições muito semelhantes.

Acurácia de δ estimada pelo modelo bayesiano com BSSVS

Na tabela 3.17 e a figura 3.23 comparamos a acurácia das indicadoras entre o método padrão (Standard) e o método de correção de viés de amostragem (BSTM).

Idealmente, também, gostaríamos de encontrar valores de acurácia das indicadoras no caso BSTM que sejam maiores que os valores do caso Standard, indicando a eficácia da correção do viés amostral.

Tabela 3.17: Acurácia de $\boldsymbol{\delta}$ estimada pelo modelo bayesiano com BSSVS

	Método	S1	S3	S4	S8
$\eta = 1$	Standard	0.51	0.47	0.44	0.52
	BSTM	0.62	0.60	0.46	0.62

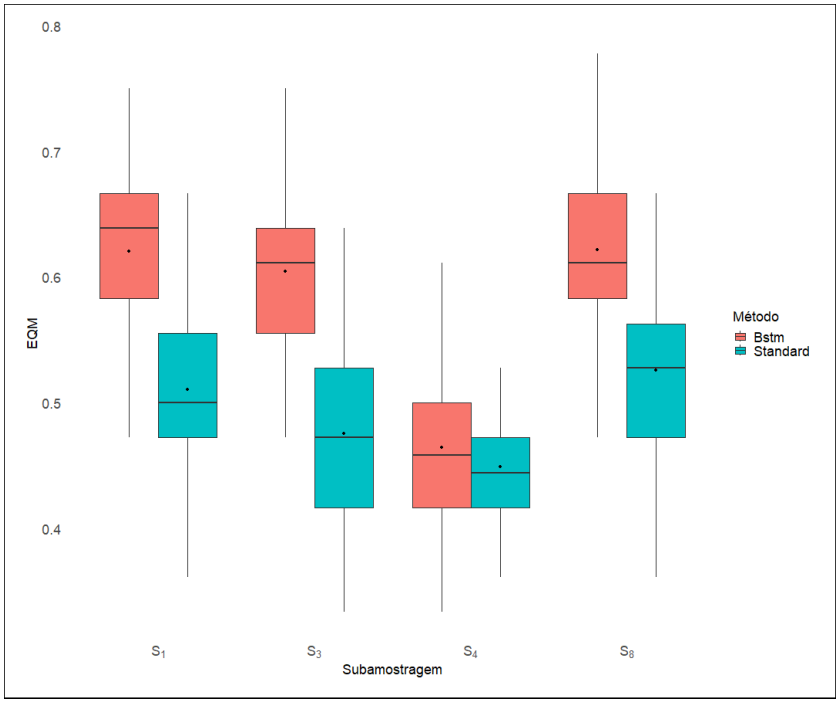

Figura 3.23: Acurácia para matriz dos indicadores $\boldsymbol{\delta}$ estimada pelo modelo bayesiano com BSSVS nos 4 vieses.

Observa-se que o método BSTM teve uma performace um pouco melhor quando comparada ao método Standard, para ambos os vieses.

Erro Quadrático Médio para a matriz de taxas $\boldsymbol{\Lambda}$

A tabela 3.18 e a figura 3.24 apresentam o EQM para $\boldsymbol{\Lambda}$ estimado pelo modelo bayesiano com BSSVS para o método padrão (Standard) e o método de correção de viés de amostragem (BSTM).

Idealmente, gostaríamos de encontrar valores de EQM para o parâmetro $\boldsymbol{\Lambda}$ no caso BSTM que sejam menores que os valores do caso Standard, indicando que o modelo de correção de viés é mais preciso.

Tabela 3.18: EQM para o $\boldsymbol{\Lambda}$ estimada pelo modelo bayesiano com BSSVS

	Método	S1	S3	S4	S8
$\eta = 1$	Standard	1.82	2.16	2.29	1.81
	BSTM	1.65	1.86	2.14	1.65

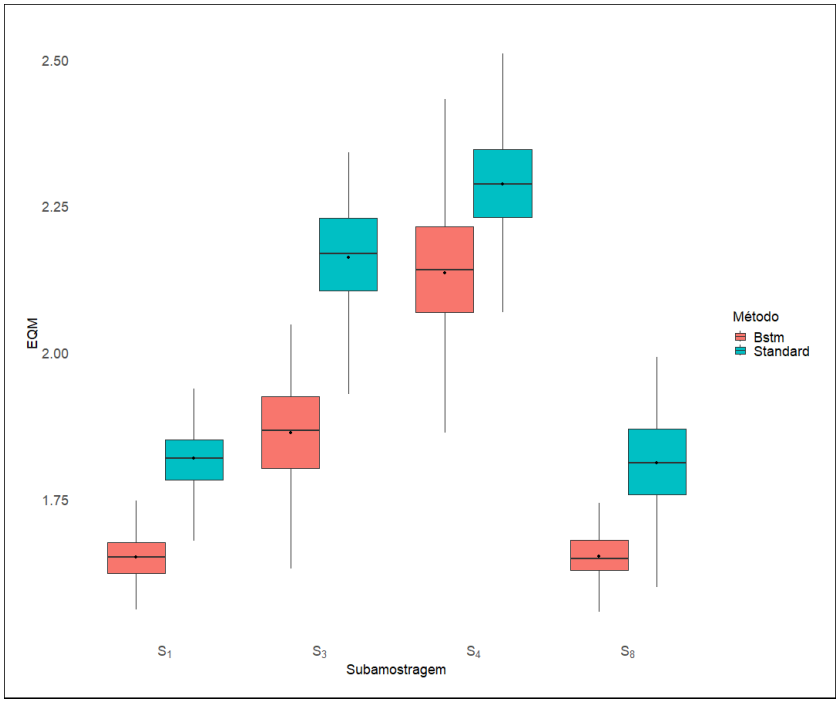

Figura 3.24: EQM para $\mathbf{\Lambda}$ estimada pelo modelo bayesiano com BSSVS nos 4 vieses.

Observa-ser que o método BSTM teve um desempenho superior ao método Standard para ambos os vieses.

3.6 Discussão

Os modelos filogeográficos discretos são amplamente utilizados em estudos relacionados à compreensão das dispersões geográficas dos vírus e dos seus processos migratórios.

Esses modelos tem sido amplamente aplicados, como por exemplo em estudos com o vírus da SARS-CoV-2 (Lemey et al., 2009) e o Ebola (De Maio et al., 2015). No entanto, apesar de sua grande utilização, esses estudos são impactados por problemas relacionados ao viés de amostragem, o que pode ocasionar problemas de inferência.

Diante desse contexto, Vargas (2023) propôs uma abordagem para corrigir o viés amostral chamada *Bias-correcting Subsampling Trait Model* (BSTM), que se baseia no conceito de modelos de mistura de subárvores amostradas de acordo com informações externas representando uma aproximação para a prevalência real das localidades. Além de atuar na correção do viés, o BSTM possui a característica de não descartar dados, baseando-se no conceito de subamostragem, sendo útil em situações de dados limitados. Para compreender e estudar os impactos do método BSTM, foram realizadas simulações comparando o método padrão, ou seja, sem a utilização de pesos para a correção de viés amostral (Standard), com o método que utiliza pesos para essa correção (BSTM). Em cada um dos métodos, foram analisados os distintos comportamentos de migração, para os quais foi multiplicado um escalar η na matriz de taxas, definindo diferentes regimes de velocidade de migração. Também foram considerandos diferentes cenários de vieses.

Nos dois cenários, foram utilizados métodos de inferência por máxima verossimilhança, inferência bayesiana para o modelo de cadeia de Markov padrão e o mesmo modelo com BSSVS. Nessas análises, foram considerados alguns indicadores: acurácia na reconstrução dos nós internos, acurácia na reconstrução da raiz, erro quadrático médio para a matriz de taxas (nos três métodos de estimação) e para as frequências de equilíbrio (realizado no modelo bayesiano original), e acurácia para as indicadoras (realizada para o modelo com BSSVS).

Nos resultados obtidos por máxima verossimilhança com valor de $\eta = 0.1$, ou seja, para comportamentos de migração muito lentos, o modelo padrão (Standard) alcançou boas estimativas na acurácia da reconstrução dos nós internos, acurácia na raiz e na matriz de taxas para todos os vieses apresentado. Nesse contexto o BSTM teve resultados inferiores.

Para valores de $\eta \geq 10$, ou seja, para comportamentos de migração muito rápidos, o método BSTM apresentou um desempenho significativamente melhor quando comparado com o método padrão (Standard), especialmente em vieses que o desequilíbrio não favorecia a localização de maior frequência populacional, para indicadores de acurácia na reconstrução dos nós internos e na raiz.

Vale destacar que o método padrão apresentou bom desempenho para o viés relacionado à distribuição de equilíbrio (ausência de viés) na acurácia da reconstrução dos nós internos em quase todos os cenários. Contudo, para $\eta = 100$, no indicador de EQM da matriz de taxas, ambos os métodos tiveram um desempenho muito ruim.

No modelo Bayesiano original, para $\eta = 0.1$. O modelo padrão alcançou melhores estimativas que o método BSTM. Exceto no erro quadrático médio da matriz das frequências de equilíbrio, nesse recorte, o viés em que o desequilíbrio não favorece a localização de maior frequência populacional o método BSTM teve melhor desempenho.

À medida que aumentamos os valores de η, ou seja, para comportamentos migratórios mais rápidos, observa-se que, nas estimativas relacionadas à acurácia na reconstrução dos nós internos, o modelo BSTM apresenta valores equivalentes ao método padrão. Nas estimativas de acurácia da raiz, o método BSTM teve melhores resultados para vieses com valores mais desequilibrados para a localização de maior frequência populacional (A). No erro quadrático médio da matriz de taxas, o método BSTM apresentou para todos os vieses ótimo desempenho. Em relação ao EQM das frequências de equilíbrio, o BSTM apresentou ótimo desempenho para os vieses com desequilíbrio contra o estado que possui maior localização de frequência populacional.

Para o modelo Bayesiano por BSSVS, não consideramos múltiplos valores de η. Em relação à acurácia na reconstrução dos nós internos e na raiz, os métodos apresentaram desempenhos iguais. Para a acurácia da matriz de indicadoras e o EQM para a matriz de taxas o método BSTM teve desempenho superior ao método padrão (Standard).

Nossas simulações evidenciam o problema de viés de amostragem. Isso pode ser visto em como a estimação de parâmetro com o método Standard sempre sofre quando é deslocada da distribuição de equilíbrio (que também corresponde às frequências populacionais na maioria dos cenários). Observa-se que a qualidade das reconstruções mostra ser menos afetada pelos vieses.

Para η muito pequeno o método BSTM tem desempenho inferior ao Standard para quase todos os cenários avaliados. Portanto, para conjuntos de dados que

sugerem pouca migração, o uso do BSTM talvez não seja o mais indicado.

Considerando os métodos de estimação, no método de máxima verossimilhança o modelo BSTM não se mostrou superior ao método padrão, de modo que o uso deste não parece ser justificado. Já para os métodos Bayesianos, quando temos uma quantidade considerável de migração, observamos que para a reconstrução da árvore o BSTM parece ser equivalente, em situações pontuais de viés, ao método padrão. Entretanto, se o objetivo é estimar os parâmetros do modelo de migração, podemos observar que o método BSTM apresentou performance claramente superior ao modelo padrão nessas situações. Assim, o uso do BSTM parece se justificar para as análises Bayesianas em conjuntos de dados com quantidades não desprezíveis de migração, principalmente se o foco é estimação de parâmetros do modelo de migração.

Referências Bibliográficas

Bengtsson, H. (2023). *matrixStats: Functions that Apply to Rows and Columns of Matrices (and to Vectors)*. R package version 1.1.0.

Bloomquist, E. W., Lemey, P., and Suchard, M. A. (2010). Three roads diverged? routes to phylogeographic inference. *Trends in Ecology & Evolution*, 25(11):626–632.

De Maio, N., Wu, C.-H., O'Reilly, K. M., and Wilson, D. (2015). New routes to phylogeography: a bayesian structured coalescent approximation. *PLoS genetics*, 11(8):e1005421.

Drummond, A. e. a. (2012). Bayesian phylogenetics with beauti and the beast 1.7. *BMC evolutionary biology*, 29:1969–1973.

Felsenstein, J. (1981). Evolutionary trees from dna sequences: a maximum-likelihood approach. *Journal of Molecular Evolution*, 17:368–376.

Grothendieck, G. (2018). *gsubfn: Utilities for Strings and Function Arguments*. R package version 0.7.

Hamner, B. and Frasco, M. (2018). *Metrics: Evaluation Metrics for Machine Learning*. R package version 0.1.4.

Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications.

Holst, K. K. and Budtz-Joergensen, E. (2013). Linear latent variable models: The lava-package. *Computational Statistics*, 28(4):1385–1452.

Holst, K. K. and Budtz-Joergensen, E. (2020). A two-stage estimation procedure for non-linear structural equation models. *Biostatistics*, 21(4):676–691.

Izrailev, S. (2023). *tictoc: Functions for Timing R Scripts, as Well as Implementations of "Stack"and "StackList"Structures*. R package version 1.2.

Kalkauskas, A., Perron, U., Sun, Y., Goldman, N., Baele, G., Guindon, S., and De Maio, N. (2021). Sampling bias and model choice in continuous phylogeography: Getting lost on a random walk. *PLOS Computational Biology*, 17(1):e1008561.

Kuhn and Max (2008). Building predictive models in r using the caret package. *Journal of Statistical Software*, 28(5):1–26.

Layan, M., Müller, N. F., Dellicour, S., De Maio, N., Bourhy, H., Cauchemez, S., and Baele, G. (2023). Impact and mitigation of sampling bias to determine viral spread: evaluating discrete phylogeography through ctmc modeling and structured coalescent model approximations. *Virus Evolution*, 9(1):vead010.

Lemey, P., Hong, S. L., Hill, V., Baele, G., Poletto, C., Colizza, V., O'toole, Á., McCrone, J. T., Andersen, K. G., Worobey, M., et al. (2020). Accommodating individual travel history and unsampled diversity in bayesian phylogeographic inference of sars-cov-2. *Nature communications*, 11(1):5110.

Lemey, P., Rambaut, A., Drummond, A. J., and Suchard, M. A. (2009). Bayesian phylogeography finds its roots. *PLoS computational biology*, 5(9):e1000520.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092.

Moreira, D. (2015). *Phylogenetic Tree*. Springer Berlin Heidelberg, Berlin, Heidelberg.

Paradis, E. and Schliep, K. (2019). ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, 35:526–528.

Plummer, M., Best, N., Cowles, K., and Vines, K. (2006). Coda: Convergence diagnosis and output analysis for mcmc. *R News*, 6(1):7–11.

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Revell, L. J. (2012). phytools: An R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution*, 3:217–223.

Rudis, B. (2020). *hrbrthemes: Additional Themes, Theme Components and Utilities for 'ggplot2'*. R package version 0.8.0.

Statisticat and LLC. (2021). *LaplacesDemon: Complete Environment for Bayesian Inference*. R package version 16.1.6.

Temple Lang, D. (2023). *XML: Tools for Parsing and Generating XML Within R and S-Plus*. R package version 3.99-0.15.

Vargas, F. (2023). Correction of sampling bias through a computationally efficient subsampling strategy in discrete bayesian phylogeographic models.

Warnes, G. R., Gorjanc, G., Magnusson, A., Andronic, L., Rogers, J., MacQueen, D., and Korosec, A. (2023). *gdata: Various R Programming Tools for Data Manipulation*. R package version 3.0.0.

Wickham, H. (2007). Reshaping data with the reshape package. *Journal of Statistical Software*, 21(12):1–20.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Wickham, H. (2023). *stringr: Simple, Consistent Wrappers for Common String Operations*. R package version 1.5.1.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., and Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686.

Wickham, H., François, R., Henry, L., Müller, K., and Vaughan, D. (2023a). *dplyr: A Grammar of Data Manipulation*. R package version 1.1.4.

Wickham, H., Hester, J., and Ooms, J. (2023b). *xml2: Parse XML*. R package version 1.3.5.

Wickham, H., Pedersen, T. L., and Seidel, D. (2023c). *scales: Scale Functions for Visualization*. R package version 1.3.0.

Yang, Z. (2006). *Computational Molecular Evolution*, volume 1. Oxford University Press.