



MÁRCIO ROGER PIAGIO

**Modelos *N-mixture* Dinâmicos para Estimação
da Subnotificação de Acidentes de Trabalho
com Exposição a Material Biológico**

Maringá - Paraná
2026

MÁRCIO ROGER PIAGIO

**Modelos *N-mixture* Dinâmicos para Estimação da
Subnotificação de Acidentes de Trabalho com Exposição
a Material Biológico**

Dissertação apresentada ao Programa de Pós-graduação em Bioestatística do Centro de Ciências Exatas da Universidade Estadual de Maringá como requisito parcial para obtenção do título de mestre em Bioestatística.

Orientador: Prof. Dr. Diogo Francisco Rossoni

Coorientador: Prof. Dr. George Lucas Moraes Pezzott

Membro do PBE: Prof. Dr. Diego Corrêa Alves

Membro Externo: Prof. Dr. Danilo Alvares da Silva

Universidade Estadual de Maringá - UEM

Departamento de Estatística - DES

Programa de Pós-Graduação em Bioestatística

Maringá - Paraná

2026

Dados Internacionais de Catalogação-na-Publicação (CIP)
(Biblioteca Central - UEM, Maringá - PR, Brasil)

P581m

Piagio, Márcio Roger

Modelos *N-mixture* dinâmicos para estimação da subnotificação de acidentes de trabalho com exposição a material biológico / Márcio Roger Piagio. -- Maringá, PR, 2026. 67 f. : il. color., figs., tabs.

Orientador: Prof. Dr. Diogo Francisco Rossoni.

Coorientador: Prof. Dr. George Lucas Moraes Pezzott.

Dissertação (mestrado) - Universidade Estadual de Maringá, Centro de Ciências Exatas, Departamento de Estatística, Programa de Pós-Graduação em Bioestatística, 2026.

1. Sub-registro de dados - Acidentes de trabalho. 2. Modelos hierárquicos. 3. Inferência Bayesiana. 4. Vigilância epidemiológica. 5. Agravos ocupacionais. I. Rossoni, Diogo Francisco, orient. II. Pezzott, George Lucas Moraes, coorient. III. Universidade Estadual de Maringá. Centro de Ciências Exatas. Departamento de Estatística. Programa de Pós-Graduação em Bioestatística. IV. Título.

CDD 23.ed. 519.5

MÁRCIO ROGER PIAGIO

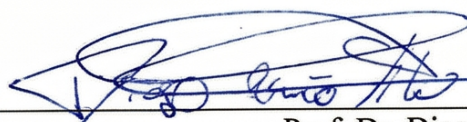
**Modelos N-mixture Dinâmicos para Estimação da Subnotificação de
Acidentes de Trabalho com Exposição a Material Biológico**

Dissertação apresentada ao Programa de Pós-Graduação em Bioestatística do Centro de Ciências Exatas da Universidade Estadual de Maringá, como requisito parcial para a obtenção do título de Mestre em Bioestatística.

BANCA EXAMINADORA



Prof. Dr. Diogo Francisco Rossoni
Universidade Estadual de Maringá – PBE/UEM



Prof. Dr. Diego Corrêa Alves
Universidade Estadual de Maringá – PBE/UEM



Prof. Dr. Danilo Alvares da Silva
MRC Biostatistics Unit University of Cambridge

Maringá, 27 de fevereiro de 2026.

Dedico esta dissertação à minha filha, Giovana, fonte de inspiração e motivo para seguir sempre em frente.

À minha mãe, Maria, por tudo o que representou em minha vida e por todos os ensinamentos que permanecem comigo.

A Guarulhos — Cumbica, lugar onde fui criado, e aos amigos que lá fiz e que permanecem parte importante da minha história.

AGRADECIMENTOS

Agradeço, inicialmente, à sociedade brasileira e, em especial, aos contribuintes do Estado do Paraná, cujo investimento na educação pública viabiliza a existência das universidades e o desenvolvimento da pesquisa científica nacional.

À cidade de Maringá, por ter me recebido e proporcionado o ambiente ideal para o meu desenvolvimento pessoal e acadêmico ao longo destes anos. Sou grato à hospitalidade desta terra e à Universidade Estadual de Maringá (UEM), bem como a todos os seus servidores, pelo suporte institucional e técnico que viabilizaram a realização deste trabalho.

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), pelo fomento concedido e pelo essencial apoio ao incentivo à formação de pesquisadores no Brasil.

Ao Prof. Dr. Diogo Francisco Rossoni, meu orientador, expresse minha sincera gratidão pela orientação, pela confiança e pelo constante incentivo.

Ao Prof. Dr. George Lucas Moraes Pezzott, agradeço pela coorientação, pelas discussões acadêmicas fundamentais e pelas valiosas contribuições ao longo de todo o desenvolvimento deste trabalho.

Aos membros da banca examinadora, Prof. Dr. Diego Corrêa Alves e Prof. Dr. Danilo Alvares da Silva, pela disponibilidade em avaliar este trabalho e pelas valiosas sugestões para o aprimoramento desta pesquisa.

Aos professores do Programa de Pós-Graduação em Bioestatística (PBE) e do Departamento de Estatística (DES) da Universidade Estadual de Maringá, pelos ensinamentos compartilhados e pela sólida contribuição à minha formação.

Aos colegas do PBE e DES, pela convivência, trocas de conhecimento e apoio mútuo. Em ordem alfabética: Alexandre (Prince), Carla, Eduardo (Pimpão), Eric (Zé), Haward, Ícaro (Ursão), Jessica, João, Juliana (Ju), Pedro, Raphael, Ricardo (CB), Rodolfo (Mogri), Rosana (Rose), Samuel (Kid).

À minha família, por ser o alicerce de toda esta caminhada. Agradeço ao meu pai, Antonio; aos meus irmãos, Magali e Marcos; e aos meus sobrinhos, Amanda, Anthony, Eduardo e Olivia, pelo amor, incentivo e compreensão constantes.

RESUMO

A subnotificação de acidentes de trabalho com exposição a material biológico constitui um desafio relevante para a vigilância epidemiológica, pois compromete a estimativa da magnitude real do agravo e a formulação de políticas públicas adequadas. Esta dissertação emprega um modelo hierárquico do tipo *N-mixture* dinâmico para estimar a ocorrência latente desses eventos e quantificar a fração não registrada no Sistema de Informação de Agravos de Notificação (SINAN), no período de 2014 a 2024. A estrutura do modelo distingue explicitamente o processo latente de ocorrência do processo observacional de notificação, adotando uma parametrização multiplicativa apropriada ao contexto de eventos incidentes mensais. A inferência foi conduzida sob abordagem Bayesiana por meio de métodos de Monte Carlo via Cadeias de Markov (MCMC), implementados no software *JAGS*. Estudos de simulação de Monte Carlo foram conduzidos em dois contextos: (i) cenários controlados, com parâmetros previamente especificados, e (ii) cenários calibrados a partir de parâmetros estimados com base nos dados reais. Em ambos os casos, observou-se desempenho inferencial satisfatório do modelo, com adequada recuperação dos parâmetros estruturais mesmo sob condições de detecção imperfeita. Na aplicação aos dados do SINAN, observaram-se níveis expressivos de subnotificação entre Profissionais de Saúde (nível superior) e Técnicos e Auxiliares da Saúde, bem como relativa estabilidade temporal do processo latente ao longo do período analisado. Os resultados indicam que o modelo *N-mixture* dinâmico constitui uma ferramenta estatisticamente consistente e epidemiologicamente informativa para investigar subnotificação em séries temporais de vigilância, contribuindo para o aprimoramento da análise quantitativa de agravos ocupacionais.

Palavras-chave: Sub-registro; Modelos hierárquicos; Inferência Bayesiana; Vigilância epidemiológica; Agravos ocupacionais; Epidemiologia ocupacional.

ABSTRACT

Underreporting of occupational accidents involving exposure to biological material represents a major challenge for epidemiological surveillance, as it compromises the estimation of the true magnitude of such events and the formulation of effective public health policies. This study proposes a hierarchical dynamic N -mixture model to estimate the latent occurrence of these events and to quantify the fraction of unreported cases in the Brazilian Notifiable Diseases Information System (SINAN), from 2014 to 2024. The model explicitly separates the latent process governing event occurrence from the observation process associated with reporting, adopting a multiplicative parameterization suitable for monthly incident counts. Inference was conducted within a Bayesian framework using Markov Chain Monte Carlo (MCMC) methods implemented in *JAGS*. A Monte Carlo simulation study was carried out under two settings: (i) controlled scenarios with predefined parameter values and (ii) scenarios calibrated using parameters estimated from real data. In both settings, the model demonstrated satisfactory inferential performance, with accurate recovery of structural parameters even under imperfect detection. Application to SINAN data revealed substantial levels of underreporting among higher-level healthcare professionals and technical/auxiliary health workers, as well as relative temporal stability in the latent process over the study period. These findings indicate that the proposed dynamic N -mixture model provides a statistically consistent and epidemiologically informative framework for investigating underreporting in surveillance time series, contributing to improved quantitative analyses of occupational health outcomes.

Key-words: Underreporting; Hierarchical models; Bayesian inference; Epidemiological surveillance; Occupational health.

LISTA DE ILUSTRAÇÕES

Figura 1 – Tendência mensal de notificações por categoria ocupacional (CBO).	37
Figura 2 – Evolução temporal por estrato para Técnicos e Auxiliares da Saúde.	38
Figura 3 – Evolução temporal por estrato para Profissionais de Saúde (nível superior).	39
Figura 4 – Série temporal das notificações observadas (Y_t), da ocorrência total estimada (N_t) e da subnotificação estimada (U_t) de acidentes de trabalho com exposição a material biológico, com intervalos de credibilidade de 95%, para dois grupos ocupacionais do setor saúde (2014–2024).	41
Figura 5 – Diagnóstico de convergência MCMC: Técnicos e Auxiliares da Saúde, Para o período de 2014–2019.	57
Figura 6 – Diagnóstico de convergência MCMC: Técnicos e Auxiliares da Saúde, para o período de 2020–2021.	58
Figura 7 – Diagnóstico de convergência MCMC: Técnicos e Auxiliares da Saúde, para o período de 2022–2024.	58
Figura 8 – Diagnóstico de convergência MCMC: Profissionais de Saúde (Nível Superior), para o período de 2014–2019.	59
Figura 9 – Diagnóstico de convergência MCMC: Profissionais de Saúde (Nível Superior), para o período de 2020–2021.	59
Figura 10 – Diagnóstico de convergência MCMC: Profissionais de Saúde (Nível Superior), para o período de 2022–2024.	60

LISTA DE TABELAS

Tabela 1 – Desempenho inferencial dos parâmetros do modelo no estudo de simulação.	34
Tabela 2 – Resultados do estudo de simulação para as quantidades agregadas N_{total} e U_{total} .	35
Tabela 3 – Distribuição das notificações de acidentes com material biológico segundo características sociodemográficas e ocupacionais.	36
Tabela 4 – Estimativas <i>a posteriori</i> dos modelos <i>N-mixture</i> dinâmicos por grupo ocupacional e período (estratos: sexo, raça e idade).	40
Tabela 5 – Diagnóstico de convergência via Gelman–Rubin ($\hat{R}$) por grupo ocupacional e período. Valores apresentados como estimativa pontual (limite superior do IC).	42
Tabela 6 – Tamanho efetivo da amostra (ESS) por grupo ocupacional e período.	43
Tabela 7 – Desempenho inferencial dos parâmetros do modelo nos cenários calibrados a partir dos parâmetros estimados nos dados reais.	44
Tabela 8 – Desempenho inferencial para os totais de abundância e subnotificação nos cenários calibrados.	45

LISTA DE ABREVIATURAS E SIGLAS

ABNT	Associação Brasileira de Normas Técnicas
CBO	Classificação Brasileira de Ocupações
COVID-19	<i>Coronavirus Disease 2019</i> (Doença pelo Coronavírus 2019)
DIC	<i>Deviance Information Criterion</i> (Critério de Informação do Desvio)
ESS	<i>Effective Sample Size</i> (Tamanho Efetivo da Amostra)
IC	Intervalo de Credibilidade
JAGS	<i>Just Another Gibbs Sampler</i>
MCMC	<i>Markov Chain Monte Carlo</i> (Cadeias de Markov Monte Carlo)
$\hat{R}$	Fator de Redução Potencial de Gelman-Rubin
SINAN	Sistema de Informação de Agravos de Notificação
SUS	Sistema Único de Saúde

SUMÁRIO

1	Introdução	12
2	Metodologia	17
2.1	Delineamento do estudo	17
2.2	Base de dados	18
2.3	Procedimentos de análise descritiva	20
2.4	Modelo <i>N-mixture</i>	20
2.4.1	Modelo para população fechada	21
2.4.2	Modelo para população aberta	21
2.4.3	Modificação adotada neste estudo	22
2.5	Implementação via JAGS	23
2.6	Estudo de simulação	29
3	Resultados	33
3.1	Estudo de Simulação	33
3.2	Aplicação aos Dados do SINAN	36
3.3	Simulação para Validação do Cenário Empírico	43
4	Discussão	46
4.1	Desempenho inferencial do modelo	47
4.2	Interpretação epidemiológica dos resultados	48
5	Conclusão	50
	Referências	53
	Apêndices	56
	APÊNDICE A Diagnóstico de Convergência das Cadeias	57
	APÊNDICE B Implementação do Modelo com dados Simulados	61

INTRODUÇÃO

Os acidentes de trabalho com exposição a material biológico constituem um importante agravo à saúde do trabalhador, especialmente no contexto dos serviços de saúde, onde a manipulação de sangue e outros fluidos potencialmente contaminados integra a rotina profissional (ARAÚJO et al., 2012; GOMES; CALDAS, 2019). Tais eventos expõem trabalhadores ao risco de infecções por agentes como HIV e vírus das hepatites B e C, além de produzirem impactos psicológicos, institucionais e econômicos relevantes (CARNEIRO et al., 2016; GOMES; CALDAS, 2017).

No Brasil, esses agravos integram a lista de notificações compulsórias registradas no Sistema de Informação de Agravos de Notificação (SINAN), instrumento central da vigilância epidemiológica no âmbito do Sistema Único de Saúde. Entre 2007 e 2014, foram registrados 284.877 acidentes envolvendo trabalhadores com ocupação conhecida (SOARES; SCHOEN; BENELLI, 2019). Embora expressivo, esse número provavelmente não reflete a totalidade dos eventos ocorridos.

A literatura reconhece que sistemas baseados em notificação passiva estão sujeitos a subregistro e subnotificação (MARINHO, 2018; FERREIRA et al., 2025). Fatores como sobrecarga de trabalho, falhas administrativas, desconhecimento de protocolos, medo de repercussões trabalhistas e baixa percepção de risco contribuem para que parte dos acidentes não seja formalmente registrada (NOGUEIRA et al., 2016). Além disso, inconsistências no preenchimento das fichas e elevada proporção de campos ignorados comprometem a qualidade da informação epidemiológica (SEBEN; MORETTO, 2022).

A subnotificação compromete a interpretação direta das contagens observadas, uma vez que o número registrado não necessariamente reflete a magnitude real do fenômeno. Do ponto de vista epidemiológico, a ausência de registro não implica ausência de ocorrência, mas sim limitação do mecanismo de captação do sistema (HOOK; REGAL, 1995). Sob perspectiva estatística, essa situação pode ser compreendida como um problema de detecção imperfeita, no qual o número observado constitui apenas uma realização parcial de um processo latente de ocorrência (DENNIS; MORGAN; RIDOUT, 2015; BARKER et al., 2018).

A distinção entre o processo que governa a ocorrência real dos eventos e o processo que determina sua observação constitui princípio central da modelagem hierárquica (WIKLE, 2003; CRESSIE; WIKLE, 2011). Modelos hierárquicos probabilísticos oferecem uma estrutura adequada para representar essa dissociação entre ocorrência e observação, permitindo explicitar o mecanismo de detecção imperfeita no processo inferencial.

No contexto de contagens sujeitas à observação incompleta, os modelos do tipo *N-mixture*, propostos por Royle (2004), formalizam essa separação ao modelar simultaneamente a abundância latente e a probabilidade de detecção a partir de contagens replicadas. A informação sobre a abundância é extraída da estrutura de correlação entre observações repetidas, mesmo na ausência de identificação individual dos casos.

A extensão dinâmica desenvolvida por Dail e Madsen (2011) incorporou variação temporal explícita na abundância latente, permitindo modelar populações abertas nas quais a quantidade real de eventos evolui ao longo do tempo. Essa estrutura possibilita distinguir variações na incidência real de alterações na sensibilidade do sistema de notificação, aspecto particularmente relevante em séries epidemiológicas longas.

Modelos estatísticos tradicionais, como a técnica de captura-recaptura (LINCOLN, 1930; OTIS et al., 1978), vêm sendo utilizados para estimar o tamanho de populações ocultas, inclusive em estudos epidemiológicos. No entanto, esses modelos baseiam-se em pressupostos difíceis de serem satisfeitos na prática, como a independência entre as fontes de dados e a possibilidade de recaptura do mesmo indivíduo, o que raramente se aplica a sistemas de notificação de base passiva. Alternativas metodológicas oriundas da Ecologia têm ganhado espaço nesse contexto. Os modelos de remoção, por exemplo, estimam a abundância com base em contagens sucessivas de indivíduos removidos (ou notificados), assumindo que a população observada diminui a cada tentativa (ZIPPIN, 1956). Já os modelos *N-mixture* utilizam uma estrutura hierárquica para estimar separadamente a abundância latente de eventos e a probabilidade de sua detecção. Essa abordagem foi ampliada por Dail e Madsen (2011) para incorporar dinâmicas populacionais ao longo do tempo, resultando em modelos dinâmicos aplicáveis a contextos com variação temporal na incidência.

Essas ferramentas, ainda que originalmente aplicadas à Ecologia, têm se mostrado promissoras para lidar com os desafios da subnotificação em Epidemiologia. Aplicações recentes incluem a estimação da prevalência real de clamídia (BRINTZ; FUENTES; MADSEN, 2018), a avaliação da sensibilidade de testes para SARS-CoV-2 (SANDERLIN et al., 2021) e a investigação da subnotificação da COVID-19 (PARKER et al., 2024). A principal vantagem desses modelos reside na sua capacidade de lidar com dados fragmentados e contagens agregadas, permitindo a estimação de parâmetros epidemiológicos mesmo em situações de observação imperfeita.

Nesta dissertação, propõe-se aplicar o modelo *N-mixture* dinâmico à análise dos acidentes de trabalho com exposição a material biológico registrados no SINAN no período de

2014 a 2024. O estudo concentra-se em dois grandes grupos ocupacionais: Profissionais de Saúde (nível superior) e Técnicos e Auxiliares da Saúde, considerando diferenças estruturais tanto no risco de exposição quanto na probabilidade de notificação.

A modelagem é conduzida sob abordagem Bayesiana, permitindo a estimação conjunta dos componentes latente e observacional por meio de métodos baseados em cadeias de Markov Monte Carlo (MCMC). Essa estratégia possibilita obter distribuições *a posteriori* completas para os parâmetros do modelo e estimativas derivadas da magnitude real do agravo. Ao modelar explicitamente a subnotificação como componente estrutural do sistema, este trabalho busca produzir estimativas mais consistentes da carga real de acidentes com exposição a material biológico no Brasil, contribuindo metodologicamente para o aprimoramento da vigilância epidemiológica em saúde do trabalhador.

Além da aplicação empírica ao contexto da vigilância epidemiológica, este estudo contribui metodologicamente ao explorar o potencial dos modelos *N-mixture* dinâmicos para a análise de sistemas de notificação sujeitos à detecção imperfeita. Ao adaptar essa estrutura de modelagem, originalmente desenvolvida no contexto da Ecologia, para dados provenientes de sistemas de vigilância em saúde, o trabalho evidencia a utilidade de modelos hierárquicos de contagem para a estimação indireta de eventos não observados. Dessa forma, a dissertação amplia o campo de aplicação desses modelos no domínio da Epidemiologia, oferecendo uma abordagem estatisticamente fundamentada para quantificar subnotificação em séries temporais baseadas em dados administrativos.

OBJETIVOS

O presente trabalho tem como objetivo desenvolver e aplicar um modelo hierárquico do tipo *N-mixture* dinâmico, sob abordagem Bayesiana, para estimar a ocorrência latente e quantificar a subnotificação de acidentes de trabalho com exposição a material biológico entre profissionais de saúde registrados no Sistema de Informação de Agravos de Notificação (SINAN) no período de 2014 a 2024. A abordagem proposta busca distinguir explicitamente o processo real de ocorrência dos eventos do processo observacional de notificação presente no sistema de vigilância.

De forma específica, pretende-se:

- Estruturar formalmente um modelo hierárquico de contagem que represente a ocorrência real dos acidentes como um processo latente dinâmico e as notificações registradas como um processo observacional sujeito à detecção imperfeita;
- Especificar e implementar o modelo *N-mixture* dinâmico sob abordagem Bayesiana, utilizando métodos de amostragem MCMC para estimar os parâmetros associados à

intensidade inicial do processo latente (λ), à dinâmica temporal (δ) e à probabilidade de detecção (p);

- Quantificar a subnotificação como a diferença entre o número estimado de eventos reais ($N_{i,t}$) e o número de eventos efetivamente registrados no sistema ($Y_{i,t}$), obtendo estimativas do total latente de ocorrências (N_{total}) e do total de eventos não notificados (U_{total});
- Avaliar o desempenho inferencial do modelo por meio de um estudo de simulação Monte Carlo, investigando a recuperação dos parâmetros estruturais e das quantidades derivadas sob diferentes cenários de detecção;
- Realizar um estudo de simulação calibrado a partir dos parâmetros estimados nos dados reais, com o objetivo de avaliar a estabilidade do procedimento inferencial em regimes paramétricos próximos à aplicação empírica;
- Aplicar o modelo aos dados do SINAN para analisar a dinâmica temporal da ocorrência latente dos acidentes e estimar a magnitude da subnotificação ao longo do período de estudo;
- Comparar os resultados entre grupos ocupacionais, particularmente entre profissionais de saúde de nível superior e técnicos ou auxiliares da saúde, avaliando diferenças na intensidade latente, na dinâmica temporal e na probabilidade de notificação.

ESTRUTURA DO TRABALHO

Esta dissertação está organizada em cinco capítulos, estruturados de modo a articular a contextualização do problema, o desenvolvimento metodológico da modelagem proposta e a análise inferencial aplicada aos dados do Sistema de Informação de Agravos de Notificação (SINAN).

O Capítulo 1 apresenta a contextualização epidemiológica dos acidentes de trabalho com exposição a material biológico, discutindo a subnotificação como fenômeno recorrente em sistemas de vigilância baseados em notificação passiva. Nesse capítulo, delimita-se o problema de pesquisa e discute-se a necessidade de abordagens estatísticas capazes de distinguir entre o processo real de ocorrência dos eventos e o processo observacional de registro no sistema de vigilância.

O Capítulo 2 apresenta o referencial metodológico da dissertação. Inicialmente, são discutidos os fundamentos dos modelos hierárquicos de contagem e a formulação original dos modelos *N-mixture*, bem como sua extensão dinâmica para populações abertas. Em seguida, é apresentada a parametrização multiplicativa adotada neste estudo, juntamente com os pressupostos probabilísticos, a estrutura latente do modelo e as quantidades derivadas de interesse

epidemiológico. Também é descrito o procedimento inferencial sob abordagem Bayesiana, implementado por meio de métodos de amostragem MCMC. Na sequência, são apresentados os dados provenientes do SINAN, incluindo os critérios de seleção, tratamento e organização da base analítica, bem como as estratégias de estratificação por grupo ocupacional e características sociodemográficas que fundamentam a modelagem proposta.

O Capítulo 3 apresenta os resultados obtidos no estudo, estruturados em três etapas complementares. Inicialmente, são detalhados os resultados do estudo de simulação Monte Carlo, conduzido para avaliar o desempenho inferencial do modelo sob diferentes cenários controlados de detecção imperfeita. Na sequência, apresenta-se a aplicação do modelo *N-mixture* dinâmico aos dados do SINAN, incluindo as estimativas dos parâmetros estruturais, a reconstrução da abundância latente e a quantificação da subnotificação no período analisado. Por fim, apresenta-se um estudo de simulação a partir dos parâmetros estimados na aplicação empírica, permitindo validar o comportamento e a robustez do modelo nos regimes paramétricos específicos observados nos dados reais.

O Capítulo 4 apresenta a discussão dos principais resultados obtidos no estudo. Nesse capítulo, os achados provenientes das simulações e da aplicação empírica são interpretados à luz da literatura sobre modelagem hierárquica e vigilância epidemiológica. São discutidas as implicações metodológicas da separação entre processo latente de ocorrência e processo observacional de notificação, bem como as interpretações epidemiológicas das estimativas obtidas para os acidentes de trabalho com exposição a material biológico. Também são analisadas as limitações da abordagem adotada e possíveis extensões da modelagem.

Por fim, o Capítulo 5 apresenta as conclusões da dissertação, sintetizando os principais resultados e destacando as contribuições metodológicas e aplicadas do estudo. Nesse capítulo, retomam-se os objetivos da pesquisa à luz das evidências obtidas e discutem-se perspectivas para trabalhos futuros, incluindo possíveis extensões da estrutura hierárquica do modelo, incorporação de covariáveis explicativas e aprimoramentos na modelagem do processo de detecção.

METODOLOGIA

2.1 DELINEAMENTO DO ESTUDO

Este estudo é observacional, quantitativo, de delineamento ecológico em série temporal. A unidade de análise é composta por agregações mensais de notificações de acidentes de trabalho com exposição a material biológico registradas no *Sistema de Informação de Agravos de Notificação* (SINAN), abrangendo todo o território brasileiro no período de janeiro de 2014 a dezembro de 2024.

O foco recai sobre profissionais de saúde em dois grandes grupos ocupacionais segundo a Classificação Brasileira de Ocupações (CBO): (i) Profissionais de Saúde de nível superior, códigos iniciados por 223; e (ii) Técnicos e Auxiliares da Saúde, códigos iniciados por 32. Essa estratificação atende à hipótese de possíveis diferenças estruturais tanto na exposição ao risco quanto na dinâmica temporal dos agravos.

A longitudinalidade dos dados permite modelar a evolução das notificações ao longo do tempo, reconhecendo que o número observado pode subestimar o efetivamente ocorrido devido à subnotificação. Para lidar com esse problema, adota-se um modelo hierárquico de contagem que distingue explicitamente o processo latente de ocorrência do agravo do processo probabilístico de detecção ou notificação.

Para fins de modelagem, assume-se homogeneidade da probabilidade de notificação entre as unidades analíticas (estratos) e ao longo do tempo. Na ausência de covariáveis informativas para o processo de detecção, supõe-se uma probabilidade comum p para todos os estratos e períodos. Essa simplificação visa garantir identificabilidade e estabilidade inferencial; assim, a interpretação das diferenças entre estratos é atribuída ao processo latente de ocorrência.

Esse delineamento permite investigar conjuntamente:

- a dinâmica temporal dos eventos latentes;
- o mecanismo probabilístico de notificação,

aplicando uma estrutura hierárquica apropriada para séries temporais agregadas de contagem.

2.2 BASE DE DADOS

FONTE DOS DADOS

Os dados utilizados neste estudo provêm do Sistema de Informação de Agravos de Notificação (SINAN), sistema oficial do Ministério da Saúde destinado ao registro de agravos de notificação compulsória no Brasil, integrado à vigilância epidemiológica do Sistema Único de Saúde (SUS).

Foram considerados registros de acidentes de trabalho com exposição a material biológico ocorridos entre janeiro de 2014 e dezembro de 2024. Esses eventos incluem exposições ocupacionais a sangue e outros fluidos potencialmente contaminados, decorrentes principalmente de perfurações por objetos perfurocortantes, cortes com instrumentos, respingos em mucosas ou contato com pele não íntegra, implicando risco de transmissão de agentes infecciosos como os vírus da hepatite B (HBV), hepatite C (HCV) e o vírus da imunodeficiência humana (HIV).

Essas ocorrências concentram-se especialmente entre profissionais e técnicos da área da saúde, em função da natureza assistencial das atividades exercidas. A análise longitudinal desses registros permite avaliar a dinâmica temporal dos eventos e investigar possíveis indícios de subnotificação ao longo do período analisado.

A base de dados foi obtida a partir de arquivos oficiais anonimizados. Por se tratar de dados secundários e de domínio público, são preservados os princípios éticos de confidencialidade e sigilo das informações individuais.

PROCESSAMENTO E CONSTRUÇÃO DA BASE ANALÍTICA

Os arquivos brutos em formato .dbc foram importados e consolidados no ambiente R, totalizando inicialmente 978.361 registros e 68 variáveis. A preparação dos dados incluiu as seguintes etapas:

1. Verificação de consistência dos registros, padronização de formatos e tratamento de valores ausentes ou inconsistentes.
2. Exclusão de registros com valores ignorados ou ausentes em variáveis consideradas essenciais: sexo, raça/cor, ocupação e idade.
3. Extração da idade a partir do campo numérico original, conversão para anos completos e categorização em faixas etárias.

4. Padronização da ocupação segundo os códigos da Classificação Brasileira de Ocupações (CBO), permitindo a identificação dos grupos investigados.
5. Conversão da data do acidente para formato de data e agregação temporal mensal (ano-mês), gerando a série temporal de contagens utilizada na análise.

Após a aplicação dessas etapas de limpeza, filtragem e padronização das informações, a base de dados final utilizada neste estudo passou a conter 410.879 observações. Esse conjunto de registros representa os casos que atenderam simultaneamente aos critérios de completude das variáveis essenciais e às restrições definidas para a população de estudo.

A agregação mensal das notificações constitui a unidade temporal da análise e sustenta a aplicação de modelos hierárquicos dinâmicos de contagem, mantendo a coerência temporal das unidades analíticas consideradas na modelagem.

DEFINIÇÃO DA POPULAÇÃO DE ESTUDO

Foram incluídos indivíduos entre 18 e 70 anos cujos códigos de ocupação se enquadram nos seguintes grupos da Classificação Brasileira de Ocupações (CBO):

- Profissionais de Saúde de nível superior (códigos da CBO iniciados por 223);
- Técnicos e Auxiliares da Saúde (códigos da CBO iniciados por 32).

A restrição etária busca representar a população economicamente ativa e reduzir inconsistências decorrentes de registros fora da faixa etária usual de exercício profissional. Adicionalmente, restringiu-se a escolaridade a níveis compatíveis com cada grupo ocupacional, de modo a garantir maior coerência interna da população analisada.

DESCRIÇÃO DAS VARIÁVEIS

As variáveis utilizadas na análise foram extraídas das fichas de notificação do SINAN e recodificadas para construção dos estratos demográficos e ocupacionais empregados na modelagem.

- **Data do acidente (DT_ACID):** utilizada para construir a variável temporal agregada em ano-mês.
- **Categoria ocupacional (CBO):** recodificada em dois grupos: Profissionais de Saúde (nível superior) e Técnicos e Auxiliares da Saúde.
- **Sexo (CS_SEXO):** feminino e masculino.

- **Raça/Cor (CS_RACA):** branca e preta/parda.
- **Faixa etária:** construída a partir da idade em anos completos. Para o sexo feminino foram consideradas as categorias 18–23, 24–28, 29–34, 35–39, 40–49 e 50–70 anos. Para o sexo masculino, as categorias foram 18–29, 30–39, 40–49 e 50–70 anos.

2.3 PROCEDIMENTOS DE ANÁLISE DESCRITIVA

Inicialmente realizou-se uma caracterização demográfica e ocupacional da população estudada, por meio de frequências absolutas e relativas segundo sexo, raça/cor, faixa etária e categoria ocupacional.

Em seguida, procedeu-se à análise temporal das notificações a partir da agregação mensal das contagens por estrato, examinando padrões de tendência, sazonalidade e possíveis alterações estruturais ao longo do período de 2014 a 2024.

Essa etapa permitiu também verificar a estabilidade da definição dos estratos ao longo do tempo, condição importante para sustentar a hipótese de homogeneidade do mecanismo de notificação. A análise exploratória orientou a especificação final do modelo dinâmico de contagem, buscando coerência entre a estrutura empírica dos dados e os pressupostos estatísticos da modelagem.

2.4 MODELO *N-MIXTURE*

A classe de modelos *N-mixture* constitui uma estrutura hierárquica para inferir o tamanho de uma população não observada a partir de contagens repetidas sujeitas à detecção imperfeita. Introduzida por Royle (2004) no contexto ecológico, essa abordagem decompõe o processo gerador dos dados em dois níveis: (i) um processo latente que descreve a abundância real de eventos e (ii) um processo de observação que representa como as contagens observadas correspondem a uma fração detectada dessa abundância.

A utilidade estatística dessa classe de modelos reside em modelar explicitamente o mecanismo de subobservação por meio de um parâmetro de detecção, ao mesmo tempo em que se estima a abundância latente. A formulação original pressupõe uma população fechada ao longo das réplicas observacionais. Posteriormente, Dail e Madsen (2011) estenderam essa estrutura para populações abertas, incorporando dinâmica temporal ao processo latente. Essa generalização permite realizar inferência mesmo quando a hipótese de fechamento não é plausível, possibilitando estimar simultaneamente parâmetros de dinâmica populacional e a abundância latente ao longo do tempo.

Nas subseções seguintes, apresentam-se brevemente as formulações para populações fechadas e abertas, seguidas da adaptação adotada neste estudo.

2.4.1 MODELO PARA POPULAÇÃO FECHADA

Considere I unidades amostrais independentes, indexadas por $i = 1, \dots, I$, e T períodos de observação em cada unidade, indexados por $t = 1, \dots, T$.

Seja N_i o número real (latente) de eventos ocorridos na unidade i durante o período considerado. Esse valor não é diretamente observado. Em cada período t , observa-se uma contagem $y_{i,t}$, que corresponde ao número de eventos detectados na unidade i no tempo t .

No modelo N -mixture clássico assume-se que o número real de eventos segue uma distribuição de Poisson,

$$N_i \sim \text{Poisson}(\lambda), \quad i = 1, \dots, I, \quad (2.1)$$

onde $\lambda > 0$ representa a abundância média do processo latente.

Condicionalmente ao valor latente N_i , as contagens observadas são modeladas por uma distribuição binomial,

$$y_{i,t} \mid N_i \sim \text{Binomial}(N_i, p), \quad t = 1, \dots, T, \quad (2.2)$$

em que $p \in (0, 1)$ representa a probabilidade de detecção de cada evento.

Nesse contexto, N_i corresponde ao número real de eventos na unidade i , enquanto $y_{i,t}$ representa a contagem observada desses eventos em cada réplica temporal t . A hipótese de população fechada implica que N_i permanece constante ao longo das réplicas observacionais.

Essa estrutura hierárquica permite separar o processo de ocorrência real do processo de observação, possibilitando estimar simultaneamente a abundância latente e a probabilidade de detecção. A verossimilhança marginal do modelo é obtida pela soma sobre todos os valores possíveis de N_i , resultando em uma mistura de distribuições binomiais ponderadas pela distribuição de Poisson.

2.4.2 MODELO PARA POPULAÇÃO ABERTA

Na formulação para população aberta, a abundância latente pode variar ao longo do tempo. Seja $N_{i,t}$ o número real de eventos na unidade amostral i no período t , com $i = 1, \dots, I$ e $t = 1, \dots, T$. A contagem observada correspondente é denotada por $y_{i,t}$.

No tempo inicial assume-se

$$N_{i,1} \sim \text{Poisson}(\lambda), \quad (2.3)$$

onde $\lambda > 0$ representa a intensidade média inicial do processo latente.

Para $t \geq 2$, a dinâmica da abundância pode ser descrita por dois componentes: sobrevivência de eventos previamente existentes e recrutamento de novos eventos. Assim, define-se

$$N_{i,t} = S_{i,t} + G_{i,t}, \quad (2.4)$$

onde, condicionalmente a $N_{i,t-1}$,

$$S_{i,t} \sim \text{Binomial}(N_{i,t-1}, \phi), \quad G_{i,t} \sim \text{Poisson}(\delta N_{i,t-1}).$$

Nesse caso, $S_{i,t}$ representa o número de eventos que persistem do período $t-1$ para o período t , enquanto $G_{i,t}$ representa o número de novos eventos recrutados no período t . O parâmetro $\phi \in (0, 1)$ corresponde à probabilidade de sobrevivência entre períodos consecutivos, e $\delta > 0$ representa a taxa de recrutamento per capita.

O processo de observação permanece descrito por

$$y_{i,t} \mid N_{i,t} \sim \text{Binomial}(N_{i,t}, p), \quad (2.5)$$

em que $p \in (0, 1)$ é a probabilidade de detecção de cada evento.

Essa estrutura mantém a separação entre o processo latente de abundância e o processo de observação. Diferentemente do modelo de população fechada, a abundância pode agora variar ao longo do tempo em função dos parâmetros de sobrevivência ϕ e recrutamento δ , permitindo representar cenários de crescimento, declínio ou estabilidade da população latente.

2.4.3 MODIFICAÇÃO ADOTADA NESTE ESTUDO

A modificação proposta busca preservar a estrutura hierárquica do modelo *N-mixture* dinâmico, adaptando-a ao contexto de séries temporais agregadas de notificações epidemiológicas.

No modelo original de população aberta proposto por Dail e Madsen (2011), o processo latente é decomposto em dois componentes: sobrevivência de indivíduos previamente presentes e recrutamento de novos indivíduos. Tal formulação é apropriada quando se modelam populações biológicas nas quais os mesmos indivíduos podem persistir ao longo do tempo.

No presente estudo, entretanto, a unidade modelada não corresponde a indivíduos persistentes, mas a eventos incidentes mensais (acidentes de trabalho). Cada período temporal representa um conjunto novo de ocorrências, não havendo mecanismo de “sobrevivência” de eventos de um período para o outro. Em outras palavras, acidentes ocorridos em um mês não permanecem ativos no mês seguinte; o processo é composto exclusivamente por novos eventos incidentes.

Dessa forma, o componente de sobrevivência foi removido da estrutura original, adotando-se uma parametrização multiplicativa direta para a dinâmica temporal do processo latente:

$$N_{i,1} \sim \text{Poisson}(\lambda), \quad (2.6)$$

$$N_{i,t} \mid N_{i,t-1} \sim \text{Poisson}(\delta N_{i,t-1}), \quad t \geq 2, \quad (2.7)$$

em que $N_{i,t}$ representa o número real de eventos na unidade i no período t , $\lambda > 0$ corresponde à intensidade inicial do processo latente e $\delta > 0$ representa um fator agregado de variação temporal na intensidade do processo.

Diferentemente da interpretação ecológica, o parâmetro δ não representa sobrevivência biológica, mas sim a variação relativa na incidência esperada entre períodos consecutivos. Assim,

- $\delta > 1$: tendência de aumento no número real de eventos ao longo do tempo;
- $\delta < 1$: tendência de redução;
- $\delta = 1$: estabilidade média da incidência latente.

No nível de observação, mantém-se a estrutura binomial,

$$y_{i,t} \mid N_{i,t} \sim \text{Binomial}(N_{i,t}, p), \quad (2.8)$$

em que $y_{i,t}$ representa a contagem observada e $p \in (0, 1)$ corresponde à probabilidade de detecção (ou notificação) de cada evento.

Assume-se $p_{i,t} = p$ para todos os estratos e períodos. Essa hipótese decorre da ausência de covariáveis observáveis para o processo de detecção e da necessidade de garantir identificabilidade entre abundância latente e probabilidade de notificação em dados agregados.

Essa parametrização preserva a separação entre processo latente e processo observacional, característica essencial dos modelos *N-mixture*, ao mesmo tempo em que reflete adequadamente a natureza de eventos incidentes mensais, nos quais não há componente de permanência ou sobrevivência entre períodos.

2.5 IMPLEMENTAÇÃO VIA JAGS

A estimação dos parâmetros foi conduzida sob uma abordagem Bayesiana utilizando métodos de Monte Carlo via cadeias de Markov (*Markov Chain Monte Carlo* – MCMC). Diversas plataformas computacionais têm sido desenvolvidas para implementação de modelos hierárquicos complexos, incluindo BUGS/OpenBUGS, JAGS, Stan, NIMBLE e abordagens baseadas

em aproximações determinísticas como o INLA (*Integrated Nested Laplace Approximation*). Essas ferramentas diferem principalmente quanto aos algoritmos de inferência utilizados, à flexibilidade na especificação dos modelos e à eficiência computacional em diferentes contextos (LUNN et al., 2012).

Os ambientes da família BUGS foram pioneiros na popularização da modelagem Bayesiana hierárquica via MCMC, permitindo especificação declarativa de modelos probabilísticos e amostragem automática por meio de algoritmos de Gibbs e Metropolis–Hastings (LUNN et al., 2009). O software JAGS (*Just Another Gibbs Sampler*) foi posteriormente desenvolvido como uma alternativa modular e extensível ao BUGS, mantendo compatibilidade conceitual com sua linguagem de modelagem e oferecendo integração direta com o ambiente R (PLUMMER, 2003).

Outra plataforma amplamente utilizada é o Stan, que emprega algoritmos baseados em *Hamiltonian Monte Carlo* (HMC) e *No-U-Turn Sampler* (NUTS), frequentemente mais eficientes em modelos contínuos de alta dimensão (CARPENTER et al., 2017). Entretanto, essa abordagem apresenta limitações quando o modelo inclui variáveis latentes discretas, como ocorre nos modelos *N-mixture*, nos quais o processo de abundância latente $N_{i,t}$ é representado por variáveis de contagem. Nesses casos, o Stan exige a marginalização analítica das variáveis discretas ou reformulações mais complexas do modelo.

Outra alternativa recente é o NIMBLE, que permite maior flexibilidade na construção de algoritmos MCMC personalizados e na compilação eficiente de modelos hierárquicos no ambiente R (VALPINE et al., 2017). Embora essa ferramenta ofereça grande potencial para aplicações avançadas, sua adoção ainda é relativamente recente em comparação com plataformas mais consolidadas.

Por sua vez, o método INLA fornece uma estratégia de inferência Bayesiana baseada em aproximações determinísticas para modelos hierárquicos latentes gaussianos (RUE; MARTINO; CHOPIN, 2009). Essa abordagem apresenta elevada eficiência computacional em diversas aplicações espaciais e temporais. Contudo, sua aplicação direta a modelos com estados latentes discretos e estrutura dinâmica, como o modelo *N-mixture* considerado neste estudo, não é imediata e frequentemente requer aproximações adicionais.

Diante dessas considerações, optou-se pela implementação do modelo no software JAGS, que permite especificação direta de modelos hierárquicos contendo variáveis latentes discretas e processos de contagem. A flexibilidade da linguagem de modelagem e a ampla utilização dessa ferramenta na literatura de modelos ecológicos e epidemiológicos baseados em *N-mixture* tornam o JAGS particularmente adequado para a estrutura considerada neste trabalho.

A implementação foi realizada no ambiente R utilizando os pacotes `rjags` e `coda`, que permitem executar o algoritmo MCMC, monitorar a convergência das cadeias e extrair estatís-

tivas da distribuição *a posteriori*. O código completo utilizado na implementação encontra-se no Apêndice B.

ESPECIFICAÇÃO DAS DISTRIBUIÇÕES *A PRIORI*

Foram adotadas distribuições *a priori* pouco informativas, de modo a permitir que a inferência fosse predominantemente orientada pelos dados observados.

O parâmetro inicial de intensidade do processo latente, λ , foi modelado na escala logarítmica. Para isso, introduziu-se o parâmetro auxiliar β_0 , definido por

$$\beta_0 \sim \mathcal{N}(0, 0.01), \quad (2.9)$$

onde o segundo parâmetro representa a precisão da distribuição normal. O parâmetro de interesse é então obtido por

$$\lambda = \exp(\beta_0), \quad (2.10)$$

garantindo positividade de λ e permitindo ampla dispersão na escala original.

Para o parâmetro dinâmico δ , que controla a variação temporal do processo latente, foi adotada especificação análoga. Define-se

$$\eta_{\text{gain}} \sim \mathcal{N}(0, 0.01), \quad (2.11)$$

$$\delta = \exp(\eta_{\text{gain}}), \quad (2.12)$$

assegurando valores positivos para δ e permitindo variação substancial na dinâmica temporal do processo.

A probabilidade de detecção p , associada ao processo observacional binomial, recebeu distribuição *a priori*

$$p \sim \text{Beta}(1, 1), \quad (2.13)$$

equivalente a uma distribuição uniforme no intervalo $(0, 1)$, representando ausência de informação prévia preferencial sobre o nível de notificação.

Essa estrutura de priors mantém coerência com a literatura de modelos *N-mixture* dinâmicos e evita impor restrições artificiais à estimação dos parâmetros estruturais.

PROCEDIMENTOS DE AMOSTRAGEM MCMC

A estimação Bayesiana foi conduzida por meio de algoritmos de Monte Carlo via Cadeias de Markov (MCMC), implementados no software JAGS e executados no ambiente R. Para todas as análises foram utilizadas três cadeias independentes, iniciadas com valores iniciais distintos para os parâmetros.

Nos estudos de simulação, cada ajuste do modelo foi realizado com fase de adaptação de 5.000 iterações, seguida de um período de aquecimento (*burn-in*) de 10.000 iterações. Após essa etapa foram geradas 100.000 iterações adicionais por cadeia, utilizando fator de *thinning* igual a 50.

Na aplicação aos dados reais adotou-se uma estratégia de amostragem mais extensa, com 10.000 iterações de adaptação e 20.000 iterações de *burn-in*, seguidas de 500.000 iterações por cadeia com *thinning* igual a 200, visando maior estabilidade das estimativas em função do tamanho da base de dados e da estrutura hierárquica do modelo.

Os procedimentos principais do algoritmo podem ser resumidos como:

- utilização de três cadeias MCMC independentes;
- fase de adaptação e período de queima (*burn-in*) para reduzir dependência das condições iniciais;
- amostragem *a posteriori* com *thinning* para atenuar autocorrelação e aumentar o tamanho efetivo da amostra;
- monitoramento dos parâmetros estruturais λ, δ, p e de quantidades derivadas epidemiológicas.

Entre as quantidades derivadas monitoradas destacam-se os totais de eventos latentes por estrato, por período e no conjunto da população:

$$N_i = \sum_{t=1}^T N_{i,t}, \quad N_t = \sum_{i=1}^I N_{i,t}, \quad N_{\text{total}} = \sum_{i=1}^I \sum_{t=1}^T N_{i,t}, \quad (2.14)$$

bem como o número estimado de eventos não notificados,

$$U_{i,t} = N_{i,t} - y_{i,t}, \quad U_i = \sum_{t=1}^T U_{i,t}, \quad U_t = \sum_{i=1}^I U_{i,t}, \quad U_{\text{total}} = \sum_{i=1}^I \sum_{t=1}^T U_{i,t}. \quad (2.15)$$

Essas quantidades permitem quantificar a magnitude da subnotificação ao longo do tempo e entre grupos ocupacionais.

CRITÉRIOS DE CONVERGÊNCIA E EFICIÊNCIA AMOSTRAL

A validade das inferências obtidas por meio de métodos de Cadeias de Markov Monte Carlo (MCMC) depende da convergência das cadeias para a distribuição posterior alvo e da eficiência da amostragem gerada pelo algoritmo. Para avaliar esses aspectos foram considerados dois diagnósticos amplamente utilizados na literatura: o fator de redução potencial de Gelman–Rubin ($\hat{R}$) e o tamanho efetivo da amostra (*Effective Sample Size* – ESS).

O diagnóstico proposto por Gelman e Rubin (1992) baseia-se na comparação entre a variabilidade entre cadeias independentes e a variabilidade dentro de cada cadeia. Seja θ um parâmetro de interesse, e considere m cadeias independentes de comprimento n . Denotando por $\bar{\theta}_{\cdot j}$ a média da cadeia j e por $\bar{\theta}_{\cdot\cdot}$ a média global entre todas as cadeias, a variância entre cadeias é dada por

$$B = \frac{n}{m-1} \sum_{j=1}^m (\bar{\theta}_{\cdot j} - \bar{\theta}_{\cdot\cdot})^2, \quad (2.16)$$

enquanto a variância média dentro das cadeias é

$$W = \frac{1}{m} \sum_{j=1}^m s_j^2, \quad (2.17)$$

em que s_j^2 representa a variância amostral da cadeia j . A estimativa da variância marginal da distribuição posterior pode ser escrita como

$$\widehat{\text{Var}}^+(\theta) = \frac{n-1}{n} W + \frac{1}{n} B. \quad (2.18)$$

O fator de redução potencial é então definido por

$$\hat{R} = \sqrt{\frac{\widehat{\text{Var}}^+(\theta)}{W}}. \quad (2.19)$$

Valores de $\hat{R}$ próximos de 1 indicam que as cadeias apresentam comportamento semelhante, sugerindo convergência para a distribuição posterior estacionária.

Além da convergência, é importante avaliar a eficiência da amostragem gerada pelas cadeias. Para esse fim utiliza-se o tamanho efetivo da amostra (ESS), que quantifica o número equivalente de amostras independentes contidas em uma sequência de amostras autocorrelacionadas. Seja N o número total de amostras geradas e ρ_k a autocorrelação na defasagem k . O ESS pode ser aproximado por

$$ESS = \frac{N}{1 + 2 \sum_{k=1}^{\infty} \rho_k}. \quad (2.20)$$

Em termos relativos, define-se a eficiência amostral como

$$ESS_{rel} = \frac{ESS}{N}. \quad (2.21)$$

Esse indicador representa a proporção de amostras efetivamente independentes contidas nas cadeias geradas. Valores elevados de ESS_{rel} indicam menor autocorrelação entre as amostras e, conseqüentemente, maior eficiência do algoritmo MCMC.

Neste estudo, a qualidade da amostragem foi avaliada com base na proximidade de $\hat{R}$ em relação à unidade e na proporção relativa do tamanho efetivo da amostra em relação ao número total de iterações geradas, assegurando precisão adequada nas estimativas *a posteriori*.

INTERVALOS DE CREDIBILIDADE

As estimativas pontuais dos parâmetros foram obtidas pelas médias *a posteriori* das cadeias MCMC. A incerteza associada às estimativas foi quantificada por meio de intervalos de credibilidade de 95%, definidos pelos quantis 2,5% e 97,5% da distribuição *a posteriori* de cada parâmetro.

Diferentemente dos intervalos de confiança frequentistas, os intervalos de credibilidade possuem interpretação probabilística direta no contexto Bayesiano. Especificamente, dado o modelo especificado, as distribuições *a priori* e os dados observados, existe probabilidade de 95% de que o valor verdadeiro do parâmetro pertença ao intervalo estimado.

Esses intervalos foram calculados tanto para os parâmetros estruturais do modelo (λ , δ e p) quanto para quantidades derivadas de interesse epidemiológico, como o número total de eventos latentes N_{total} e o número estimado de casos não notificados U_{total} . Dessa forma, os intervalos de credibilidade permitem avaliar não apenas a incerteza associada aos parâmetros do processo latente e do mecanismo de detecção, mas também à magnitude da subnotificação estimada ao longo do período de estudo.

A utilização desses intervalos é particularmente relevante em contextos de detecção imperfeita, como sistemas de vigilância epidemiológica baseados em notificação passiva, nos quais a variabilidade das estimativas reflete simultaneamente a incerteza sobre o processo de ocorrência dos eventos e sobre o processo de notificação.

AVALIAÇÃO DO MODELO

A avaliação do modelo dinâmico *N-mixture* considerou três dimensões principais:

1. diagnóstico de convergência das cadeias MCMC;

2. verificação da coerência estrutural entre os estados latentes estimados e as contagens observadas;
3. análise comparativa das distribuições *a posteriori* entre grupos ocupacionais e períodos epidemiológicos.

ANÁLISE COMPARATIVA ENTRE CENÁRIOS

A comparação entre grupos ocupacionais e períodos epidemiológicos foi realizada por meio das distribuições *a posteriori* dos parâmetros estruturais λ , δ e p , bem como das quantidades derivadas $N_{i,t}$ e $U_{i,t}$.

As diferenças entre cenários foram interpretadas considerando a sobreposição dos intervalos de credibilidade e a magnitude das estimativas centrais. Essa abordagem privilegia a interpretação direta das distribuições *a posteriori*, evitando depender exclusivamente de critérios de informação baseados em penalização de complexidade.

Particular atenção foi dedicada ao período associado à pandemia de COVID-19, com o objetivo de investigar possíveis alterações estruturais tanto na ocorrência real dos eventos quanto no processo de notificação.

2.6 ESTUDO DE SIMULAÇÃO

Com o objetivo de avaliar o desempenho inferencial do modelo sob diferentes níveis de detecção imperfeita, foi conduzido um estudo de simulação Monte Carlo. O estudo buscou verificar a capacidade do modelo em recuperar os valores verdadeiros dos parâmetros λ , δ e p , bem como avaliar a precisão e a calibração dos intervalos de credibilidade obtidos a partir das distribuições *a posteriori*.

PROCEDIMENTO DE GERAÇÃO DOS DADOS

Para cada cenário considerado foram geradas séries sintéticas contendo I estratos e T períodos temporais, seguindo a mesma estrutura probabilística adotada no modelo aplicado aos dados reais.

O processo latente foi gerado segundo

$$N_{i,1} \sim \text{Poisson}(\lambda), \quad (2.22)$$

$$N_{i,t} \mid N_{i,t-1} \sim \text{Poisson}(\delta N_{i,t-1}), \quad t = 2, \dots, T, \quad (2.23)$$

enquanto o processo observacional foi definido por

$$y_{i,t} \mid N_{i,t} \sim \text{Binomial}(N_{i,t}, p). \quad (2.24)$$

Assumiu-se $p_{i,t} = p$, em consonância com a especificação adotada na aplicação empírica do modelo.

Para cada combinação de parâmetros (λ, δ, p) foram geradas B réplicas independentes. Em cada réplica, o modelo foi ajustado utilizando o mesmo procedimento MCMC, as mesmas distribuições *a priori* e os mesmos critérios de convergência empregados na análise dos dados reais.

MÉTRICAS DE DESEMPENHO

Para um parâmetro genérico $\theta \in \{\lambda, \delta, p\}$, seja $\hat{\theta}_b$ a média *a posteriori* obtida na réplica b do estudo de simulação, com valor verdadeiro θ . As seguintes métricas foram consideradas:

Viés relativo

$$\text{RelBias}(\theta) = \frac{1}{B} \sum_{b=1}^B \frac{\hat{\theta}_b - \theta}{\theta}, \quad (2.25)$$

medindo o desvio relativo médio entre as estimativas e o valor verdadeiro do parâmetro.

Viés

$$\text{Bias}(\theta) = \frac{1}{B} \sum_{b=1}^B (\hat{\theta}_b - \theta), \quad (2.26)$$

avaliando a presença de tendência sistemática do estimador.

Erro quadrático médio (RMSE)

$$\text{RMSE}(\theta) = \sqrt{\frac{1}{B} \sum_{b=1}^B (\hat{\theta}_b - \theta)^2}, \quad (2.27)$$

quantificando simultaneamente o viés e a variabilidade das estimativas.

Cobertura dos intervalos de credibilidade

A cobertura empírica dos intervalos de credibilidade de 95% foi calculada por

$$\text{Cobertura}(\theta) = \frac{1}{B} \sum_{b=1}^B \mathbb{I}\{\theta \in IC_{95\%}^{(b)}(\theta)\}, \quad (2.28)$$

em que $\mathbb{I}\{\cdot\}$ representa a função indicadora e $IC_{95\%}^{(b)}(\theta)$ denota o intervalo de credibilidade de 95% obtido na réplica b . Valores de cobertura próximos de 0,95 indicam calibração adequada dos intervalos de credibilidade.

Avaliação de quantidades derivadas

Além dos parâmetros estruturais λ , δ e p , examinou-se também a recuperação de quantidades epidemiologicamente relevantes, como o número total de eventos latentes

$$N_{\text{total}} = \sum_{i=1}^I \sum_{t=1}^T N_{i,t}, \quad (2.29)$$

e o número total estimado de eventos não notificados

$$U_{\text{total}} = \sum_{i=1}^I \sum_{t=1}^T (N_{i,t} - y_{i,t}), \quad (2.30)$$

avaliando se o modelo foi capaz de recuperar adequadamente a magnitude total de eventos e da subnotificação.

CRITÉRIOS DE AVALIAÇÃO

O modelo foi considerado bem calibrado quando apresentou:

- viés próximo de zero;
- RMSE relativamente pequeno em relação à magnitude do parâmetro;
- cobertura empírica próxima de 95%.

A avaliação conjunta dessas métricas permitiu examinar a estabilidade inferencial, a precisão e a robustez do modelo sob diferentes cenários de subnotificação e dinâmica temporal, fornecendo suporte metodológico para sua aplicação aos dados reais com detecção imperfeita.

SIMULAÇÃO CALIBRADA A PARTIR DAS ESTIMATIVAS DOS DADOS REAIS

Com o objetivo de avaliar o desempenho do modelo em cenários realistas, foi conduzido um estudo de simulação Monte Carlo calibrado a partir das estimativas obtidas na aplicação do modelo aos dados reais do SINAN.

Nesse procedimento, os valores utilizados para os parâmetros estruturais do modelo, intensidade inicial do processo latente (λ), fator de dinâmica temporal (δ) e probabilidade de

detecção (p), foram definidos com base nas estimativas *a posteriori* obtidas para cada grupo ocupacional e período epidemiológico analisado.

A geração dos dados sintéticos seguiu a mesma estrutura hierárquica apresentada anteriormente para o modelo dinâmico *N-mixture* (Seção 2.4), preservando a separação entre o processo latente de ocorrência dos eventos e o processo observacional associado à notificação.

Para cada cenário considerado foram geradas séries sintéticas contendo $n = 20$ estratos e T períodos temporais, sendo T definido de acordo com a duração do período epidemiológico correspondente. Em cada cenário foram geradas $B = 100$ réplicas independentes.

Em cada réplica, os dados simulados foram utilizados para ajustar novamente o modelo utilizando inferência Bayesiana via MCMC implementada no software JAGS, mantendo-se a mesma especificação de distribuições *a priori*, número de cadeias e critérios de convergência empregados na análise dos dados reais.

Além da recuperação dos parâmetros λ , δ e p , foram avaliadas quantidades derivadas epidemiologicamente relevantes, como o número total de eventos latentes e a magnitude total da subnotificação estimada pelo modelo. A comparação entre os valores verdadeiros utilizados na geração dos dados e as estimativas obtidas permitiu avaliar a robustez do método em cenários compatíveis com os padrões observados no sistema real de notificação.

RESULTADOS

3.1 ESTUDO DE SIMULAÇÃO

Com o objetivo de avaliar o desempenho inferencial do modelo *N-mixture* dinâmico sob diferentes configurações paramétricas, foi conduzido um estudo de simulação Monte Carlo com múltiplas replicações independentes. A estrutura probabilística utilizada para geração e ajuste dos dados segue o modelo hierárquico descrito no Capítulo 2.

CONFIGURAÇÃO DOS CENÁRIOS

Foram avaliadas diferentes combinações dos parâmetros de intensidade inicial (λ), dinâmica temporal (δ) e probabilidade de detecção (p). Consideraram-se dois níveis para λ (100 e 500), três valores para δ (0,95, 1,00 e 1,05) e três níveis para p (0,50, 0,75 e 1,00), totalizando 18 cenários distintos.

Em todos os cenários, fixou-se o número de estratos em $n = 20$ e o número de períodos em $T = 24$. Para cada combinação paramétrica foram geradas 100 replicações independentes. O ajuste do modelo em cada réplica foi realizado utilizando o mesmo procedimento de inferência Bayesiana descrito na Seção 2.4.

Além dos parâmetros estruturais, foram avaliadas quantidades derivadas associadas ao processo latente, incluindo o total acumulado de eventos e o número de eventos não observados, definidos conforme as Equações (2.14) e (2.15).

DESEMPENHO INFERENCIAL DOS PARÂMETROS

A Tabela 1 apresenta o resumo das métricas de desempenho para os parâmetros λ , δ e p nos diferentes cenários avaliados.

Tabela 1 – Desempenho inferencial dos parâmetros do modelo no estudo de simulação.

λ	δ	p	λ				δ				p			
			Bias	RBias	RMSE	Cov	Bias	RBias	RMSE	Cov	Bias	RBias	RMSE	Cov
100	0.95	0.50	2.97	0.0297	13.2	0.96	0.000224	0.000236	0.00617	0.94	-0.00132	-0.00264	0.0554	0.97
100	0.95	0.75	1.05	0.0105	8.05	0.94	-0.0000588	-0.0000619	0.00547	0.98	0.000708	0.000945	0.0572	0.93
100	0.95	1.00	4.53	0.0453	5.58	0.78	0.000427	0.000449	0.00590	0.94	-0.0397	-0.0397	0.0441	0.00
100	1.00	0.50	1.83	0.0183	12.1	0.94	0.0000276	0.0000276	0.00443	0.96	0.00292	0.00584	0.0539	0.95
100	1.00	0.75	1.83	0.0183	7.31	1.00	-0.0000927	-0.0000927	0.00479	0.94	-0.00624	-0.00832	0.0487	0.98
100	1.00	1.00	4.32	0.0432	5.09	0.86	-0.000234	-0.000234	0.00486	0.92	-0.0398	-0.0398	0.0429	0.00
100	1.05	0.50	1.54	0.0154	13.6	0.92	0.000636	0.000605	0.00389	0.92	0.00501	0.0100	0.0632	0.92
100	1.05	0.75	0.454	0.00454	8.16	0.96	-0.000945	-0.000900	0.00353	0.95	0.00618	0.00824	0.0576	0.94
100	1.05	1.00	3.77	0.0377	5.05	0.81	0.000127	0.000121	0.00384	0.96	-0.0385	-0.0385	0.0436	0.00
500	0.95	0.50	13.8	0.0277	82.0	0.95	0.000333	0.000351	0.00263	0.96	0.00194	0.00387	0.0618	0.95
500	0.95	0.75	9.86	0.0197	39.7	0.96	0.000591	0.000622	0.00252	0.96	-0.00466	-0.00621	0.0543	0.97
500	0.95	1.00	22.2	0.0445	25.5	0.73	0.0000728	0.0000767	0.00241	0.97	-0.0411	-0.0411	0.0457	0.00
500	1.00	0.50	6.25	0.0125	63.4	0.91	0.0000486	0.0000486	0.00204	0.99	0.00701	0.0140	0.0609	0.91
500	1.00	0.75	2.47	0.00493	38.1	0.92	0.0000448	0.0000448	0.00213	0.93	0.00469	0.00625	0.0538	0.94
500	1.00	1.00	22.4	0.0447	25.6	0.69	-0.0000625	-0.0000625	0.00182	0.97	-0.0409	-0.0409	0.0449	0.00
500	1.05	0.50	5.50	0.0110	54.0	0.95	-0.000138	-0.000131	0.00185	0.90	0.00657	0.0131	0.0516	0.95
500	1.05	0.75	8.16	0.0163	36.0	0.94	-0.0000513	-0.0000488	0.00147	0.98	-0.00450	-0.00600	0.0507	0.94
500	1.05	1.00	23.6	0.0473	26.4	0.59	-0.0000517	-0.0000493	0.00172	0.93	-0.0440	-0.0440	0.0477	0.00

Desempenho para λ .

O viés associado à estimativa de λ manteve-se relativamente reduzido na maioria dos cenários com detecção imperfeita moderada ($p = 0,5$ e $p = 0,75$). Observa-se que o erro quadrático médio (RMSE) aumenta quando $\lambda = 500$, o que é esperado, uma vez que o erro absoluto tende a crescer com a magnitude do parâmetro.

A cobertura empírica aproximou-se do nível nominal de 95% na maior parte dos cenários com detecção imperfeita. Entretanto, quando $p = 1$, verificou-se redução sistemática da cobertura, particularmente nos cenários com maior intensidade inicial. Esse resultado sugere impacto da ausência de variabilidade observacional na separação entre o processo latente e o processo de detecção.

Desempenho para δ .

O parâmetro de dinâmica temporal apresentou desempenho consistente em todos os cenários avaliados. O viés foi praticamente nulo, o RMSE manteve-se da ordem de 10^{-3} e a cobertura empírica permaneceu próxima de 95%. Esses resultados indicam que o modelo recupera de forma robusta a dinâmica temporal do processo latente, independentemente do regime de crescimento considerado.

Desempenho para p .

Nos cenários com detecção imperfeita ($p = 0,5$ e $p = 0,75$), o modelo apresentou viés reduzido, RMSE estável e cobertura próxima ao nível nominal, indicando adequada capacidade de estimar a probabilidade de detecção sob subnotificação.

Por outro lado, quando $p = 1$, a cobertura empírica foi igual a zero em todos os cenários. Esse comportamento decorre da perda de identificabilidade do componente observacional quando não há subnotificação, uma vez que o processo observado coincide com o processo latente.

RECUPERAÇÃO DE QUANTIDADES DERIVADAS

Além da avaliação dos parâmetros estruturais, investigou-se também o desempenho do modelo na recuperação de quantidades derivadas associadas ao processo latente. A Tabela 2 apresenta os resultados para o total de eventos N_{total} e para o total de eventos não observados U_{total} .

Tabela 2 – Resultados do estudo de simulação para as quantidades agregadas N_{total} e U_{total} .

λ	δ	p	Médias simuladas			N_{total}				U_{total}			
			$\bar{Y}$	$\bar{N}$	$\bar{U}$	Bias	RBias	RMSE	Cov	Bias	RBias	RMSE	Cov
100	0.95	0.50	14216	28454	14238	873	0.030	3700	0.97	873	0.061	3700	0.97
100	0.95	0.75	21259	28336	7077	314	0.011	2292	0.93	314	0.045	2292	0.93
100	0.95	1.00	28613	28613	0	1221	0.043	1379	0.00	1221	–	1379	0.00
100	1.00	0.50	23835	47682	23847	1000	0.021	5680	0.95	1000	0.042	5680	0.95
100	1.00	0.75	36110	48133	12022	891	0.019	3342	0.98	891	0.075	3342	0.98
100	1.00	1.00	47989	47989	0	2050	0.043	2236	0.00	2050	–	2236	0.00
100	1.05	0.50	44929	89851	44922	1881	0.021	12193	0.92	1881	0.041	12193	0.92
100	1.05	0.75	66297	88379	22083	294	0.004	7125	0.94	294	0.015	7125	0.94
100	1.05	1.00	89880	89880	0	3744	0.041	4318	0.00	3744	–	4318	0.00
500	0.95	0.50	70862	141755	70894	4219	0.030	23460	0.95	4219	0.060	23460	0.95
500	0.95	0.75	107311	143076	35765	2543	0.018	11237	0.97	2543	0.071	11237	0.97
500	0.95	1.00	141793	141793	0	6320	0.044	7126	0.00	6320	–	7126	0.00
500	1.00	0.50	119885	239862	119977	3381	0.014	30413	0.92	3381	0.028	30413	0.92
500	1.00	0.75	180288	240420	60132	1039	0.004	18012	0.94	1039	0.017	18012	0.94
500	1.00	1.00	240315	240315	0	10617	0.044	11819	0.00	10617	–	11819	0.00
500	1.05	0.50	222737	445363	222626	5149	0.012	48252	0.95	5149	0.023	48252	0.95
500	1.05	0.75	333850	445144	111294	7263	0.016	32280	0.94	7263	0.065	32280	0.94
500	1.05	1.00	445280	445280	0	21227	0.048	23240	0.00	21227	–	23240	0.00

Desempenho para N_{total} .

De modo geral, o modelo apresentou boa capacidade de recuperação do número total de eventos N_{total} . O viés relativo permaneceu reduzido na maioria dos cenários e a cobertura empírica aproximou-se do nível nominal de 95% nos cenários com detecção imperfeita.

Observa-se que o RMSE aumenta conforme cresce a intensidade inicial λ , refletindo o aumento natural da variabilidade absoluta em cenários com maior magnitude do processo latente.

Desempenho para U_{total} .

Resultados semelhantes foram observados para o total de eventos não observados U_{total} . Nos cenários com detecção imperfeita ($p = 0,5$ e $p = 0,75$), o modelo apresentou viés relativamente pequeno e cobertura empírica próxima do nível nominal.

Nos cenários com $p = 1$, o número de eventos não observados é nulo por construção ($U_{total} = 0$), de modo que o viés relativo não é definido nesses casos e a cobertura empírica também assume valor nulo.

De forma geral, os resultados indicam que o modelo *N-mixture* dinâmico apresenta bom desempenho na recuperação tanto dos parâmetros estruturais quanto das quantidades derivadas associadas ao processo latente, especialmente nos cenários em que há detecção imperfeita.

3.2 APLICAÇÃO AOS DADOS DO SINAN

ANÁLISE DESCRITIVA

A base analítica compreende 410.879 notificações de acidentes de trabalho com exposição a material biológico registradas no período de 2014 a 2024. A distribuição sociodemográfica e ocupacional dos registros encontra-se apresentada na Tabela 3.

Tabela 3 – Distribuição das notificações de acidentes com material biológico segundo características sociodemográficas e ocupacionais.

Variável	Categoria	n (%)
Categoria Ocupacional	Técnicos e Auxiliares da Saúde	275.243 (67,0)
	Profissionais de Saúde (Superior)	135.636 (33,0)
Sexo	Feminino	338.351 (82,3)
	Masculino	72.528 (17,7)
Raça/Cor	Branca	244.346 (59,5)
	Preta/Parda	166.533 (40,5)
Faixa Etária (Feminino)	18–23	33.882 (10,0)
	24–28	70.166 (20,7)
	29–34	77.251 (22,8)
	35–39	56.165 (16,6)
	40–49	69.633 (20,6)
	50–70	31.254 (9,2)
Faixa Etária (Masculino)	18–29	28.959 (39,9)
	30–39	25.582 (35,3)
	40–49	11.704 (16,1)
	50–70	6.283 (8,7)

Fonte: Sistema de Informação de Agravos de Notificação (SINAN), Brasil, 2014–2024. Percentuais calculados dentro de cada variável.

PERFIL DOS REGISTROS

Observa-se predominância de notificações entre Técnicos e Auxiliares da Saúde, responsáveis por 67,0% dos registros. A distribuição por sexo revela forte concentração no sexo feminino (82,3%), padrão consistente com a composição da força de trabalho nos serviços de saúde. Em relação à raça/cor, a maior proporção de notificações refere-se a indivíduos autodeclarados brancos (59,5%).

Quanto à faixa etária, verifica-se maior concentração de casos entre mulheres de 24 a 34 anos, enquanto entre os homens predominam as faixas de 18 a 39 anos. Esses resultados indicam maior incidência do agravo em grupos etariamente ativos no exercício profissional.

DINÂMICA TEMPORAL AGREGADA

A Figura 1 apresenta a série temporal mensal agregada por categoria ocupacional. Observa-se tendência geral de crescimento no número de notificações ao longo do período analisado, acompanhada de oscilações sazonais e variações abruptas em determinados intervalos temporais. A magnitude das notificações entre Técnicos e Auxiliares da Saúde permanece sistematicamente superior à observada entre Profissionais de Saúde (nível superior).

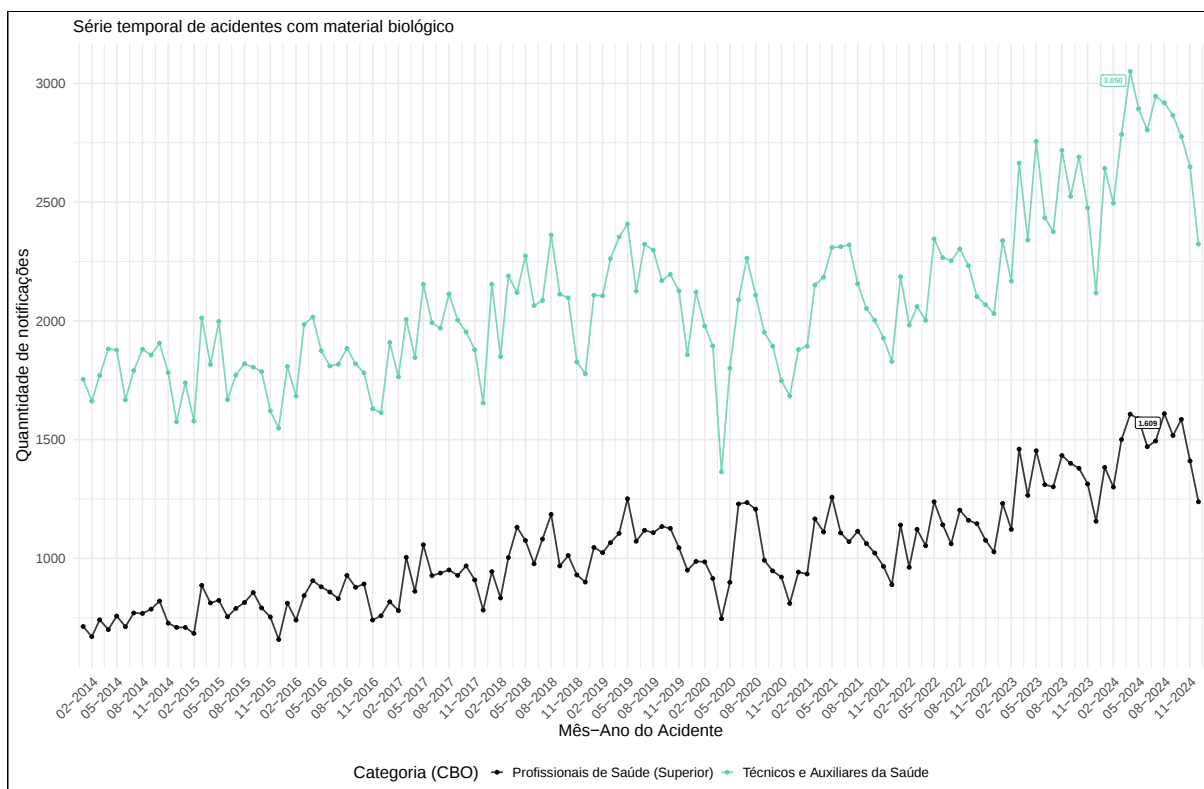


Figura 1 – Tendência mensal de notificações por categoria ocupacional (CBO).

Destaca-se uma inflexão estrutural no período entre 2020 e 2022, possivelmente associada às mudanças organizacionais e assistenciais decorrentes da pandemia de COVID-19.

Esse comportamento sugere que parte das variações observadas pode refletir tanto alterações na ocorrência real dos acidentes quanto mudanças no processo de notificação.

HETEROGENEIDADE ENTRE ESTRATOS

As Figuras 2 e 3 apresentam a evolução temporal estratificada por sexo, raça/cor e faixa etária para cada grupo ocupacional. Observa-se heterogeneidade significativa entre os estratos, tanto em termos de nível médio quanto de variabilidade temporal.

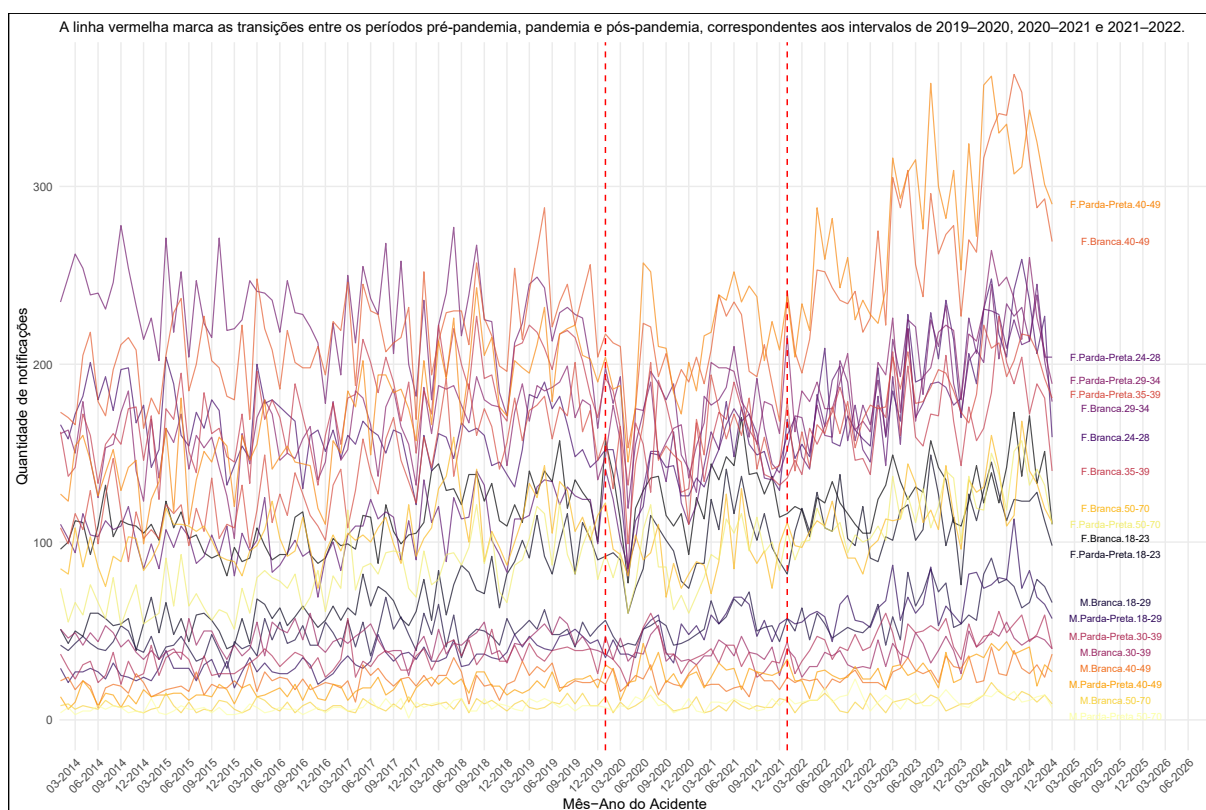


Figura 2 – Evolução temporal por estrato para Técnicos e Auxiliares da Saúde.

Alguns estratos concentram volumes substancialmente superiores de notificações, enquanto outros apresentam séries mais estáveis e de menor magnitude. Além disso, observa-se ocorrência de flutuações simultâneas em múltiplos estratos, sugerindo a presença de componentes sistêmicos que afetam conjuntamente diferentes segmentos ocupacionais.

Esses padrões reforçam a necessidade de uma abordagem de modelagem hierárquica capaz de acomodar variações entre estratos e dependência temporal na abundância latente.

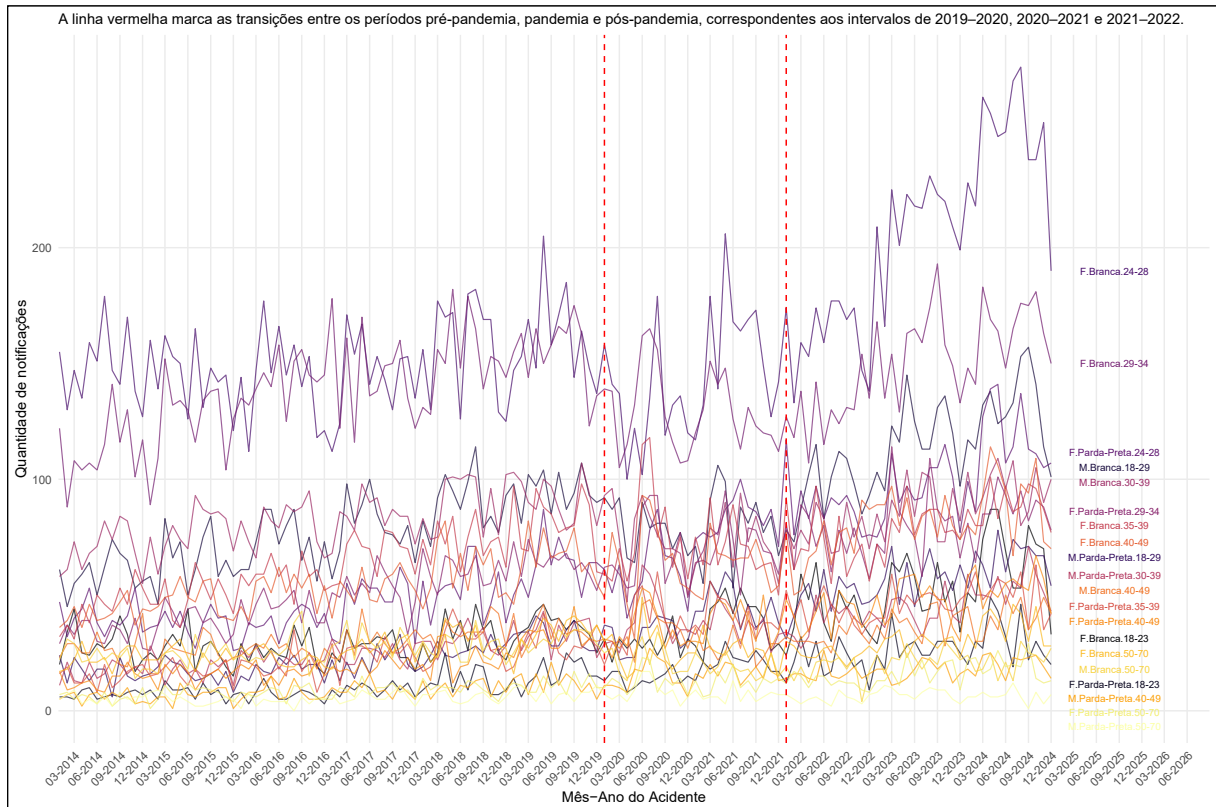


Figura 3 – Evolução temporal por estrato para Profissionais de Saúde (nível superior).

APLICAÇÃO DO MODELO AOS DADOS REAIS DO SINAN

Os modelos *N-mixture* dinâmicos foram ajustados separadamente para os grupos ocupacionais de Profissionais de Saúde (nível superior) e Técnicos e Auxiliares da Saúde. Além disso, a análise foi conduzida considerando três períodos epidemiológicos distintos: até 2019, 2020–2021 e a partir de 2022. Essa segmentação temporal foi adotada para permitir a investigação de possíveis mudanças estruturais associadas ao período da pandemia de COVID-19 e ao contexto subsequente.

O procedimento de estimação foi realizado sob abordagem Bayesiana utilizando amostragem MCMC implementada no JAGS, conforme descrito na Seção 2.5 do Capítulo 2. A partir das amostras *a posteriori*, foram obtidas estimativas pontuais dos parâmetros do modelo e intervalos de credibilidade de 95%.

As estimativas *a posteriori* dos parâmetros estruturais e das quantidades derivadas encontram-se apresentadas na Tabela 4. Os valores correspondem às médias *a posteriori* acompanhadas dos respectivos intervalos de credibilidade de 95%.

Tabela 4 – Estimativas *a posteriori* dos modelos *N-mixture* dinâmicos por grupo ocupacional e período (estratos: sexo, raça e idade).

Grupo	Período	$\hat{\lambda}$	$\hat{\delta}$	$\hat{p}$	$\hat{N}_{\text{total}}$	$\hat{U}_{\text{total}}$
Técnicos e Auxiliares da Saúde						
	2014–2019	158,8 (147,7 ; 170,6)	1,002 (0,998 ; 1,006)	0,520 (0,488 ; 0,553)	266.800 (251.000 ; 283.800)	128.000 (112.200 ; 145.000)
	2020–2021	143,2 (134,2 ; 153,6)	0,996 (0,988 ; 1,004)	0,711 (0,666 ; 0,752)	67.370 (63.700 ; 71.760)	19.460 (15.790 ; 23.850)
	2022–2024	165,7 (154,6 ; 177,8)	1,004 (0,998 ; 1,010)	0,624 (0,585 ; 0,663)	141.800 (133.400 ; 151.200)	53.200 (44.870 ; 62.660)
Profissionais de Saúde (Nível Superior)						
	2014–2019	83,5 (75,1 ; 93,5)	1,006 (1,001 ; 1,011)	0,395 (0,357 ; 0,434)	162.400 (147.900 ; 179.800)	98.110 (83.660 ; 115.600)
	2020–2021	81,9 (74,2 ; 90,7)	0,998 (0,988 ; 1,009)	0,578 (0,524 ; 0,630)	42.410 (38.900 ; 46.680)	17.900 (14.390 ; 22.160)
	2022–2024	106,4 (96,3 ; 117,6)	1,006 (0,999 ; 1,013)	0,497 (0,453 ; 0,543)	94.210 (86.110 ; 103.400)	47.360 (39.250 ; 56.500)

De forma geral, o parâmetro $\hat{\lambda}$, associado à intensidade inicial do processo latente, apresentou valores sistematicamente superiores no grupo de Técnicos e Auxiliares da Saúde em comparação aos Profissionais de Saúde de nível superior, indicando maior magnitude do processo latente nesse grupo ocupacional.

O parâmetro de dinâmica temporal $\hat{\delta}$ manteve-se próximo de 1 em todos os modelos ajustados, sugerindo relativa estabilidade no crescimento do processo latente ao longo do tempo, com variações discretas entre períodos e grupos ocupacionais.

A probabilidade de detecção $\hat{p}$ apresentou variação relevante entre os períodos analisados. Observa-se aumento substancial durante o intervalo 2020–2021 em ambos os grupos ocupacionais, seguido de redução a partir de 2022. Esse comportamento sugere alterações no mecanismo de notificação ao longo do período analisado.

As estimativas das quantidades derivadas $\hat{N}_{\text{total}}$ e $\hat{U}_{\text{total}}$ indicam diferenças marcantes entre períodos. Em ambos os grupos ocupacionais, observa-se redução do total estimado de eventos no intervalo 2020–2021, seguida de aumento a partir de 2022. A magnitude de $\hat{U}_{\text{total}}$ evidencia volume expressivo de eventos não notificados em todos os períodos considerados.

A Figura 4 apresenta a série temporal das notificações observadas (Y_t), das estimativas da ocorrência total (N_t) e da subnotificação estimada (U_t), obtidas a partir do modelo *N-mixture* dinâmico para os dois grupos ocupacionais analisados.

Observa-se que, ao longo de todo o período analisado, as estimativas da ocorrência total (N_t) permanecem sistematicamente superiores às notificações observadas (Y_t). Essa diferença corresponde ao número estimado de eventos não notificados (U_t), indicando presença persistente de subnotificação nas séries temporais analisadas.

Para o grupo de Técnicos e Auxiliares da Saúde, as notificações observadas apresentam tendência geral de crescimento entre 2014 e 2019, variando aproximadamente entre 1.600 e 2.100 registros mensais. No mesmo período, as estimativas da ocorrência total situam-se em níveis mais elevados, oscilando aproximadamente entre 3.200 e 4.300 eventos mensais.

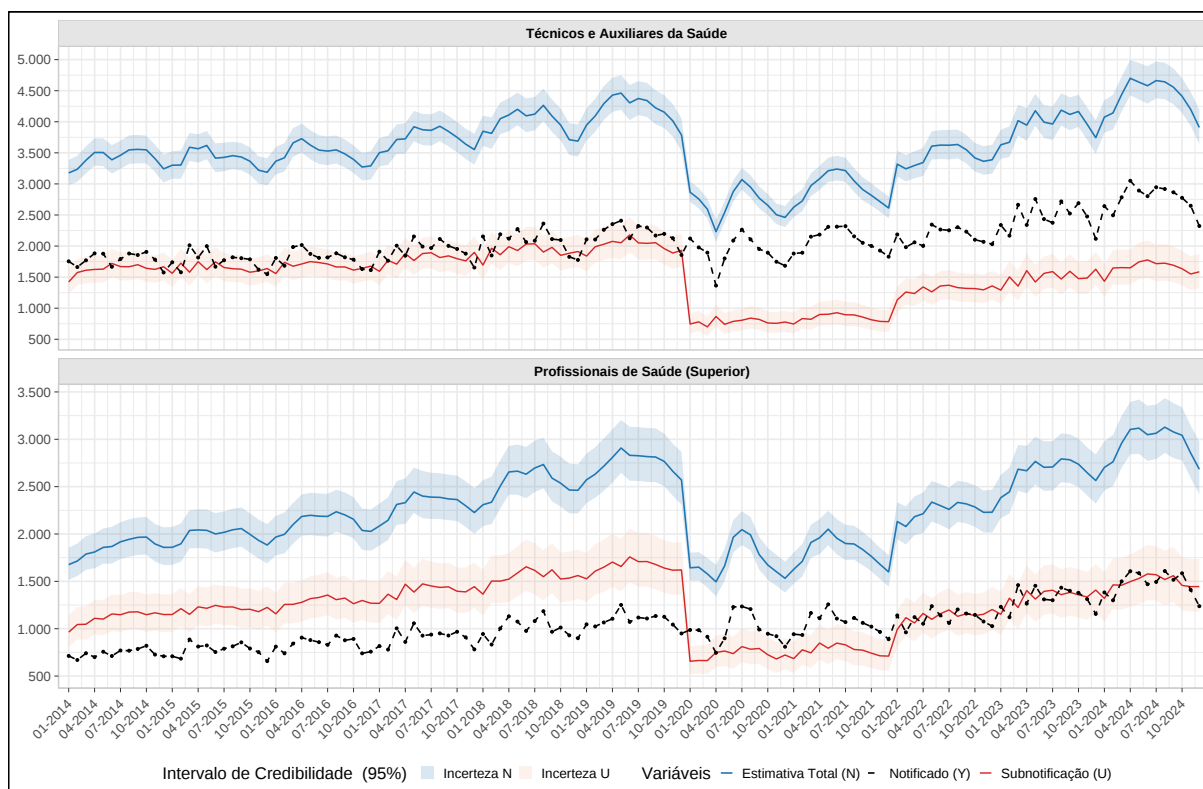


Figura 4 – Série temporal das notificações observadas (Y_t), da ocorrência total estimada (N_t) e da subnotificação estimada (U_t) de acidentes de trabalho com exposição a material biológico, com intervalos de credibilidade de 95%, para dois grupos ocupacionais do setor saúde (2014–2024).

Entre 2020 e 2021 observa-se redução acentuada nas notificações observadas, acompanhada por diminuição nas estimativas da ocorrência total. A partir de 2022, as notificações voltam a apresentar tendência de crescimento gradual, comportamento também refletido nas estimativas do processo latente.

Para o grupo de Profissionais de Saúde (Nível Superior) observa-se padrão temporal semelhante, porém com níveis absolutos inferiores de ocorrência total estimada. Ao longo de toda a série temporal, as estimativas de subnotificação permanecem substanciais, refletindo a diferença consistente entre os eventos estimados e aqueles efetivamente registrados no sistema de vigilância.

DIAGNÓSTICO DE CONVERGÊNCIA E EFICIÊNCIA AMOSTRAL

A convergência das cadeias de Markov e a eficiência da amostragem foram avaliadas por meio do fator de redução potencial de Gelman–Rubin ($\hat{R}$) e do tamanho efetivo da amostra (*Effective Sample Size* – ESS). Esses diagnósticos permitem verificar se as cadeias atingiram o regime estacionário da distribuição *a posteriori* e se o número de amostras efetivamente independentes é suficiente para garantir precisão nas estimativas obtidas via MCMC.

Os resultados do diagnóstico de convergência são apresentados na Tabela 5. Observa-

se que os valores estimados de $\hat{R}$ permaneceram muito próximos de 1 em todos os parâmetros e cenários analisados. As estimativas pontuais variaram aproximadamente entre 1,00 e 1,02, enquanto os limites superiores dos respectivos intervalos de confiança permaneceram inferiores a 1,05.

Esses resultados indicam adequada mistura entre as cadeias independentes e sugerem que o algoritmo MCMC atingiu convergência para a distribuição *a posteriori*. De acordo com recomendações recentes da literatura, valores de $\hat{R}$ muito próximos da unidade indicam boa convergência das cadeias e baixo erro de Monte Carlo nas estimativas inferenciais (VATS; KNUDSON, 2021).

Tabela 5 – Diagnóstico de convergência via Gelman–Rubin ($\hat{R}$) por grupo ocupacional e período. Valores apresentados como estimativa pontual (limite superior do IC).

Grupo	Período	$\hat{R}_{\lambda}$	$\hat{R}_{\delta}$	$\hat{R}_{\hat{p}}$	$\hat{R}_{\hat{N}_{total}}$
Técnicos e Auxiliares da Saúde					
	2019	1.00 (1.00)	1.00 (1.00)	1.00 (1.01)	1.00 (1.01)
	2020–2021	1.00 (1.01)	1.00 (1.00)	1.00 (1.01)	1.00 (1.01)
	2022	1.01 (1.02)	1.00 (1.01)	1.02 (1.01)	1.00 (1.01)
Profissionais de Saúde (Nível Superior)					
	2019	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	1.01 (1.00)
	2020–2021	1.00 (1.01)	1.00 (1.00)	1.00 (1.01)	1.00 (1.01)
	2022	1.00 (1.01)	1.00 (1.00)	1.01 (1.02)	1.01 (1.01)

A eficiência da amostragem foi avaliada por meio do tamanho efetivo da amostra (ESS), cujos resultados encontram-se apresentados na Tabela 6. O ESS quantifica o número equivalente de amostras independentes presentes nas cadeias autocorrelacionadas geradas pelo algoritmo MCMC.

De modo geral, observa-se que os valores de ESS obtidos foram elevados para todos os parâmetros analisados. A maioria das estimativas apresentou ESS superior a 1.500, indicando nível adequado de independência entre as amostras e boa precisão na aproximação da distribuição *a posteriori*. Na literatura recente, valores de ESS superiores a aproximadamente 400 já são considerados suficientes para garantir estimativas confiáveis em inferência Bayesiana via MCMC (VATS; KNUDSON, 2021).

Destaca-se particularmente o parâmetro associado à dinâmica temporal do processo latente (δ), que apresentou valores de ESS próximos ao limite máximo observado após o procedimento de *thinning*, atingindo aproximadamente 7.500 em diversos cenários. Esse valor corresponde aproximadamente ao número total de amostras *a posteriori* disponíveis após a combinação das três cadeias independentes, indicando níveis muito baixos de autocorrelação nas cadeias correspondentes.

Mesmo nos casos em que o ESS apresentou valores relativamente menores, como

observado em alguns parâmetros no período de 2019, os valores permaneceram acima de aproximadamente 1.000, o que ainda assegura boa precisão nas estimativas de Monte Carlo.

Em conjunto, os diagnósticos de $\hat{R}$ e ESS indicam que as cadeias MCMC apresentaram comportamento estacionário e eficiência amostral satisfatória, garantindo robustez às estimativas *a posteriori* obtidas pelos modelos ajustados.

Tabela 6 – Tamanho efetivo da amostra (ESS) por grupo ocupacional e período.

Grupo	Período	ESS($\hat{N}_{\text{total}}$)	ESS($\hat{U}_{\text{total}}$)	ESS($\hat{\delta}$)	ESS($\hat{\lambda}$)	ESS($\hat{p}$)
Técnicos e Auxiliares da Saúde						
	2019	1357.295	1357.295	6744.805	1929.412	1364.051
	2020–2021	3209.393	3209.393	7465.115	3912.127	3226.427
	2022	1978.183	1978.183	7500.000	2513.059	1989.010
Profissionais de Saúde (Nível Superior)						
	2019	1095.740	1095.740	7500.000	1531.640	1118.564
	2020–2021	2904.819	2904.819	7500.000	3745.900	2926.588
	2022	1817.412	1817.412	7182.154	2353.443	1823.061

Por fim, os diagnósticos gráficos de convergência apresentados no Apêndice A corroboram os indicadores numéricos discutidos anteriormente. As cadeias exibem comportamento estacionário ao longo das iterações e adequada sobreposição entre as cadeias independentes, reforçando a evidência de convergência e confirmando a robustez das estimativas obtidas pelos modelos ajustados.

3.3 SIMULAÇÃO PARA VALIDAÇÃO DO CENÁRIO EMPÍRICO

Com o objetivo de avaliar o desempenho inferencial do modelo em configurações paramétricas consistentes com a aplicação empírica, foi conduzido um estudo de simulação calibrado a partir das estimativas obtidas na aplicação aos dados do SINAN.

Especificamente, os valores estimados para os parâmetros λ , δ e p foram utilizados como referência para a geração de dados sintéticos segundo a estrutura hierárquica do modelo *N-mixture* dinâmico. Essa estratégia permite avaliar se o procedimento de estimação é capaz de recuperar os parâmetros em cenários que reproduzem a magnitude e a dinâmica observadas no conjunto de dados empírico.

Os resultados apresentados na Tabela 7 indicam que o modelo apresenta bom desempenho inferencial nos cenários calibrados a partir das estimativas obtidas na aplicação aos dados reais. De modo geral, os valores de viés e viés relativo permaneceram reduzidos para todos os parâmetros considerados, indicando que o procedimento de estimação foi capaz de recuperar adequadamente os valores verdadeiros utilizados na geração dos dados.

Para o parâmetro λ , associado à intensidade inicial do processo latente, observa-se viés relativo inferior a aproximadamente 3% em todos os cenários avaliados. Esse resultado sugere boa precisão na recuperação da magnitude do processo de ocorrência latente, mesmo em cenários com diferentes níveis de detecção.

O parâmetro de dinâmica temporal δ apresentou desempenho particularmente estável. Os valores de viés foram praticamente nulos, enquanto o RMSE permaneceu muito baixo em todos os cenários considerados. Esse comportamento indica que o modelo recupera com elevada precisão a dinâmica multiplicativa do processo latente.

Para a probabilidade de detecção p , os resultados também indicam desempenho satisfatório. Os valores de viés relativo permaneceram moderados e a cobertura empírica dos intervalos de credibilidade manteve-se próxima do nível nominal de 95%, evidenciando adequada calibração inferencial para o componente observacional do modelo.

Tabela 7 – Desempenho inferencial dos parâmetros do modelo nos cenários calibrados a partir dos parâmetros estimados nos dados reais.

Períodos	Valores de Referência			$\hat{\lambda}$				$\hat{\delta}$				$\hat{p}$			
	λ	δ	p	Bias	RBias	RMSE	Cob	Bias	RBias	RMSE	Cob	Bias	RBias	RMSE	Cob
72	159,0	1,000	0,520	-0,744	-0,005	9,96	0,95	-0,0001	-0,0001	0,0021	0,93	0,0054	0,0104	0,0308	0,97
24	143,0	0,996	0,711	2,710	0,019	13,10	0,93	0,0004	0,0004	0,0042	0,95	-0,0048	-0,0067	0,0593	0,92
36	166,0	1,000	0,624	4,310	0,026	15,30	0,93	0,0000	0,0000	0,0024	0,97	-0,0064	-0,0103	0,0493	0,94
72	83,5	1,010	0,395	0,631	0,007	7,34	0,96	0,0001	0,0001	0,0024	0,95	0,0026	0,0065	0,0301	0,98
24	81,9	0,998	0,578	0,931	0,011	8,69	0,96	0,0003	0,0003	0,0047	0,95	0,0056	0,0097	0,0588	0,96
36	106,0	1,010	0,497	3,340	0,031	10,20	0,99	0,0001	0,0001	0,0035	0,97	-0,0079	-0,0159	0,0418	0,98

Os resultados referentes às quantidades agregadas do processo latente encontram-se apresentados na Tabela 8. Além das métricas de desempenho inferencial, são apresentadas também as médias das quantidades simuladas $\bar{Y}$, $\bar{N}$ e $\bar{U}$ em cada cenário considerado.

De modo geral, observa-se que o modelo apresentou boa capacidade de recuperação tanto do número total de eventos N_{total} quanto do total de eventos não observados U_{total} . Os valores de viés relativo permaneceram reduzidos na maioria dos cenários, enquanto a cobertura empírica dos intervalos de credibilidade manteve-se próxima ao nível nominal de 95%.

Os resultados mostram que a magnitude da subnotificação simulada (U_{total}) é consistente com os níveis observados na aplicação aos dados reais do SINAN. Em particular, verifica-se que a diferença entre o número total de eventos simulados (N_{total}) e as contagens observadas (Y_{total}) permanece substancial em todos os cenários, refletindo a presença de subnotificação estrutural no processo de geração dos dados.

Em conjunto, esses resultados indicam que o modelo *N-mixture* dinâmico apresenta desempenho inferencial adequado quando aplicado a cenários calibrados a partir das estimativas obtidas nos dados reais. Esse comportamento fornece evidência adicional de que a estrutura hierárquica proposta é capaz de representar adequadamente a magnitude e a dinâmica temporal do processo de ocorrência e do mecanismo de detecção observados na base do SINAN.

Tabela 8 – Desempenho inferencial para os totais de abundância e subnotificação nos cenários calibrados.

Períodos	Valores de Referência			Médias Simuladas			$\hat{N}_{total}$				$\hat{U}_{total}$			
	λ	δ	p	$\bar{Y}$	$\bar{N}$	$\bar{U}$	Bias	RBias	RMSE	Cob	Bias	RBias	RMSE	Cob
72	159,0	1,000	0,520	127.159	244.497	117.338	-592	-0,002	14.103	0,97	-592	-0,005	14.103	0,97
24	143,0	0,996	0,711	47.069	66.211	19.142	1.360	0,021	5.985	0,91	1.360	0,073	5.985	0,91
36	166,0	1,000	0,624	80.328	128.709	48.381	2.994	0,024	11.054	0,94	2.994	0,063	11.054	0,94
72	83,5	1,010	0,395	59.777	151.258	91.481	1.037	0,006	11.823	0,98	1.037	0,011	11.823	0,98
24	81,9	0,998	0,578	22.463	38.867	16.404	427	0,011	4.077	0,96	427	0,027	4.077	0,96
36	106,0	1,010	0,497	42.363	85.276	42.913	2.923	0,033	8.248	0,99	2.923	0,067	8.248	0,99

DISCUSSÃO

Este capítulo apresenta a interpretação integrada dos resultados obtidos ao longo do estudo, articulando os achados provenientes do estudo de simulação, da simulação calibrada a partir dos dados reais e da aplicação empírica do modelo aos dados do Sistema de Informação de Agravos de Notificação (SINAN). A discussão busca examinar o desempenho inferencial do modelo proposto, interpretar os padrões epidemiológicos observados e discutir limitações estruturais da abordagem adotada.

SÍNTESE INTEGRADA DOS ACHADOS

Os resultados obtidos ao longo deste trabalho fornecem evidências consistentes sobre o comportamento inferencial do modelo *N-mixture* dinâmico e sua aplicabilidade à análise de acidentes de trabalho com exposição a material biológico.

No estudo de simulação teórico, conduzido sob diferentes combinações dos parâmetros λ , δ e p , observou-se recuperação adequada dos parâmetros estruturais do modelo e das quantidades derivadas associadas ao processo latente. Os baixos níveis de viés, juntamente com coberturas empíricas próximas ao nível nominal de 95%, indicam que o modelo apresenta bom desempenho inferencial em cenários caracterizados por detecção imperfeita.

A simulação para validação do cenário empírico reforçou esse resultado. Ao reproduzir cenários que refletem a magnitude e a dinâmica observadas na base do SINAN, verificou-se que o modelo manteve desempenho inferencial estável, com viés relativo reduzido e cobertura adequada tanto para os parâmetros estruturais quanto para as quantidades agregadas do processo latente. Esses resultados indicam que o modelo permanece robusto mesmo em regimes paramétricos próximos daqueles observados na aplicação empírica.

Por fim, a aplicação do modelo aos dados do SINAN revelou padrões estruturais relevantes na dinâmica dos acidentes de trabalho com exposição a material biológico. Em particular, três resultados se destacam: (i) maior intensidade latente estimada entre Técnicos e Auxiliares da Saúde; (ii) estabilidade relativa da trajetória temporal do processo latente dentro

de cada período epidemiológico analisado; e (iii) variações expressivas na probabilidade de detecção ao longo do período considerado.

Esses resultados evidenciam a utilidade da decomposição entre processo latente de ocorrência e processo observacional de detecção, princípio central da estrutura *N-mixture*. Em sistemas de vigilância epidemiológica baseados em notificação passiva, nos quais a subnotificação constitui fenômeno plausível, essa separação permite interpretar de forma mais realista a dinâmica das séries temporais observadas.

4.1 DESEMPENHO INFERENCIAL DO MODELO

RECUPERAÇÃO DOS PARÂMETROS ESTRUTURAIS

Os resultados do estudo de simulação indicam que o modelo apresentou desempenho satisfatório na recuperação dos parâmetros estruturais λ , δ e p . De modo geral, os valores de viés e viés relativo permaneceram reduzidos, enquanto a cobertura empírica dos intervalos de credibilidade manteve-se próxima ao nível nominal.

O parâmetro associado à dinâmica temporal do processo latente, δ , apresentou desempenho particularmente robusto. As estimativas exibiram viés praticamente nulo e baixos valores de erro quadrático médio, indicando que a estrutura de dependência temporal do modelo é bem identificada a partir das contagens observadas.

Esse resultado sugere que a informação contida na sequência temporal das notificações é suficiente para recuperar adequadamente a dinâmica multiplicativa do processo latente, mesmo na presença de detecção imperfeita.

ESTIMATIVA DA INTENSIDADE DO PROCESSO LATENTE

A intensidade inicial do processo latente, representada pelo parâmetro λ , também apresentou recuperação satisfatória nos cenários considerados. Embora o erro quadrático médio aumente em cenários caracterizados por maior magnitude do processo latente, esse comportamento é esperado, uma vez que o erro absoluto tende a crescer com a escala do parâmetro.

Quando analisado em termos relativos, o desempenho permanece adequado, indicando que o modelo é capaz de estimar de forma consistente o nível basal do processo de ocorrência.

IDENTIFICABILIDADE DO COMPONENTE DE DETECÇÃO

Um aspecto importante observado no estudo de simulação refere-se ao comportamento do parâmetro de detecção em cenários nos quais $p = 1$. Nessa situação limite, a cobertura empírica dos intervalos de credibilidade torna-se nula, refletindo perda de identificabilidade do mecanismo de detecção.

Quando todas as ocorrências são observadas, isto é, quando $Y_{i,t} = N_{i,t}$, não existe informação suficiente nos dados para distinguir entre o processo latente e o mecanismo observacional. Esse resultado não representa falha do procedimento de estimação, mas constitui consequência direta da estrutura probabilística do modelo.

Esse comportamento reforça que a abordagem *N-mixture* é particularmente apropriada em contextos nos quais existe plausibilidade de detecção imperfeita, situação frequentemente observada em sistemas de vigilância epidemiológica.

4.2 INTERPRETAÇÃO EPIDEMIOLÓGICA DOS RESULTADOS

DIFERENÇAS OCUPACIONAIS

A aplicação do modelo indicou uma intensidade latente de ocorrência superior entre técnicos e auxiliares de saúde em comparação aos profissionais de nível superior. Tal disparidade pode ser atribuída à maior exposição a procedimentos invasivos, à natureza das atividades assistenciais e à presença contínua em ambientes hospitalares críticos, onde o risco biológico é mais acentuado.

Contudo, deve-se considerar que diferenças observadas entre grupos ocupacionais podem refletir não apenas variações no processo de ocorrência, mas também possíveis diferenças nos padrões de notificação.

DINÂMICA TEMPORAL

As estimativas do parâmetro δ permaneceram próximas da unidade em praticamente todos os modelos ajustados. Esse resultado sugere que, dentro de cada período epidemiológico analisado, o processo latente apresentou trajetória aproximadamente estável ao longo do tempo.

Essa estabilidade sugere que parte das variações observadas nas séries de notificações pode estar associada a mudanças institucionais ou operacionais no processo de registro dos eventos, e não necessariamente a alterações persistentes na ocorrência real dos acidentes.

SUBNOTIFICAÇÃO

Um dos resultados mais relevantes da análise empírica refere-se à magnitude estimada da subnotificação. As estimativas do número de eventos não observados indicam diferença sistemática entre o número total de ocorrências latentes e as notificações registradas no sistema.

Essa diferença sugere que a subnotificação constitui componente relevante na interpretação das séries temporais de acidentes com material biológico. A quantificação dessa

discrepância entre ocorrência real e registros oficiais representa uma contribuição importante para a interpretação epidemiológica de dados provenientes de sistemas de vigilância passiva.

LIMITAÇÕES DO ESTUDO

Apesar dos resultados consistentes obtidos, algumas limitações estruturais devem ser consideradas.

A principal limitação refere-se à hipótese de probabilidade de detecção constante dentro de cada período epidemiológico. Em sistemas reais de vigilância, é plausível que a probabilidade de notificação varie entre regiões, instituições ou características demográficas dos trabalhadores.

Outra limitação diz respeito à estrutura de dependência temporal adotada no modelo. A dinâmica considerada assume dependência de primeira ordem entre períodos consecutivos. Embora essa estrutura capture adequadamente a persistência temporal do processo, ela pode não representar completamente choques sistêmicos ou efeitos de memória mais longa.

CONSIDERAÇÕES FINAIS

Os resultados obtidos neste trabalho indicam que o modelo *N-mixture* dinâmico constitui uma ferramenta metodologicamente consistente para investigar processos de ocorrência sujeitos à detecção imperfeita em sistemas de vigilância epidemiológica.

O estudo de simulação demonstrou que o modelo recupera adequadamente parâmetros estruturais e quantidades derivadas em cenários caracterizados por subnotificação, enquanto a simulação com cenário empírico evidenciou estabilidade inferencial em regimes paramétricos próximos aos observados nos dados reais.

A aplicação aos dados do SINAN revelou padrões relevantes na dinâmica dos acidentes de trabalho com exposição a material biológico, indicando diferenças ocupacionais e evidenciando a presença de subnotificação persistente.

Em conjunto, esses resultados reforçam o potencial de abordagens hierárquicas que separam explicitamente o processo latente de ocorrência do mecanismo observacional de registro, contribuindo para interpretações mais robustas de séries epidemiológicas baseadas em dados de vigilância passiva.

CONCLUSÃO

Este trabalho teve como objetivo investigar a ocorrência latente de acidentes de trabalho com exposição a material biológico e quantificar possíveis níveis de subnotificação nos registros do Sistema de Informação de Agravos de Notificação (SINAN), por meio da aplicação de um modelo hierárquico do tipo *N-mixture* dinâmico. A estrutura de modelagem adotada permite separar explicitamente o processo latente de ocorrência do agravo do processo observacional de notificação, possibilitando inferência probabilística tanto sobre o número real de eventos quanto sobre a fração efetivamente registrada pelo sistema de vigilância.

Para avaliar o comportamento inferencial do modelo, foram conduzidas diferentes estratégias de investigação. Inicialmente, um estudo de simulação Monte Carlo foi realizado sob distintos cenários paramétricos, variando a intensidade inicial do processo latente, a dinâmica temporal e a probabilidade de detecção. De modo geral, os resultados indicaram bom desempenho inferencial em cenários caracterizados por detecção imperfeita, com baixos níveis de viés, erro quadrático médio moderado e cobertura empírica próxima ao nível nominal dos intervalos de credibilidade. Em particular, o parâmetro associado à dinâmica temporal do processo latente apresentou estimativas bastante estáveis, evidenciando que a estrutura temporal do modelo é bem identificada a partir das contagens observadas.

Complementarmente, foi conduzido um estudo de simulação calibrado a partir das estimativas obtidas na aplicação aos dados reais. Nessa etapa, os valores estimados para os parâmetros do modelo foram utilizados como referência para gerar dados sintéticos que reproduzem a magnitude e a dinâmica observadas na base empírica do SINAN. Os resultados dessa simulação indicaram que o modelo manteve desempenho inferencial consistente mesmo em regimes paramétricos próximos à aplicação empírica, reforçando a robustez da abordagem quando utilizada em contextos epidemiológicos reais.

Na aplicação empírica aos dados do SINAN, referentes ao período de 2014 a 2024, os resultados revelaram padrões estruturais relevantes na dinâmica dos acidentes de trabalho com exposição a material biológico. Observou-se que a intensidade inicial do processo latente estimada foi sistematicamente maior entre técnicos e auxiliares da saúde quando com-

parada ao grupo de profissionais de saúde de nível superior, sugerindo maior magnitude basal de ocorrência nesse grupo ocupacional. Além disso, o parâmetro associado à dinâmica temporal permaneceu próximo da unidade nos diferentes períodos analisados, indicando relativa estabilidade da trajetória latente dentro de cada bloco epidemiológico considerado.

Outro resultado importante refere-se à variação observada na probabilidade de detecção ao longo do tempo. As estimativas indicaram aumento da probabilidade média de notificação durante o período de 2020–2021, seguido de redução parcial a partir de 2022. Esse comportamento pode refletir mudanças institucionais e organizacionais no sistema de vigilância associadas ao contexto da pandemia de COVID-19 e às adaptações subsequentes nos processos de registro dos eventos.

As estimativas das quantidades derivadas do processo latente indicaram a presença de subnotificação persistente ao longo da série temporal analisada. A diferença sistemática entre o número estimado de ocorrências latentes e as notificações observadas sugere que os registros do sistema podem representar apenas uma fração da ocorrência real do agravo. Ao quantificar explicitamente essa diferença e sua incerteza associada, a abordagem proposta contribui para uma interpretação mais realista das séries de dados provenientes de sistemas de vigilância epidemiológica baseados em notificação passiva.

Apesar dos resultados consistentes obtidos, algumas limitações devem ser consideradas. A principal refere-se à hipótese de probabilidade de detecção homogênea dentro de cada período epidemiológico analisado, adotada em função de restrições de identificabilidade e da ausência de covariáveis informativas sobre o processo de notificação. Além disso, a estrutura temporal considerada assume dependência de primeira ordem entre períodos consecutivos, o que pode não capturar completamente possíveis efeitos estruturais de longo prazo ou heterogeneidades adicionais presentes nas séries observadas.

Como perspectivas de aprofundamento, destaca-se a possibilidade de incorporar covariáveis explicativas tanto no processo latente quanto no componente de detecção, permitindo investigar fatores organizacionais, regionais ou institucionais associados à ocorrência e à notificação dos eventos. Extensões que permitam probabilidades de detecção variáveis no tempo ou entre estratos, bem como a inclusão de efeitos aleatórios ou componentes espaciais, podem ampliar a flexibilidade da modelagem e melhorar a representação das heterogeneidades presentes em sistemas reais de vigilância.

Adicionalmente, uma extensão metodológica promissora consiste na incorporação de esquemas de observação baseados em processos de remoção, nos quais os eventos detectados são progressivamente retirados do conjunto de casos disponíveis para observação. Essa formulação aproxima a estrutura probabilística do modelo da dinâmica operacional de muitos sistemas de notificação e pode contribuir para melhorar a identificabilidade entre abundância latente e probabilidade de detecção.

Em síntese, os resultados obtidos indicam que o modelo *N-mixture* dinâmico constitui uma ferramenta estatisticamente consistente e epidemiologicamente informativa para investigar subnotificação em séries temporais provenientes de sistemas de vigilância. Ao integrar evidências provenientes de estudos de simulação e aplicação empírica, este trabalho demonstra o potencial de abordagens hierárquicas que distinguem explicitamente o processo de ocorrência real do mecanismo de registro observacional, contribuindo para o aprimoramento da análise quantitativa de agravos ocupacionais e para o fortalecimento da produção de evidências em saúde pública.

REFERÊNCIAS

- ARAÚJO, T. M. d. et al. Acidentes de trabalho com exposição a material biológico entre os profissionais de enfermagem. *Revista de Enfermagem Referência*, v. 3, n. 7, p. 7–14, 2012.
- BARKER, R. J. et al. On the reliability of n-mixture models for count data. *Biometrics*, v. 74, n. 1, p. 369–377, 2018.
- BRINTZ, B.; FUENTES, C.; MADSEN, L. An asymptotic approximation to the N-mixture model for the estimation of disease prevalence. *Biometrics*, v. 74, n. 4, p. 1512–1518, 2018.
- CARNEIRO, T. M. S. et al. Acidentes de trabalho com exposição a material biológico: descrição dos casos na bahia. *Revista de Epidemiologia e Controle de Infecção*, v. 6, n. 2, p. 50–56, 2016.
- CARPENTER, B. et al. Stan: A probabilistic programming language. *Journal of Statistical Software*, v. 76, n. 1, p. 1–32, 2017.
- CRESSIE, N.; WIKLE, C. K. *Statistics for Spatio-Temporal Data*. Hoboken, NJ: Wiley, 2011. ISBN 9780471692744.
- DAIL, D.; MADSEN, L. Models for estimating abundance from repeated counts of an open metapopulation. *Biometrics*, v. 67, n. 2, p. 577–587, 2011.
- DENNIS, E. B.; MORGAN, B. J. T.; RIDOUT, M. S. Computational aspects of n-mixture models. *Biometrics*, v. 71, n. 1, p. 237–246, 2015.
- FERREIRA, W. F. da S. et al. Consequências ao meio ambiente e a saúde com descarte irregular de perfurocortantes no estado do paran . *ID on line. Revista de psicologia*, v. 19, n. 76, p. 140–158, 2025.
- GELMAN, A.; RUBIN, D. B. Inference from iterative simulation using multiple sequences. *Statistical Science*, v. 7, n. 4, p. 457–472, 1992.
- GOMES, S. C. S.; CALDAS, A. d. J. M. Qualidade dos dados do sistema de informa  o sobre acidentes de trabalho com exposi  o a material biol gico no brasil. *Revista Brasileira de Medicina do Trabalho*, v. 15, n. 3, p. 200–208, 2017.
- GOMES, S. C. S.; CALDAS, A. d. J. M. Incid ncia de acidentes de trabalho com exposi  o a material biol gico em profissionais de sa de no brasil, 2010–2016. *Revista Brasileira de Medicina do Trabalho*, v. 17, n. 2, p. 188–200, 2019.

- HOOK, E. B.; REGAL, R. R. Capture-recapture methods in epidemiology: methods and limitations. *Epidemiologic Reviews*, v. 17, n. 2, p. 243–264, 1995.
- LINCOLN, F. C. *Calculating Waterfowl Abundance on the Basis of Banding Returns*. Washington, DC, 1930.
- LUNN, D. et al. *The BUGS Book: A Practical Introduction to Bayesian Analysis*. [S.l.]: CRC Press, 2012.
- LUNN, D. J. et al. The bugs project: Evolution, critique and future directions. *Statistics in Medicine*, v. 28, n. 25, p. 3049–3067, 2009.
- MARINHO, D. P. V. *Acidentes de trabalho com exposição a material biológico entre trabalhadores da saúde no Brasil*. Dissertação (Dissertação (Mestrado em Saúde Coletiva)) — Universidade Estadual de Feira de Santana (UEFS), Feira de Santana, BA, jul 2018. Área de concentração: Epidemiologia. Orientadora: Prof. Dr. Maura Maria Guimarães de Almeida.
- NOGUEIRA, S. A. et al. Prevalência e notificações de acidentes de trabalho com exposição a material biológico na odontologia. *Revista Ciência Plural*, v. 2, n. 1, p. 102–119, 2016.
- OTIS, D. L. et al. Statistical inference from capture data on closed animal populations. *Wildlife Monographs*, n. 62, p. 3–135, 1978.
- PARKER, M. R. P. et al. Faster asymptotic solutions for N-mixtures on large populations. *Journal of Agricultural, Biological and Environmental Statistics*, p. 1–16, 2024.
- PLUMMER, M. Jags: A program for analysis of bayesian graphical models using gibbs sampling. In: *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*. Vienna, Austria: [s.n.], 2003. p. 1–10.
- R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2024. Disponível em: <<https://www.R-project.org>>.
- ROYLE, J. A. N-mixture models for estimating population size from spatially replicated counts. *Biometrics*, v. 60, n. 1, p. 108–115, 2004.
- RUE, H.; MARTINO, S.; CHOPIN, N. Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the Royal Statistical Society: Series B*, v. 71, n. 2, p. 319–392, 2009.
- SANDERLIN, J. S. et al. Occupancy modeling and resampling overcomes low test sensitivity to produce accurate SARS-CoV-2 prevalence estimates. *BMC Public Health*, v. 21, p. 1095, 2021.
- SEBEN, Y. P.; MORETTO, C. F. Estratégias de enfrentamento em acidentes de trabalho com exposição ao material biológico. *Psicologia: Ciência e Profissão*, v. 42, p. e181772, 2022.
- SOARES, R. Z.; SCHOEN, A. S.; BENELLI, K. d. R. G. Análise dos acidentes de trabalho com exposição a material biológico notificados por profissionais da saúde. *Revista Brasileira de Medicina do Trabalho*, v. 17, n. 2, p. 201–208, 2019.
- VALPINE, P. de et al. Programming with models: writing statistical algorithms for general model structures with nimble. *Journal of Computational and Graphical Statistics*, v. 26, n. 2, p. 403–413, 2017.

VATS, D.; KNUDSON, C. Revisiting the Gelman–Rubin diagnostic. *Statistical Science*, Institute of Mathematical Statistics, v. 36, n. 4, p. 518–529, 2021.

WIKLE, C. K. Hierarchical bayesian models for ecological processes. *Ecological Applications*, v. 13, n. 3, p. 1–23, 2003.

ZIPPIN, C. An evaluation of the removal method of estimating animal populations. *Biometrics*, v. 12, n. 2, p. 163–189, 1956.

Apêndices

DIAGNÓSTICO DE CONVERGÊNCIA DAS CADEIAS

Neste apêndice, apresentam-se os diagnósticos de convergência MCMC para os parâmetros λ , δ , p e U_{total} dos modelos ajustados para os dados reais do SINAN. Para cada parâmetro, são exibidos o gráfico de traço (*traceplot*), a estimativa da densidade *a posteriori* e as funções de autocorrelação (ACF) para cada uma das três cadeias independentes.

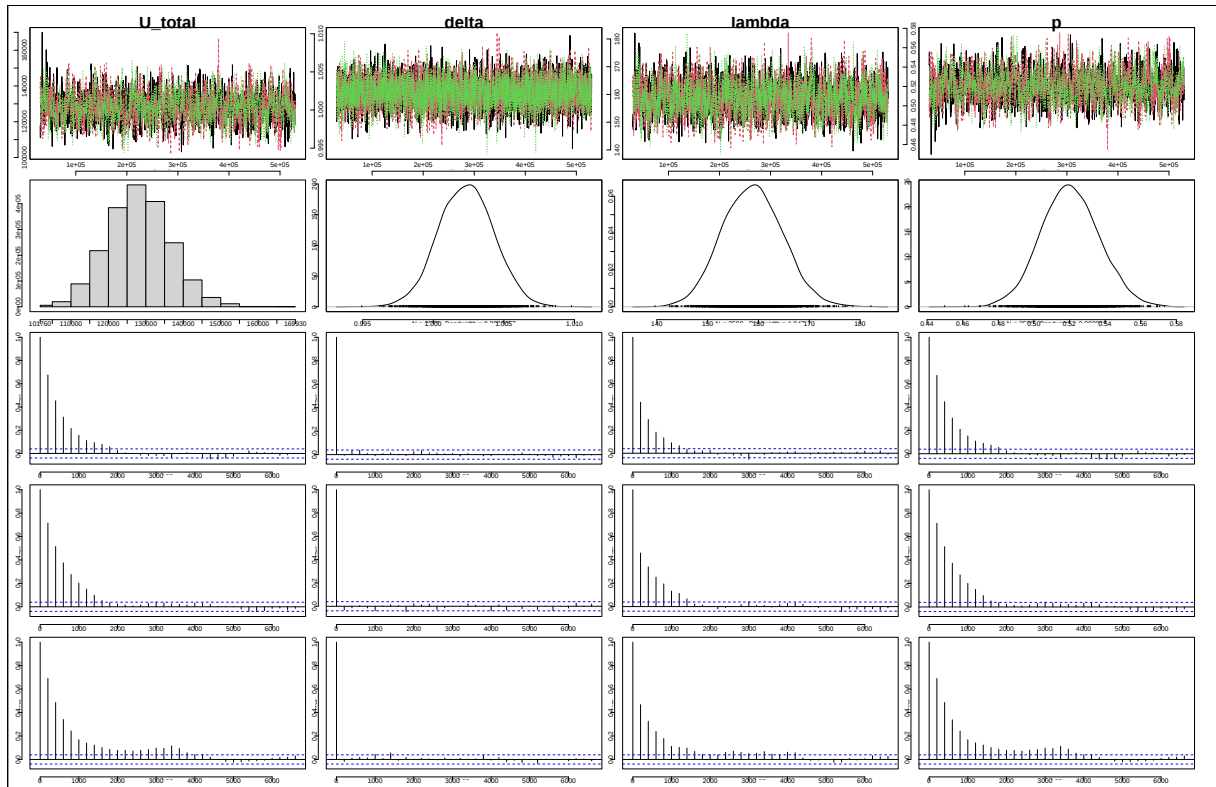


Figura 5 – Diagnóstico de convergência MCMC: Técnicos e Auxiliares da Saúde, Para o período de 2014–2019.

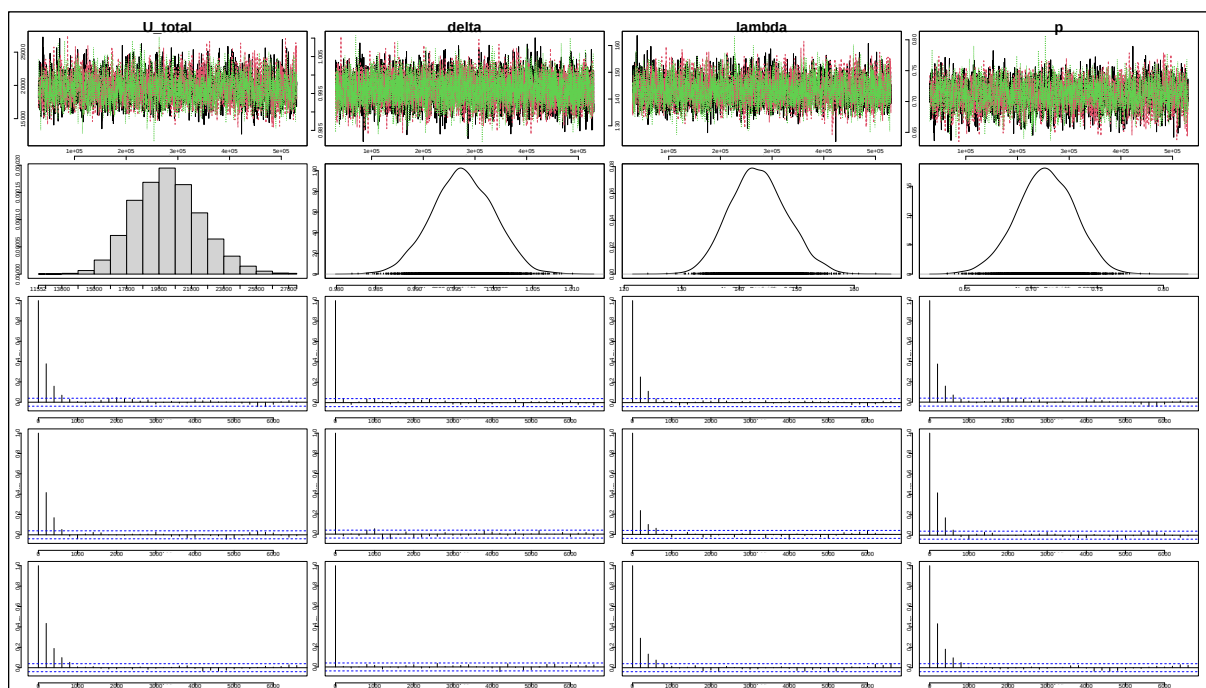


Figura 6 – Diagnóstico de convergência MCMC: Técnicos e Auxiliares da Saúde, para o período de 2020–2021.

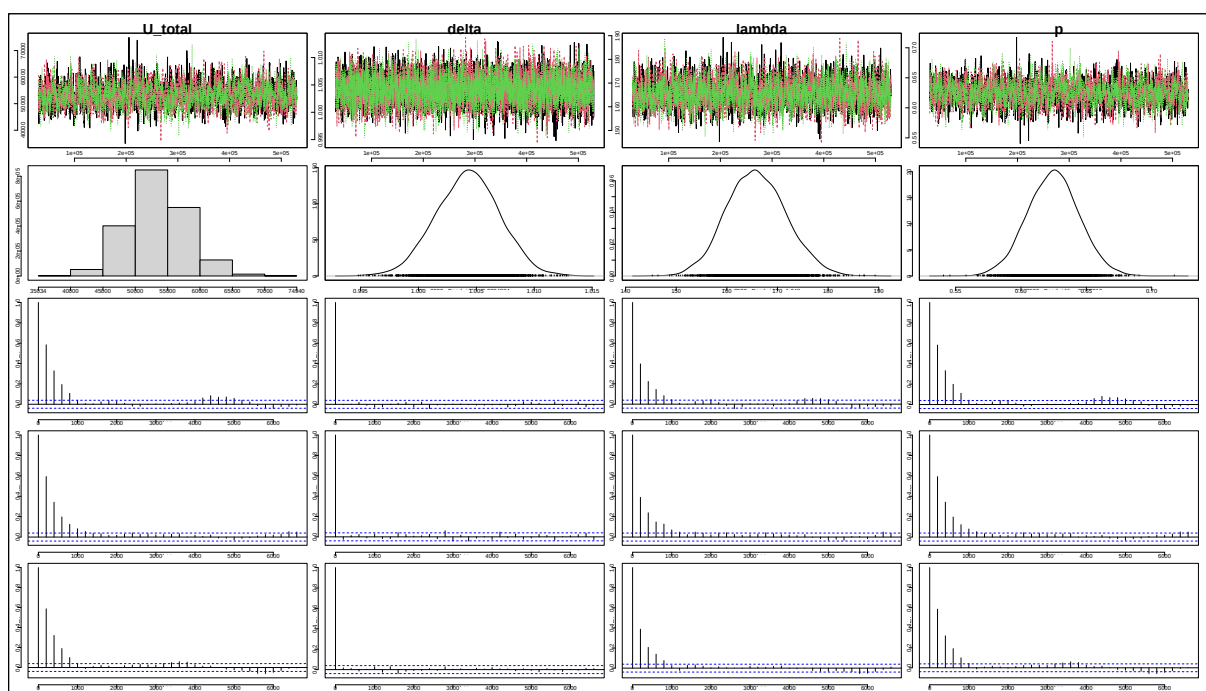


Figura 7 – Diagnóstico de convergência MCMC: Técnicos e Auxiliares da Saúde, para o período de 2022–2024.

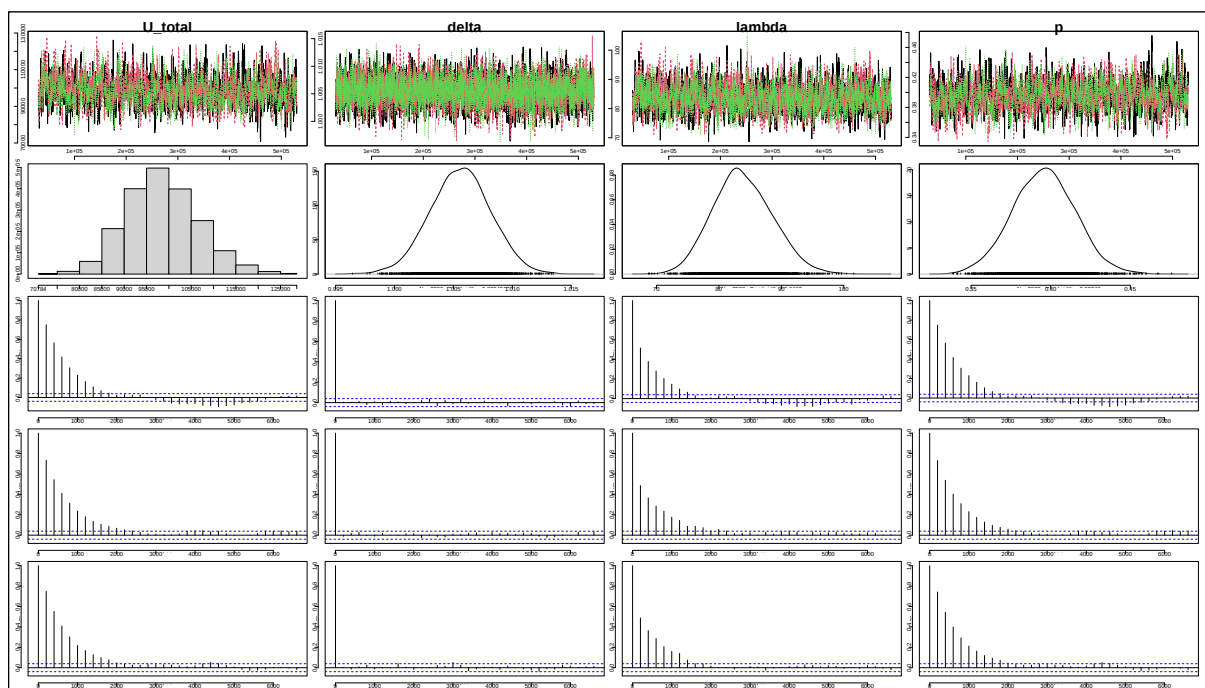


Figura 8 – Diagnóstico de convergência MCMC: Profissionais de Saúde (Nível Superior), para o período de 2014–2019.

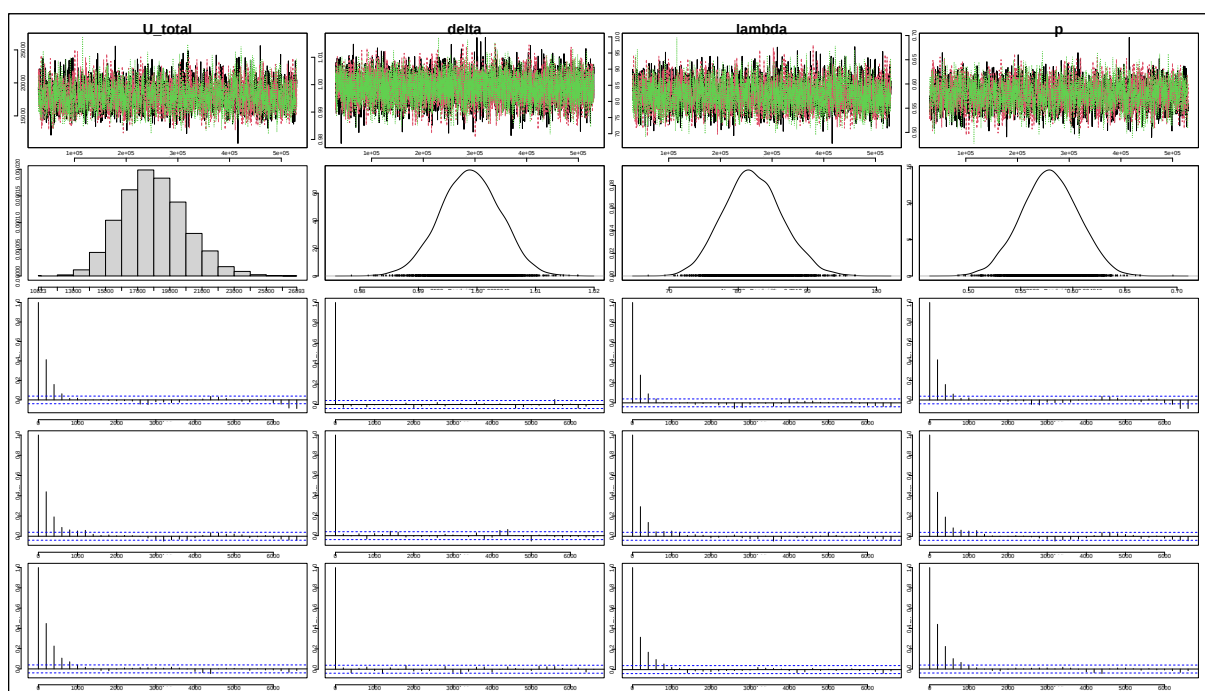


Figura 9 – Diagnóstico de convergência MCMC: Profissionais de Saúde (Nível Superior), para o período de 2020–2021.

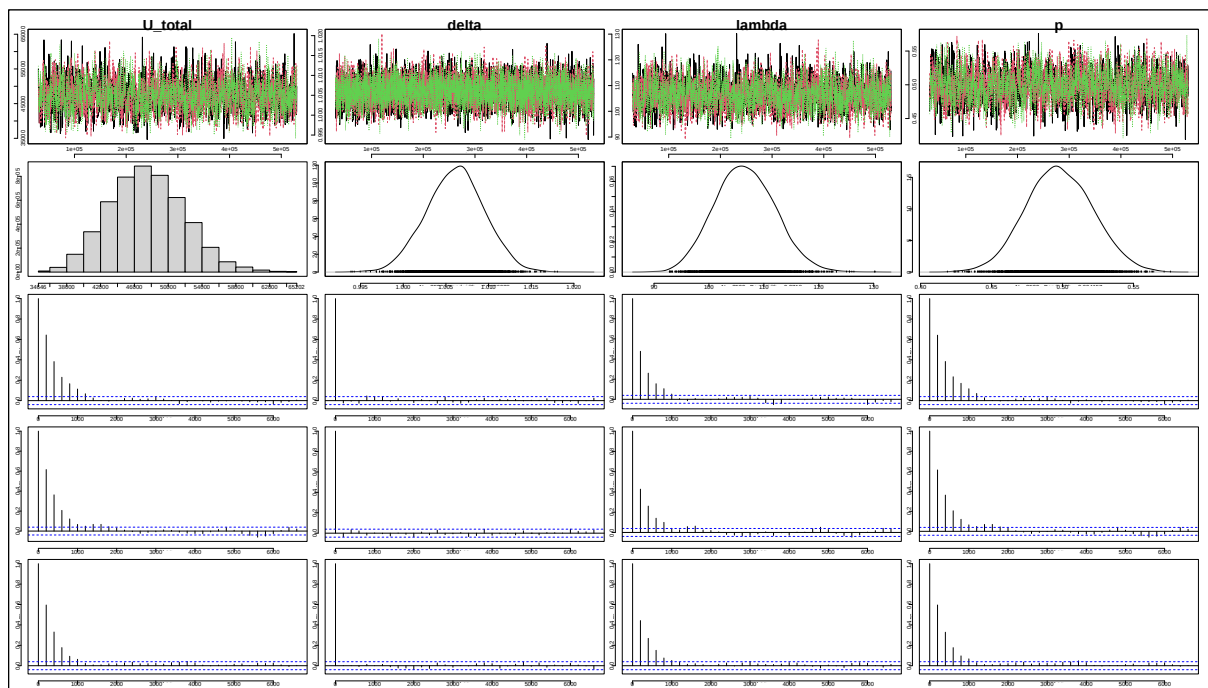


Figura 10 – Diagnóstico de convergência MCMC: Profissionais de Saúde (Nível Superior), para o período de 2022–2024.

IMPLEMENTAÇÃO DO MODELO COM DADOS SIMULADOS

Neste apêndice, apresenta-se o código desenvolvido em linguagem R (R Core Team, 2024) utilizando o pacote `rjags` (RUE; MARTINO; CHOPIN, 2009) para a implementação do modelo N-mixture dinâmico e realização do estudo de simulação.

```

1 # =====
2 # SIMULACAO N-MIXTURE DINAMICO - MONTE CARLO (reps=100)
3 # com Ntotal_true, Utotal_true e Utotal_est (IC 95%)
4 # =====
5
6 dyn.load("/miniconda3/lib/libjags.so.4")
7
8 # =====
9 # PACOTES
10 # =====
11
12 library(rjags)
13 library(coda)
14 library(parallel)
15 library(dplyr)
16 library(tidyr)
17
18 # =====
19 # CENARIOS
20 # =====
21
22 cenarios <- expand.grid(
23   lambda = c(100, 500),
24   delta  = c(0.95, 1, 1.05),
25   p      = c(0.5, 0.75, 1)
26 )
27

```

```

28 reps <- 100
29
30 grid_sim <- do.call(rbind,
31                     lapply(1:nrow(cenarios), function(i){
32                         cbind(cenarios[i,], seed = (1:reps) + 665*i)
33                     })
34 )
35
36
37 # =====
38 # MODELO JAGS
39 # =====
40
41 modelo_string <- "
42 model {
43
44     for (i in 1:n){
45
46         N[i, 1] ~ dpois(lambda)
47         y[i, 1] ~ dbin(p, N[i, 1])
48
49         for (t in 2:R){
50             N[i, t] ~ dpois(delta * N[i, t-1])
51             y[i, t] ~ dbin(p, N[i, t])
52         }
53
54         N_i[i] <- sum(N[i, 1:R])
55     }
56
57     beta0 ~ dnorm(0, 0.01)
58     lambda <- exp(beta0)
59
60     p ~ dbeta(1,1)
61
62     eta_gain ~ dnorm(0, 0.01)
63     delta <- exp(eta_gain)
64
65     N_total <- sum(N_i[1:n])
66 }
67 "
68
69 # =====
70 # FUNCAO DE SIMULACAO (100 replica)
71 # =====
72
73 rodar_simulacao <- function(params){
74

```

```
75 lambda_true <- params$lambda
76 delta_true  <- params$delta
77 p_true      <- params$p
78 seed        <- params$seed
79
80 set.seed(seed)
81
82 n <- 20
83 T <- 24
84
85 # -----
86 # GERACAO DOS DADOS
87 # -----
88 N <- matrix(NA, n, T)
89 Y <- matrix(NA, n, T)
90
91 for(i in 1:n){
92   N[i,1] <- rpois(1, lambda_true)
93   Y[i,1] <- rbinom(1, N[i,1], p_true)
94
95   for(t in 2:T){
96     N[i,t] <- rpois(1, delta_true * N[i,t-1])
97     Y[i,t] <- rbinom(1, N[i,t], p_true)
98   }
99 }
100
101 Y_total_obs <- sum(Y)
102 Ntotal_true <- sum(N)
103 Utotal_true <- Ntotal_true - Y_total_obs
104
105 Y_t_obs <- colSums(Y)
106 names(Y_t_obs) <- paste0("Yt_", 1:T)
107
108 # -----
109 # AJUSTE JAGS
110 # -----
111 dados_jags <- list(y = Y, n = n, T = T)
112
113 inits <- function() list(
114   beta0 = log(lambda_true),
115   p = 0.5,
116   eta_gain = log(delta_true),
117   N = Y + 5
118 )
119
120 resultado <- tryCatch({
121
```

```
122 mod <- jags.model(  
123   textConnection(modelo_string),  
124   data = dados_jags,  
125   inits = inits,  
126   n.chains = 3,  
127   n.adapt = 5000  
128 )  
129  
130 update(mod, 10000)  
131  
132 amostras <- coda.samples(  
133   mod,  
134   variable.names = c("lambda", "delta", "p", "N_total"),  
135   n.iter = 100000,  
136   thin = 50  
137 )  
138  
139 resumo <- summary(amostras)  
140 stats <- resumo$statistics  
141 quant <- resumo$quantiles  
142  
143 # -----  
144 # parametros  
145 # -----  
146 out_base <- c(  
147   lambda_est = stats["lambda", "Mean"],  
148   delta_est  = stats["delta", "Mean"],  
149   p_est      = stats["p", "Mean"],  
150  
151   lambda_inf = quant["lambda", "2.5%"],  
152   lambda_sup = quant["lambda", "97.5%"],  
153  
154   delta_inf  = quant["delta", "2.5%"],  
155   delta_sup  = quant["delta", "97.5%"],  
156  
157   p_inf      = quant["p", "2.5%"],  
158   p_sup      = quant["p", "97.5%"]  
159 )  
160  
161 # -----  
162 # N_total  
163 # -----  
164 Ntotal_est <- stats["N_total", "Mean"]  
165 Ntotal_inf <- quant["N_total", "2.5%"]  
166 Ntotal_sup <- quant["N_total", "97.5%"]  
167  
168 # -----
```

```

169 # U_total (posterior) = N_total - Y_total_obs
170 # (Y_total_obs e fixo na replica)
171 # -----
172 Utotal_est <- Ntotal_est - Y_total_obs
173 Utotal_inf <- Ntotal_inf - Y_total_obs
174 Utotal_sup <- Ntotal_sup - Y_total_obs
175
176 c(
177   # verdadeiros (da geracao)
178   Ntotal_true = Ntotal_true,
179   Utotal_true = Utotal_true,
180   Y_total     = Y_total_obs,
181
182   # estimativas
183   out_base,
184   Ntotal_est = Ntotal_est,
185   Ntotal_inf = Ntotal_inf,
186   Ntotal_sup = Ntotal_sup,
187
188   Utotal_est = Utotal_est,
189   Utotal_inf = Utotal_inf,
190   Utotal_sup = Utotal_sup
191 )
192
193 }, error = function(e){
194   rep(NA, 3 + 9 + 3 + 3) # 18 columnas
195 })
196
197 return(resultado)
198 }
199
200 # =====
201 # PARALELIZACAO
202 # =====
203
204 n_cores <- max(1, floor(detectCores() * 3/4))
205 cl <- makeCluster(n_cores)
206
207 clusterExport(cl, c("modelo_string", "rodar_simulacao"))
208 clusterEvalQ(cl, {dyn.load("/miniconda3/lib/libjags.so.4"); library(
209   rjags); library(coda) })
210
211 resultados <- parLapply(
212   cl,
213   split(grid_sim, 1:nrow(grid_sim)),
214   rodar_simulacao
215 )

```

```

215
216 stopCluster(cl)
217
218 # =====
219 # ORGANIZACAO
220 # =====
221
222 result_matrix <- do.call(rbind, resultados)
223
224 resultados_df <- cbind(grid_sim, result_matrix) %>%
225   as.data.frame()
226
227 # converter tudo pra numerico
228 resultados_df <- resultados_df %>%
229   mutate(across(everything(), as.numeric)) %>%
230   filter(!is.na(lambda_est))
231
232 # Remover simulacoes com erro
233 resultados_df <- resultados_df %>%
234   filter(!is.na(lambda_est))
235
236 save(
237   resultados_df,
238   file = "resultados_simulacao_Nmixture.RData"
239 )
240
241 # =====
242 # CALCULO DE COBERTURA
243 # =====
244
245 load('resultados_simulacao_Nmixture.RData')
246
247 resultados_df <- resultados_df %>%
248   mutate(
249     cover_lambda = ifelse(lambda >= lambda_inf & lambda <= lambda_sup, 1, 0),
250     cover_delta  = ifelse(delta >= delta_inf & delta <= delta_sup, 1, 0),
251     cover_p      = ifelse(p >= p_inf & p <= p_sup, 1, 0)
252   )
253
254 resumo_sim <- resultados_df %>%
255   group_by(lambda, delta, p) %>%
256   summarise(
257     # --- LAMBDA ---
258     Bias_lambda      = mean(lambda_est - lambda),
259     RBias_lambda     = mean((lambda_est - lambda) / lambda),
260     RMSE_lambda      = sqrt(mean((lambda_est - lambda)^2)),
261     Cob_lambda       = mean(cover_lambda),

```

```
262
263   # --- DELTA ---
264   Bias_delta      = mean(delta_est - delta),
265   RBias_delta     = mean((delta_est - delta) / delta),
266   RMSE_delta      = sqrt(mean((delta_est - delta)^2)),
267   Cob_delta       = mean(cover_delta),
268
269   # --- P ---
270   Bias_p          = mean(p_est - p),
271   RBias_p         = mean((p_est - p) / p),
272   RMSE_p          = sqrt(mean((p_est - p)^2)),
273   Cob_p           = mean(cover_p),
274
275   .groups = "drop"
276 )
```

Listing B.1 – Código em R para simulação e ajuste do modelo via JAGS.