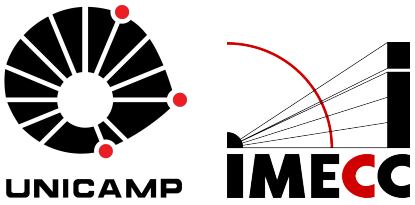


UNIVERSIDADE ESTADUAL DE CAMPINAS

Instituto de Matemática, Estatística e Computação Científica

Disciplina: ME720 – Modelos Lineares Generalizados

Modelagem Bayesiana da Inativação Térmica de Bacillus simplex Através de um Modelo Linear Generalizado com Distribuição de Poisson

Autor:

João Vitor Nunes da Silva – 175072

Professora:

Profª. Drª. Hildete Prisco Pinheiro

2º semestre de 2025

Sumário

1 Introdução

A segurança dos alimentos está diretamente relacionada ao entendimento de como os microrganismos se comportam em diferentes condições de processamento, especialmente quando expostos ao calor Lammerding and Fazil (2000). Para garantir que os processos sejam seguros e eficientes, é comum utilizar modelos matemáticos que descrevem como os microrganismos são inativados ao longo do tempo Gil et al. (2017). Esses modelos ajudam tanto no planejamento do processamento térmico quanto na avaliação de riscos ao longo da cadeia produtiva.

Tradicionalmente, esses modelos são ajustados aos dados experimentais por meio do método dos mínimos quadrados, que parte da suposição de que os erros seguem uma distribuição normal van Boekel (2020). No entanto, esse tipo de abordagem nem sempre é o mais adequado para dados microbiológicos, já que as contagens de bactérias são números inteiros e nunca negativos. A distribuição normal, por outro lado, é contínua e pode gerar valores negativos, o que não faz sentido nesse contexto. Além disso, métodos estatísticos tradicionais costumam fornecer apenas estimativas pontuais dos parâmetros, sem representar bem a incerteza dos resultados ou as variações naturais do experimento Nauta (2000).

Uma alternativa mais adequada é o uso de Modelos Lineares Generalizados (GLM) dentro de uma abordagem bayesiana. Esse tipo de modelo permite utilizar distribuições mais compatíveis com dados de contagem, como a Poisson, e também quantificar de forma explícita a incerteza dos parâmetros através das distribuições de probabilidade obtidas na inferência bayesiana Pouillot et al. (2003); Koyama et al. (2019). Além disso, essa metodologia consegue lidar naturalmente com observações de contagens iguais a zero, algo comum em experimentos microbiológicos.

Neste trabalho, aplicamos a metodologia proposta por Hiura et al. (2021) para analisar os dados de inativação térmica de *Bacillus simplex* a 94°C publicados por Abe et al. (2019). Foi utilizado o modelo de Weibull com distribuição de Poisson para os resíduos, com o objetivo de: (i) estimar os parâmetros cinéticos δ e p por meio de inferência bayesiana via MCMC usando dados de concentração inicial de 10^5 células; (ii) validar o modelo com simulações estocásticas em concentrações menores (10^3, 10^2 e 10^1 células), comparando as previsões com 60 experimentos independentes; e (iii) avaliar a incerteza dos parâmetros a partir das distribuições posteriores. Essa aplicação busca demonstrar, de forma prática, como a abordagem bayesiana pode ser uma ferramenta útil e mais robusta na análise de dados de inativação microbiana, contribuindo para estudos de segurança alimentar.

2 Materiais e Métodos

2.1 Dataset

Os dados utilizados neste estudo foram obtidos do trabalho publicado por Abe et al. (2019) e disponibilizados publicamente por Hiura et al. (2021). O dataset consiste em observações de inativação térmica de esporos de *Bacillus simplex*, uma bactéria psicrofílica formadora de esporos originária de leite pasteurizado da Organização de Pesquisa de Hokkaido (Japão).

Os dados de treinamento compreendem três replicações independentes de inativação térmica a 94°C com concentração inicial de 10^5 células. As suspensões bacterianas foram submetidas a tratamento térmico isotérmico, e as contagens viáveis foram estimadas através de plaqueamento em ágar nutriente (Eiken, Tóquio, Japão) após incubação a 30°C por 2 dias. As medições foram realizadas em nove pontos temporais ao longo do processo de inativação, totalizando 27 observações (3 replicatas $\times$ 9 tempos).

Para validação do modelo, foram utilizados três conjuntos adicionais de dados, cada um contendo 60 replicações independentes de inativação bacteriana com concentrações iniciais reduzidas: aproximadamente 850 células (10^3), 90 células (10^2) e 8 células (10^1). Estes datasets de validação permitem avaliar a capacidade do modelo de capturar a variabilidade estocástica observada em concentrações celulares baixas, onde a heterogeneidade individual das células desempenha papel preponderante no comportamento de inativação.

Os dados estão organizados em quatro arquivos CSV: `bayesian_fitting.csv` (dados de treinamento com 10^5 células), `850cell.csv`, `90cell.csv` e `8cell.csv` (datasets de validação). Cada arquivo contém as colunas de tempo de inativação (minutos) e o número de células sobreviventes observado.

2.2 Modelo

Para análise dos dados de inativação térmica, foi aplicado um Modelo Linear Generalizado (GLM) Bayesiano em substituição ao método convencional de mínimos quadrados. Diferentemente dos métodos frequentistas que assumem distribuição normal para os erros, o GLM Bayesiano permite incorporar distribuições de probabilidade discretas apropriadas para dados de contagem, além de quantificar explicitamente a incerteza paramétrica Hiura et al. (2021).

2.2.1 Modelo Cinético de Weibull

O modelo cinético de Weibull tem sido amplamente utilizado para descrever curvas de inativação não-lineares em microbiologia preditiva. A forma logarítmica do modelo é expressa por:

$$\log_{10}\left(\frac{N_t}{N_0}\right) = -\left(\frac{t}{\delta}\right)^p \tag{1}$$

onde t representa o tempo de inativação (min), N_t é a população bacteriana no tempo t, N_0 é o número inicial de células, δ é o parâmetro de escala (scale parameter) e p é o parâmetro de forma (shape parameter). Rearranjando a Equação 1, obtém-se a forma não-logarítmica:

$$N_t = N_0 \times 10^{-\left(\frac{t}{\delta}\right)^p} \tag{2}$$

2.2.2 Modelo Linear Generalizado com Distribuição de Poisson

No método de mínimos quadrados tradicional, assume-se que os erros no número logarítmico de células seguem uma distribuição normal. No entanto, como o número de células bacterianas é um dado de contagem discreto, a distribuição de Poisson é mais apropriada para descrever a variabilidade observada.

Assumindo que: (i) o número inicial de células segue uma distribuição de Poisson; (ii) a taxa de inativação é igual para todas as células; e (iii) os eventos de inativação são independentes entre si, o número de células sobreviventes também seguirá uma distribuição de Poisson. Portanto, o número observado de bactérias sobreviventes no tempo t pode ser descrito como:

$$N_t \sim \text{Poisson}\left(\lambda_t\right) \tag{3}$$

onde o parâmetro λ_t (taxa média da distribuição de Poisson) é dado pela Equação 2:

$$\lambda_t = N_0 \times 10^{-\left(\frac{t}{\delta}\right)^p} \tag{4}$$

Neste modelo, os parâmetros δ, p e N_0 foram estimados a partir dos dados observados de tempo de aquecimento (t) e número de células sobreviventes (N_t). A variável aleatória do número de células é distribuída segundo Poisson, o que corresponde ao resíduo do modelo na variável dependente do GLM.

2.2.3 Inferência Bayesiana

A inferência bayesiana permite que os parâmetros do modelo sejam representados como distribuições de probabilidade em vez de estimativas pontuais. Neste estudo, foram utilizadas distribuições a priori não-informativas (uniformes) para os parâmetros, visto que não havia informação prévia disponível. A estimação foi realizada através de Markov Chain Monte Carlo (MCMC) utilizando o software Stan, conforme implementação de Hiura et al. (2021).

2.2.4 Computação

A estimação dos parâmetros foi realizada utilizando inferência bayesiana via algoritmo Markov Chain Monte Carlo (MCMC) implementado no software Stan. Na inferência bayesiana, os dados observados são considerados como gerados a partir de uma distribuição de probabilidade, e todos os parâmetros são estimados como distribuições de probabilidade. Como não havia informação prévia disponível, foram utilizadas distribuições uniformes como priors não-informativos para todos os parâmetros (δ, p e N_0).

Para cada modelo, a inferência foi conduzida com 10.000 iterações em quatro cadeias MCMC independentes. As primeiras 5.000 iterações foram descartadas como período de aquecimento (*warm-up*), e as 5.000 iterações restantes foram utilizadas para estimação dos parâmetros posteriores. Parâmetros de controle do algoritmo NUTS (*No-U-Turn Sampler*) foram ajustados para `adapt_delta = 0.95` e `max_treedepth = 15` para garantir convergência adequada e minimizar divergências na amostragem.

A convergência foi verificada através da inspeção visual dos traços das cadeias MCMC e do diagnóstico de Gelman-Rubin (R-hat). Valores de R-hat próximos a 1.0 indicam convergência satisfatória das cadeias. Adicionalmente, foram gerados gráficos de diagnóstico das distribuições posteriores e correlações entre parâmetros.

2.3 Simulação

Após a estimação dos parâmetros via inferência bayesiana, foram obtidos 2×10^4 conjuntos de pares de parâmetros (δ, p) das amostras posteriores (5.000 iterações $\times$ 4 cadeias). Estes conjuntos foram utilizados para simular o comportamento de inativação com concentrações iniciais reduzidas de aproximadamente 10^n (n = 1, 2, 3) células, correspondendo aos valores médios de 8, 90 e 850 células, respectivamente.

Para cada concentração inicial e cada conjunto de parâmetros amostrados da distribuição posterior, o procedimento de simulação foi conduzido da seguinte forma: (i) o número inicial de células (N_0) foi amostrado de uma distribuição de Poisson com média correspondente à concentração desejada; (ii) para cada ponto temporal específico presente nos datasets de validação, o número de células sobreviventes foi calculado utilizando a Equação 4 com os parâmetros δ e p amostrados; (iii) o número observado de células foi então amostrado de uma distribuição de Poisson com média λ_t. Este procedimento foi repetido para todos os tempos observados experimentalmente em cada dataset de validação, resultando em 20.000 trajetórias simuladas de inativação para cada concentração inicial.

Os intervalos de predição de 95% foram construídos ordenando os 2×10^4 resultados simulados em cada ponto temporal e identificando os valores correspondentes aos percentis 2.5% (limite inferior) e 97.5% (limite superior). As 60 observações experimentais de cada concentração foram então comparadas com estes intervalos de predição.

2.3.1 Avaliação da Acurácia

O procedimento de avaliação das predições foi baseado em metodologia previamente descrita em Hiura et al. (2020). Para cada ponto temporal, os 2×10^4 resultados de predição foram ordenados em ordem crescente. Uma observação experimental foi considerada dentro do intervalo de predição se o valor de contagem de colônias observado fosse maior que a predição correspondente ao percentil 2.5% e menor que a predição correspondente ao percentil 97.5%. A acurácia foi calculada como a razão entre o número de observações dentro do intervalo de predição de 95% e o total de 60 observações para cada concentração:

$$\text{Acurácia (\%)} = \frac{\text{N° observações dentro do intervalo}}{\text{Total de observações}} \times 100 \tag{5}$$

3 Resultados

A Figura 1 apresenta os traços das quatro cadeias MCMC independentes para os parâmetros δ, p e N_0. As cadeias demonstram boa convergência e mistura adequada, sem evidências. Os valores de R-hat foram inferiores a 1.01 para todos os parâmetros, confirmando convergência satisfatória.

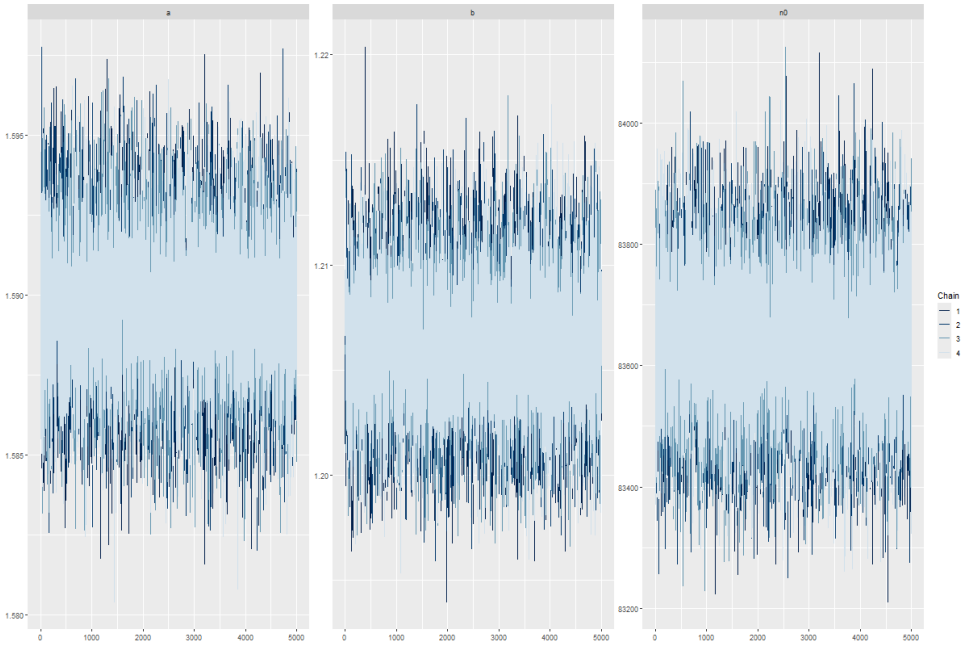


Figura 1: Traços das cadeias MCMC para os três parâmetros estimados (δ, p e N_0). As quatro cadeias independentes demonstram boa convergência e mistura adequada.

A Figura 2 apresenta as distribuições posteriores dos três parâmetros do modelo: δ (parâmetro de escala), p (parâmetro de forma) e N_0 (concentração inicial). Utilizando priors uniformes (não-informativos) para todos os parâmetros, a inferência bayesiana resultou em distribuições posteriores aproximadamente simétricas e unimodais, indicando que os dados observados contêm informação suficiente para estimação precisa dos parâmetros.

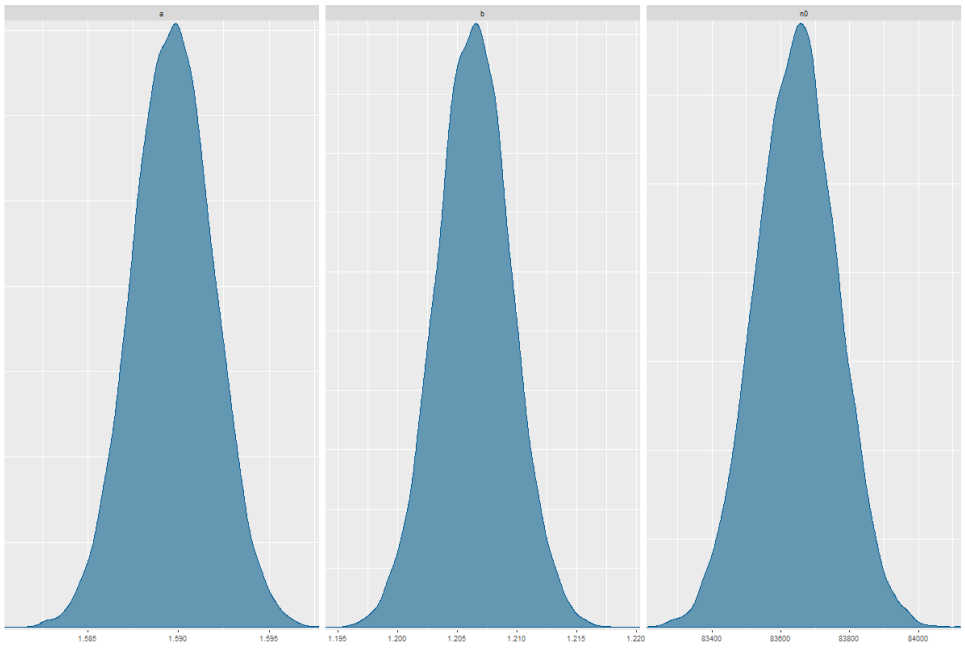


Figura 2: Distribuições posteriores dos parâmetros estimados: (a) δ, (b) p e (c) N_0.

A Tabela 1 sumariza as estimativas posteriores com seus respectivos intervalos de credibilidade de 95%. O parâmetro de escala δ foi estimado em 1.590 min (IC 95%: [1.585, 1.594]), o parâmetro de forma p em 1.206 (IC 95%: [1.200, 1.213]), e a concentração inicial N_0 em 83645 células (IC 95%: [83414, 83874]). Os valores de $\hat{R}$ (estatística de Gelman-Rubin) iguais a 1.0 para todos os parâmetros confirmam excelente convergência das cadeias MCMC. As distribuições posteriores vistas na Figura 2 refletem baixa incerteza paramétrica, resultado da combinação entre o modelo apropriado (GLM com distribuição de Poisson).

Tabela 1: Parâmetros estimados do modelo de inativação através de inferência bayesiana MCMC.

Parâmetro	**Média**	**Desvio Padrão**	**Q2.5%**	**Q97.5%**	**IC 95%**	$\hat{R}$
δ	1.590	0.002	1.585	1.594	[1.585, 1.594]	1
p	1.206	0.003	1.200	1.213	[1.200, 1.213]	1
N_0	83645	118	83414	83874	[83414, 83874]	1

A Figura 3 apresenta os gráficos de dispersão das amostras posteriores dos três parâmetros, juntamente com suas distribuições marginais na diagonal. Os padrões de dispersão aproximadamente circulares indicam baixa correlação entre os pares de parâmetros (δ-p, δ-N_0 e p-N_0), sugerindo boa identificabilidade do modelo. A ausência de correlações fortes entre δ, p e N_0 indica que estes podem ser estimados de forma independente, o que favorece tanto a interpretabilidade do modelo quanto a eficiência do algoritmo MCMC.

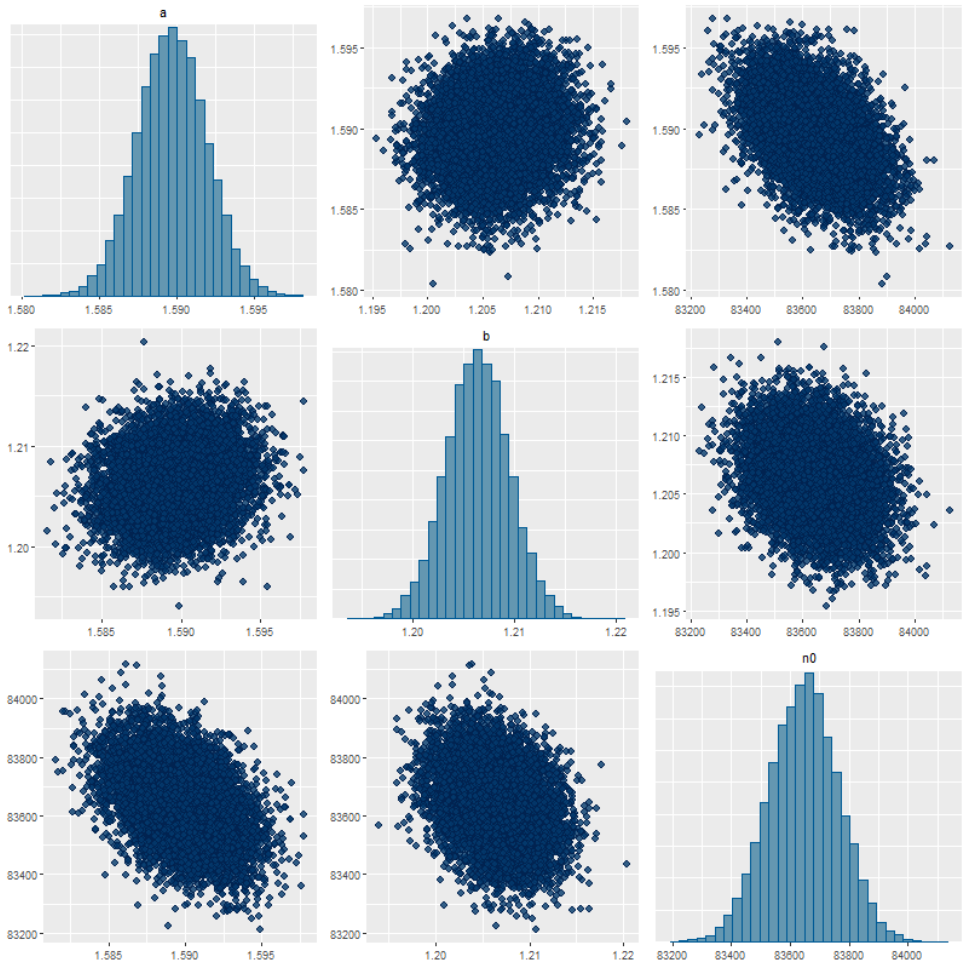


Figura 3: Correlações entre os parâmetros estimados δ, p e N_0. As distribuições marginais (diagonal) e os gráficos de dispersão revelam baixa correlação entre os parâmetros.

A Figura 4 apresenta o ajuste do modelo aos dados experimentais obtidos com concentração inicial de 10^5 células em três replicatas. A linha azul representa a curva mediana de inativação (a mediana e a média são iguais), enquanto a área sombreada corresponde ao intervalo de credibilidade de 95%. O modelo demonstra excelente ajuste, com a maioria dos pontos experimentais localizados dentro do intervalo de credibilidade. O eixo vertical em escala logarítmica representa o log10 das unidades formadoras de colônias (CFU, *colony forming units*), permitindo visualizar o decaimento de aproximadamente cinco ordens de magnitude (de 10^5 a menos de 10^1 CFU) ao longo dos 5 minutos de tratamento térmico a 94°C. O ajuste satisfatório evidencia que o modelo GLM Bayesiano captura adequadamente a cinética de inativação térmica de *Bacillus simplex*.

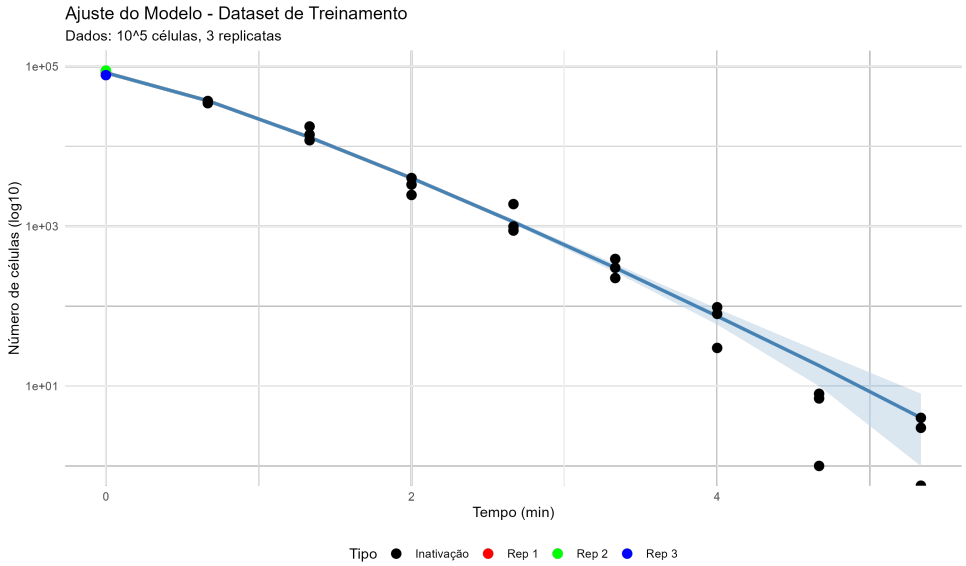
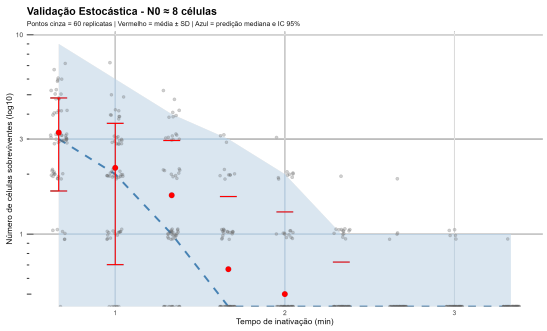
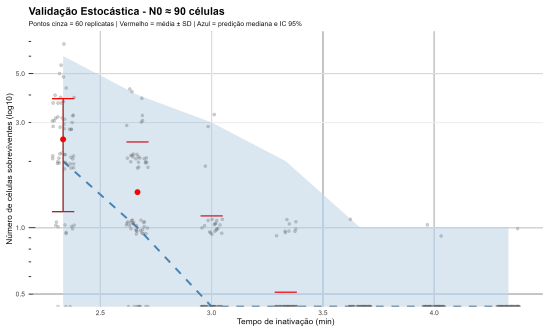


Figura 4: Ajuste do modelo GLM Bayesiano aos dados de treinamento (10^5 células, 3 replicatas). A linha representa a curva de inativação estimada e a área sombreada o intervalo de credibilidade de 95%.

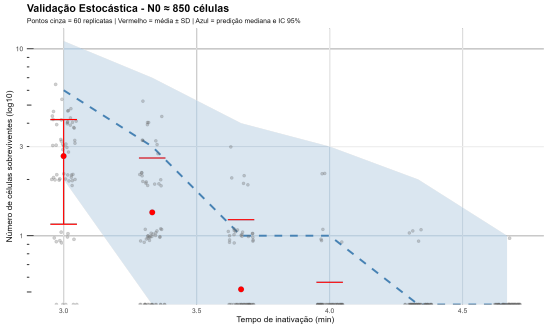
O modelo foi validado através de simulações estocásticas com concentrações iniciais reduzidas. A Figura 5 apresenta as comparações entre as predições do modelo e os dados experimentais para três concentrações: aproximadamente 8 (10^1), 90 (10^2) e 850 (10^3) células, cada uma com 60 replicações independentes Abe et al. (2019). Para cada concentração, foram realizadas 20.000 simulações utilizando os parâmetros das distribuições posteriores, e os intervalos de predição de 95% foram calculados a partir dos percentis 2.5% e 97.5%.



(a) 8 células (10^1)



(b) 90 células (10^2)



(c) 850 células (10^3)

Figura 5: Validação do modelo GLM Bayesiano através de simulações estocásticas com diferentes concentrações iniciais. As linhas representam os intervalos de predição de 95%, e os pontos indicam as 60 observações experimentais para cada concentração. Acurácia: (a) 98.9%, (b) 99.8%, (c) 96.4%.

A Tabela 2 apresenta as acurácias de validação. O modelo demonstrou excelente capacidade preditiva, com acurácia geral de 98.4% (1300 de 1320 observações dentro do intervalo de predição de 95%). As acurácias individuais foram de 98.9% para 8 células, 99.8% para 90 células e 96.4% para 850 células. Em concentrações muito baixas, a variabilidade estocástica é pronunciada, como evidenciado pela ampla dispersão das observações na Figura 5a, e o modelo capturou adequadamente essa variabilidade. Estes resultados confirmam que o modelo GLM Bayesiano com distribuição de Poisson descreve apropriadamente tanto a cinética de inativação determinística quanto a variabilidade estocástica em baixas concentrações celulares.

Tabela 2: Acurácia de predição do modelo para diferentes concentrações iniciais.

Concentração Inicial	**Acurácia (%)**	**N predito/Total**
8 células (10^1)	98.9	534/540
90 células (10^2)	99.8	419/420
850 células (10^3)	96.4	347/360
Total	**98.4**	**1300/1320**

4 Discussão

Este estudo aplicou a metodologia GLM Bayesiana proposta por Hiura et al. (2021) aos dados de inativação térmica de *Bacillus simplex* publicados por Abe et al. (2019), demonstrando a viabilidade e eficácia dessa abordagem para modelagem de processos de inativação microbiana. Os resultados obtidos reforçam a adequação da distribuição de Poisson para descrever dados de contagem bacteriana, superando limitações dos métodos tradicionais baseados em mínimos quadrados e distribuição normal.

O modelo GLM Bayesiano com distribuição de Poisson demonstrou excelente desempenho tanto no ajuste aos dados de treinamento quanto na validação com concentrações reduzidas. A convergência perfeita das cadeias MCMC, evidenciada pelos valores de $\hat{R} = 1.0$ para todos os parâmetros, confirma a robustez do procedimento de estimação. As distribuições posteriores estreitas e aproximadamente simétricas indicam que os dados experimentais de alta concentração (10^5 células, 3 replicatas) forneceram informação suficiente para estimação dos parâmetros $\delta = 1.590$ e $p = 1.206$, mesmo utilizando priores uniformes não-informativas.

A baixa correlação entre os parâmetros δ, p e N_0 observada nos gráficos de pares demonstra boa identificabilidade do modelo, o que é crucial para interpretabilidade dos parâmetros e eficiência do algoritmo MCMC. Esta característica contrasta com situações onde alta correlação entre parâmetros pode indicar problemas de identificabilidade ou necessidade de reparametrização do modelo.

A principal vantagem da distribuição de Poisson em relação à distribuição normal tradicionalmente utilizada é sua capacidade de lidar naturalmente com dados de contagem discretos, incluindo observações com valores zero.

A validação do modelo com concentrações reduzidas (8, 90 e 850 células) revelou sua excelente capacidade de capturar a variabilidade estocástica observada experimentalmente. A acurácia geral de 98.4% indica que os intervalos de predição de 95% construídos através de 20.000 simulações Monte Carlo representam adequadamente a incerteza nas predições. Este resultado é notável considerando que o modelo foi ajustado exclusivamente com dados de alta concentração.

Os resultados deste estudo confirmam que o modelo GLM Bayesiano com distribuição de Poisson é uma ferramenta apropriada e eficiente para modelagem de processos de inativação microbiana. A metodologia

demonstrou excelente capacidade de ajuste aos dados de treinamento, convergência robusta das cadeias MCMC, boa identificabilidade de parâmetros e alta acurácia preditiva em concentrações reduzidas.

5 Códigos e Dataset Utilizado

As análises foram implementadas em linguagem R utilizando os pacotes `rstan` Team (2024) para interface com Stan, `tidyverse` Wickham (2023) para manipulação de dados e visualização, e `bayesplot` Gabry and Mahr (2024) para o diagnóstico das distribuições posteriores. O código completo e os datasets utilizados estão disponíveis no repositório GitHub: `https://github.com/udinh0/bayesian-inactivation-analysis`.

Referências

Abe, H., Koyama, K., Kawamura, S., and Koseki, S. (2019). Stochastic modeling of variability in survival behavior of *Bacillus simplex* spore population during isothermal inactivation at the single cell level using a monte carlo simulation. *Food Microbiology*, 82:436–444.

Gabry, J. and Mahr, T. (2024). *bayesplot: Plotting for Bayesian Models*. R package version 1.14.0.

Gil, M. M., Miller, F. A., Brandão, T. R. S., and Silva, C. L. M. (2017). Mathematical models for prediction of temperature effects on kinetic parameters of microorganisms' inactivation: tools for model comparison and adequacy in data fitting. *Food and Bioprocess Technology*, 10:2208–2225.

Hiura, S., Abe, H., Koyama, K., and Koseki, S. (2020). Transforming kinetic model into a stochastic inactivation model: statistical evaluation of stochastic inactivation of individual cells in a bacterial population. *Food Microbiology*, 91:103508.

Hiura, S., Abe, H., Koyama, K., and Koseki, S. (2021). Bayesian generalized linear model for simulating bacterial inactivation/growth considering variability and uncertainty. *Frontiers in Microbiology*, 12:674364.

Koyama, K., Aspridou, Z., Koseki, S., and Koutsoumanis, K. (2019). Describing uncertainty in *Salmonella* thermal inactivation using bayesian statistical modeling. *Frontiers in Microbiology*, 10:2239.

Lammerding, A. M. and Fazil, A. (2000). Hazard identification and exposure assessment for microbial food safety risk assessment. *International Journal of Food Microbiology*, 58:147–157.

Nauta, M. J. (2000). Separation of uncertainty and variability in quantitative microbial risk assessment models. *International Journal of Food Microbiology*, 57:9–18.

Pouillot, R., Albert, I., Cornu, M., and Denis, J. B. (2003). Estimation of uncertainty and variability in bacterial growth using bayesian inference. application to *Listeria monocytogenes*. *International Journal of Food Microbiology*, 81:87–104.

Team, S. D. (2024). *rstan: R Interface to Stan*. R package version 2.32.7.

van Boekel, M. A. J. S. (2020). On the pros and cons of bayesian kinetic modeling in food science. *Trends in Food Science & Technology*, 99:181–193.

Wickham, H. (2023). *tidyverse: Easily Install and Load the 'Tidyverse'*. R package version 2.0.0.