**ABG CONSULTORIA ESTATÍSTICA**

# Previsão de séries epidemiológicas incorporando atraso na notificação

Danielly Santos Severino

Orientador: Dani Gamerman

Co-Orientadore: Izabel Nolau

Universidade Federal de Minas Gerais

Instituto de Ciências Exatas

2023

# Resumo

A gestão do risco de doenças infecciosas pelos sistemas de vigilância é dificultada pela presença de atrasos na notificação dos casos relatados. O atraso na notificação ocorre devido a diversos fatores como confirmação laboratorial, dificuldades de estrutura e problemas logísticos. É crucial corrigir a informação disponível o mais rápido possível para auxiliar a tomada de decisões como emissões de aviso às autoridades locais e ao público local.

São propostos modelos hierárquicos bayesianos para dados de contagem como uma forma flexível de corrigir o atraso nas notificações e quantificar a incerteza associada. A modelagem proposta assume que os dados de contagem seguem uma distribuição de probabilidade Poisson incorporando componentes temporais referentes ao período de ocorrência e ao tempo de atraso na notificação. Essa abordagem é ilustrada em dados de casos de incidência de dengue no Rio de Janeiro, Brasil. Além disso, o método de Monte Carlo via cadeias de Markov (MCMC) foi utilizado como técnica de extração de amostras para aproximar as densidades a posteriori das quantidades de interesse e a operacionalização foi realizada via STAN e R (versão 4.3.1).

**Palavras-chaves**: Estatística, inferência Bayesiana, séries temporais, dados de contagem, atraso na notificação, *Nowcasting*, *Forecasting*, dengue.

# Agradecimentos

Algumas épocas da vida a gente trabalha e estuda tanto que o tempo com a família é raro, quando estamos em casa só queremos aproveitar. O engraçado e mais contraditório é que é tudo por eles. Espero estar fazendo o melhor e dar orgulho para quem amo.

Não reclamo porque eu vivo um sonho. Todos os dias posso acessar minha fonte de felicidade, quero ser forte como a Tia Lú (Vera Lúcia) em tempos difíceis. Agradeço por poder olhar nos olhos do meu irmão Lucas e entender melhor a vida, aprender a me dedicar mais. Sou grata pela minha família e amigos que incentivam minha jornada. Me sinto mais livre sempre que me aproximo mais das minhas raízes. Deus, obrigada por ter me tornado a continuação de um sonho do meu pai e da minha mãe, de todos que vieram antes de mim.

Aos meu orientador Dani Gamerman e minha co-orientadora Izabel Nolau, agradeço pelos esforços, dedicação, inspiração, apoio e paciência em me auxiliar a concretizar esse trabalho. Agradeço também aos professores que contribuíram para a minha formação, especialmente Cristina Eterovik e Helinton A. Lopes. Por fim, agradeço a Fundação Mendes Pimental (FUMP) por me proporcionar condições financeiras para cursar o início da graduação.

# Sumário

[Previsão de séries epidemiológicas incorporando atraso na notificação 1](#_Toc137149896)

[Resumo 2](#_Toc137149897)

[Sumário 3](#_Toc137149898)

[Sumário de Tabelas 3](#_Toc137149899)

[Sumário de Figuras 3](#_Toc137149900)

[Introdução 4](#_Toc137149901)

[Estrutura típica dos dados com atraso na notificação 5](#_Toc137149902)

[Metodologia 6](#_Toc137149903)

[Inferência bayesiana 6](#_Toc137149904)

[Estimação pontual e intervalar 6](#_Toc137149905)

[Previsão 7](#_Toc137149906)

[Modelagem 7](#_Toc137149907)

[Aplicação 8](#_Toc137149908)

[Descrição dos casos de dengue 8](#_Toc137149909)

[Resultados 8](#_Toc137149910)

[Estimativas com estrutura de atraso independente 8](#_Toc137149911)

[Estimativas com estrutura de atraso conjunta 8](#_Toc137149912)

[Comparação dos modelos 8](#_Toc137149913)

[Anexo 8](#_Toc137149914)

[Referências 9](#_Toc137149915)

# Sumário de Tabelas

[Tabela 1 - Estrutura da base de dados 5](#_Toc135805133)

# Sumário de Figuras

[Figura 1 - Número total de casos de dengue relatados ao longo do tempo 7](#_Toc135826468)

# 

# Introdução

O sistema de vigilância de doenças infecciosas possui algumas abordagens comuns que muitos países adotam (M. Steven, R. Elliott, 2000). Na maioria dos países a notificação é obrigatória, exigindo que a incidência de doenças infecciosas como tuberculose, hepatite, HIV/AIDS e sarampo seja informada às autoridades de saúde. Os profissionais de saúde, hospitais e laboratórios são responsáveis por relatar os casos suspeitos ou confirmados. Portanto, existe um sistema de notificação e coleta de dados que se caracteriza por estabelecer informações essenciais para monitorar a disseminação e identificar tendências das doenças. Cabe destacar que os sistemas de vigilância epidemiológica ajudam a identificar surtos, avaliar a eficácia das medidas de controle e orientar políticas públicas de saúde. Além disso, a pandemia de COVID-19 trouxe à tona a importância da vigilância de doenças e a necessidade de fortalecer os sistemas de saúde em todo o mundo.

Os casos de doenças infecciosas que são relatados aos profissionais de saúde podem ser analisados através de contagens, sendo que, o número de casos comumente sofre atraso na notificação. Existem diversas situações que podem levar ao atraso na notificação de doenças como tempo até a confirmação do diagnóstico, desafios logísticos, confirmação de exames e barreiras culturais. Dessa forma, os casos notificados distorcem a relação entre a incidência da doença relatada e a verdadeira incidência da doença.

O atraso na notificação de doenças infecciosas pode ter consequências significativas para a saúde pública e para a eficácia de medidas de controle. É extremamente importante corrigir o atraso nas notificações para diminuir o risco de propagação da doença, não retardar o tratamento e cuidados adequados e nem dificultar a detecção de tendências e padrões. Sem informações adequadas da gravidade da situação atual em relação à doença as autoridades têm dificuldade em agir de forma eficaz para identificar e responder rapidamente aos surtos, permitindo que a infecção entre a população aumente.

Para diversos casos de doenças é comum que os dados sejam potencialmente subnotificados, ou seja, casos de doenças que não foram detectados ou que foram detectados, mas não notificados. Isso resulta em uma representação imprecisa da verdadeira contagem de doenças. Corrigir a subnotificação de doenças é um desafio complexo que requer uma abordagem abrangente e a implementação de várias estratégias, exigindo fontes adicionais de informação como conhecimento prévio da taxa de subnotificação.

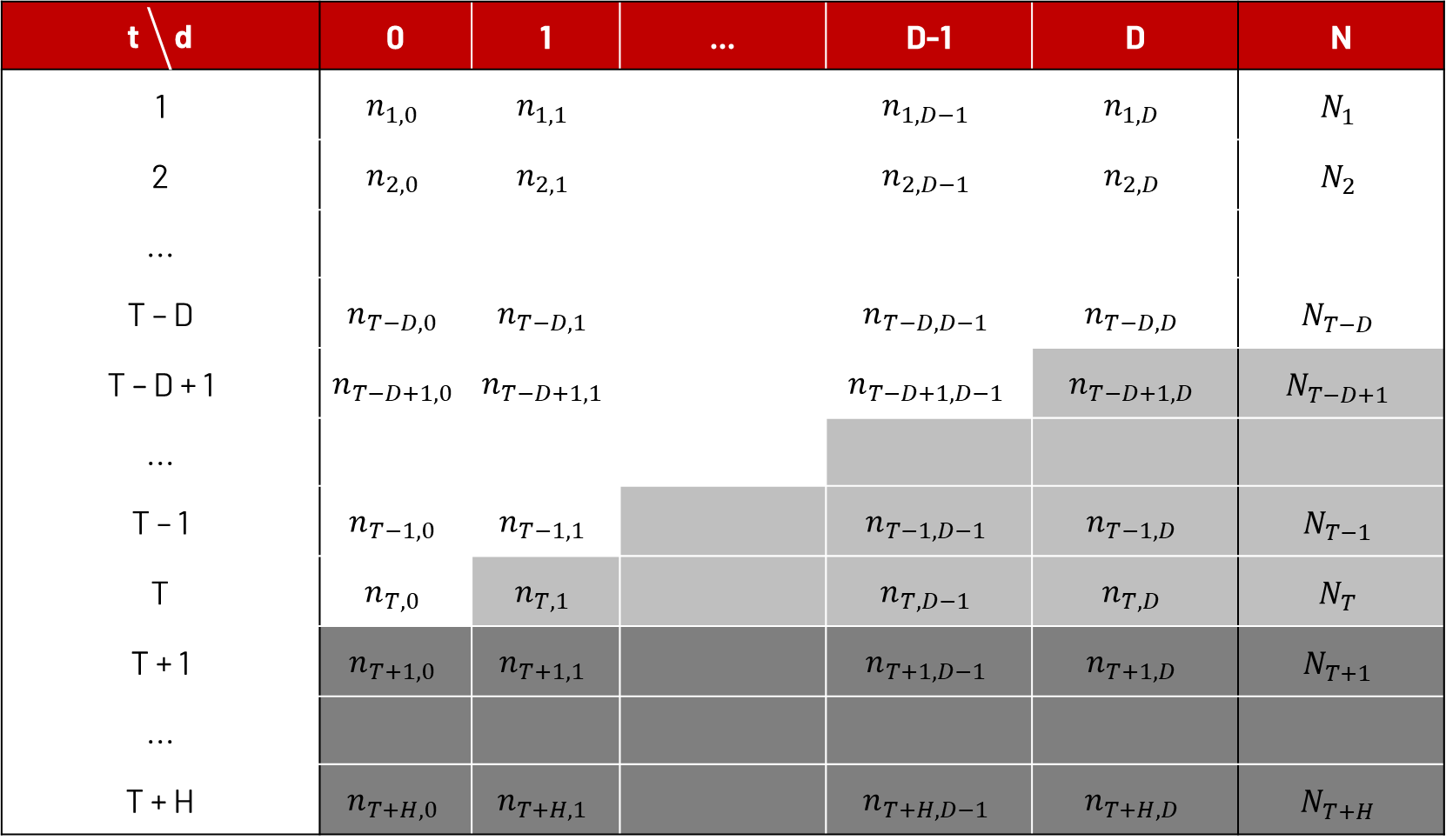
O foco neste trabalho é corrigir o atraso na notificação dos dados. Para isso, são propostos modelos hierárquicos bayesianos para dados de contagem como uma forma flexível de corrigir o atraso nas notificações e quantificar a incerteza associada. A modelagem proposta assume que os dados de contagem seguem uma distribuição de probabilidade Poisson incorporando componentes temporais referentes ao período de ocorrência e ao tempo de atraso na notificação.

# Estrutura típica dos dados com atraso na notificação

Existe uma estrutura típica que é utilizada para lidar com dados de contagem com problema de atraso na notificação e através dessa estrutura é possível capturar e analisar adequadamente as informações disponíveis (Bastos et al, 2019). A Tabela 1 - exibe a estrutura típica de uma base de dados qualquer com atraso na notificação. As linhas da tabela correspondem ao tempo , enquanto as colunas indicam o tempo de atraso na notificação, onde é o tempo atual, é o tempo máximo de previsão e é o tempo de atraso máximo relevante. Fixado qualquer intervalo de tempo (linha), a quantidade total de casos notificados é , sendo que é o número de eventos ocorridos no tempo registrados após unidades de tempo.

Fixado o tempo atual existem eventos ocorridos que serão registrados no futuro. As contagens até o tempo são integralmente observadas (valores em branco na tabela), enquanto os valores de até são parcialmente conhecidos (valores em cinza claro) e os valores de até ainda irão ocorrer (valores em cinza escuro).

### Estrutura típica dos dados com atraso na notificação



# Metodologia

A distribuição e a densidade de probabilidades são conceitos fundamentais na teoria das probabilidades. Ambas representam conceitos semelhantes, mas tem diferenças importantes. A distribuição de probabilidade caracteriza eventos, dados e variáveis aleatórias discretos. Já a função de densidade de probabilidade descreve valores contínuos. Apesar das análises e definições utilizadas serem válidas para distribuições e funções de densidade de probabilidade, as descrições foram apresentadas utilizando o conceito de distribuição de probabilidade, uma vez que os dados de contagem são discretos.

## Inferência Bayesiana

A inferência Bayesiana (G. Andrew et al, 2021) é uma abordagem estatística utilizada para fazer inferência sobre parâmetros desconhecidos ou eventos futuros com base em evidências observadas e conhecimento prévio. A inferência bayesiana parte do princípio de que as probabilidades são representações subjetivas de incerteza. Ela permite a inclusão de informações subjetivas e conhecimento especializado sobre o problema analisado através da especificação de uma distribuição ou densidade a priori. A inferência bayesiana envolve três elementos fundamentais para obter estimativas e realizar previsões: uma distribuição a priori; uma função de verossimilhança; e uma distribuição a posteriori.

Uma distribuição a priori, , é uma expressão probabilística que representa o conhecimento prévio ou falta de conhecimento sobre o problema. A função de verossimilhança, , é uma medida estatística que representa a relação entre os dados observados com os diferentes valores possíveis dos parâmetros desconhecidos. Por sua vez, a distribuição a posteriori, , é uma distribuição de probabilidade atualizada após a observação dos dados, obtida através da combinação da distribuição a priori e função de verossimilhança usando o teorema de Bayes:

A inferência Bayesiana é particularmente útil quando existem informações prévias relevantes ou quando a incerteza precisa ser modelada e atualizada à medida que novos dados são disponibilizados. Além disso, essa abordagem é aplicada em diversas áreas como ciências da saúde, ciências sociais, engenharia, economia e aprendizado de máquina.

Estimação e previsão

Sob o contexto da inferência Bayesiana, a estimação pontual é o processo de obter um único valor para estimar o parâmetro desconhecido a partir da distribuição a posteriori. Há diferentes métricas utilizadas para sintetizar a distribuição a posteriori e obter uma estimativa pontual como calcular o valor esperado (média), moda e mediana.

A estimação intervalar é utilizada para quantificar a incerteza em relação ao parâmetro desconhecido com a construção de intervalos de credibilidade. Os intervalos de credibilidade fornecem uma medida da probabilidade de o parâmetro estar dentro do intervalo, sendo que, o intervalo resultante representa uma faixa de valores mais prováveis para o parâmetro desconhecido considerando as informações disponíveis nos dados observados e conhecimento prévio do problema.

O intervalo HPD (*Highest Posterior Density*) se caracteriza por ser o intervalo de credibilidade mais estreito que contém uma proporção específica baseado na distribuição a posteriori. Esse intervalo (C. Ming-Hui, S. Qi-Man, 1988) não depende de pressupostos assintóticos para a sua construção. Além do mais, o intervalo pode ser assimétrico, acomodando distribuições a posteriori assimétricas. Em rigor, o procedimento para calcular o HPD envolve obter a aproximação da distribuição a posteriori, ordenar os valores, calcular a distribuição de probabilidade acumulada, e determinar a faixa mais estreita que contém a proporção desejada da aproximação da distribuição a posteriori.

Dentro do paradigma bayesiano, a previsão permite gerar estimativas pontuais e intervalares sobre eventos futuros com base nas informações disponíveis da distribuição a posteriori. Para realizar previsões deve-se gerar amostragens da distribuição a posteriori para obter uma amostra de possíveis valores dos parâmetros. A partir da amostra dos valores dos parâmetros aplica-se o modelo probabilístico para gerar valores preditos com base nos parâmetros amostrados. Depois disso são feitas as análises dos resultados das previsões.

Nowcasting

Forecasting

Computação

É possível obter uma amostra de possíveis valores dos parâmetros usando a técnica de amostragem de *Markov Chain Monte Carlo* (MCMC). Essa é uma técnica computacional comumente usada em inferência bayesiana para realizar amostragem de distribuições a posteriori que podem ser muito complexas. Além do mais, o MCMC é muito útil quando a distribuição a posteriori não pode ser calculada diretamente ou quando não há uma solução analítica disponível.

O MCMC permite gerar uma sequência de amostras que, em equilíbrio, representa uma aproximação da distribuição desejada. O processo de amostragem da distribuição a posteriori usando MCMC envolve a construção de uma cadeia de Markov irredutível e aperiódica que tem a distribuição estacionária equivalente a distribuição a posteriori. A cadeia de Markov é construída por meio de uma série de iterações que consiste em propor um novo estado, geralmente por meio de amostragem aleatória baseada no estado atual, e decidir sobre a aceitação ou rejeição desse novo estado. Em rigor, um novo estado proposto pode ser aceito ou rejeitado com base em uma regra de aceitação, sendo que, essa regra geralmente envolve calcular a razão de verossimilhança e a probabilidade de aceitar o novo estado é determinada com base nessa razão.

As técnicas de MCMC geralmente usadas incluem o Algoritmo de Metropolis-Hastings e a Amostragem de Gibbs. Ambas as técnicas fornecem uma forma de gerar amostras da distribuição a posteriori e obter estimativas robustas para os parâmetros desconhecidos.

Cabe destacar que a previsão na inferência bayesiana é flexível pois permite incorporar informações a priori, atualizar o conhecimento a partir dos dados observados e fornecer estimativas probabilísticas completas. Além disso, ela é amplamente aplicada em diversas áreas para previsões do tempo, previsões de eventos raros e previsão de séries temporais.

Convergencia

## Detalhamento da modelagem

Sob o paradigma bayesiano, os modelos hierárquicos são uma abordagem utilizada para lidar com a incerteza e a variabilidade em vários níveis dos dados. Esses modelos podem ser especialmente úteis quando há uma estrutura hierárquica ou agrupamento dos dados, permitindo que as informações de cada nível sejam compartilhadas e combinadas para obter estimativas mais precisas dos parâmetros desconhecidos.

Em um modelo hierárquico, os dados são organizados em vários níveis, sendo que cada nível do modelo possui seus próprios parâmetros e a incerteza sobre esses parâmetros é modelada usando distribuições a priori. A principal ideia de um modelo hierárquico é que as distribuições a priori de cada nível são influenciadas pelas informações dos níveis superiores.

Os modelos hierárquicos podem fornecer estimativas mais precisas e robustas, principalmente quando os dados em cada nível são limitados ou possuem ruídos. Além do mais, a estrutura hierárquica possibilita que os níveis mais baixos sejam regularizados pelos dados nos níveis superiores. Portanto, modelos hierárquicos aderem uma modelagem mais realista em relação à incerteza e variabilidade em dados complexos.

Especificações dos modelos

Foram propostos modelos hierárquicos com duas estruturas diferentes para lidar com dados de contagem incorporando o atraso na notificação. Ambos os modelos permitem mudanças suaves na variação temporal do número total de casos e no mecanismo de atraso. Cabe destacar que dados de contagem podem ser modelados através de múltiplas distribuições de probabilidade discretas, como a distribuição Poisson.

Ressalta-se que é a quantidade de casos ocorridos no tempo que foram registrados após unidades de tempo, onde é o último tempo para o qual os dados estão disponíveis e é o atraso máximo relevante.

Existe uma configuração de dependência entre os dados uma vez que , ou seja, o número total de casos ocorridos no tempo é exatamente a soma do número de casos ocorridos no tempo para todos os possíveis atrasos considerados. Portanto, não é possível especificar um modelo para e acatando todos os atrasos ao mesmo tempo respeitando a disposição dos dados. Salienta-se que são os casos notificados sem nenhum atraso e, por isso, pode ser mais interessante especificar modelos para e com e .

Bastos et al (2019) apresenta com mais detalhes os modelos bayesianos hierárquicos que foram utilizados. Na primeira abordagem a estrutura de atraso na notificação dos dados é independente, sendo assim, foram estimados os parâmetros para cada tempo de atraso. Enquanto a segunda abordagem abrange uma estrutura conjunta para os atrasos na notificação.

**Modelo com estrutura de atraso na notificação independente**

Modelo proposto para os dados:

Para e .

Onde,

Distribuições a priori:

**Modelo com estrutura de atraso na notificação conjunta**

Modelo proposto para os dados:

Para e .

Onde,

Distribuições a priori:

Comparação de modelos

Comparar os resultados de diferentes modelos é uma etapa importante para selecionar o modelo que melhor se adequou aos dados e ao problema determinado. O desempenho dos modelos pode ser avaliado através de múltiplas métricas e técnicas, dentre elas, o MAE (*Mean Absolute Error*) e o RMSE (*Root Mean Squared Error*). Ambos foram utilizados para avaliar os erros de estimação e previsão, comparando as estimativas e previsões geradas pelos modelos com os valores reais observados.

O MAE é uma métrica calculada como a média das diferenças absolutas entre as previsões/estimações e os valores reais observados. Ele fornece uma medida média do erro absoluto sem considerar a direção do erro. Já o RMSE calcula a raiz quadrada da média dos erros quadrados dos valores previstos e quantidades observadas. Ele fornece uma medida da raiz quadrada do erro quadrático e é sensível a erros maiores. Além do mais, ambas as métricas são úteis para comparar diferentes modelos.

O MAE e o RMSE são calculados através das seguintes fórmulas:

Onde, corresponde ao total de valores comparados.

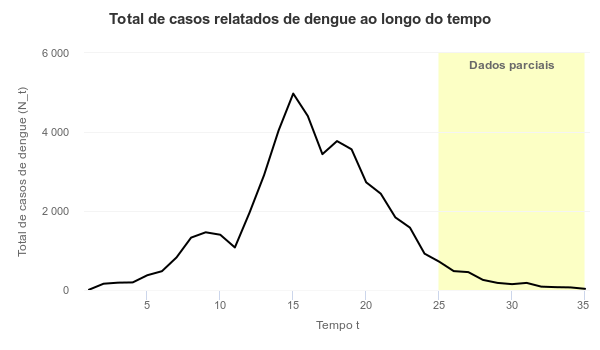
# Aplicação

## Descrição dos casos de dengue

A aplicação dos modelos foi realizada utilizando séries temporais de casos notificados de dengue no Rio de Janeiro entre janeiro de 2011 e abril de 2012. A estrutura dos dados foi organizada semanalmente para 35 semanas com no máximo 10 semanas de atraso na notificação. Dessa forma, as contagens foram integralmente observadas até 25 semanas e, a partir de então as contagens eram parciais.

A Figura 1 - ilustra o total de casos relatados de dengue no decorrer das semanas. Além disso, observa-se que houve um pico de casos de dengue entre a 12ª e 22ª semanas.

##### Número total de casos de dengue relatados ao longo do tempo



Resultados do modelo com estrutura de atraso independente

Resultados do modelo com estrutura de atraso conjunta

Comparação dos modelos

**Erros de estimação e previsão**

# Conclusões

# Anexo

Convergência dos parâmetros

# Referências

**Livro vigilância de doenças:**

Principles and Practice of Public Health Surveillance. Steven M. 2000.

<https://books.google.com.br/books?hl=pt-BR&lr=&id=R1n5Yrcld1UC&oi=fnd&pg=PR9&dq=Principles+and+Practice+of+Public+Health+Surveillance&ots=ShKMs5objB&sig=3ZV3iCDDsMlZ2EsFBBOPpD0PRgQ#v=onepage&q=Principles%20and%20Practice%20of%20Public%20Health%20Surveillance&f=false>

**Atraso na notificação:**

Dorigatti, I., Cauchemez, S., Pugliese, A., & Ferguson, N. M. (2011). "Characterizing the epidemiology of the 2009 influenza A/H1N1 pandemic in Mexico."

https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1000436

**Livro Inferencia bayesiana:**

Andrew Gelman, John Carlin, Hal Stern, David Dunson, Aki Vehtari, and Donald Rubin. Bayesian Data Analysis. 3 ed, 2013. <http://www.stat.columbia.edu/~gelman/book/>

**Artigo MCMC:**

Galin L. Jones and Qian Qin. Markov Chain Monte Carlo in Practice. 2021

<https://www.annualreviews.org/doi/abs/10.1146/annurev-statistics-040220-090158>

**LIVRO MCMC**: (so uma parte gratuita)

W.R. Gilks, S. Richardson e D.J. Spiegelhalter. Markov Chain Monte Carlo in Practice. 1996

**MAE RMSE:**

Shmueli, G. (2010). "To explain or to predict?"

https://projecteuclid.org/journals/statistical-science/volume-25/issue-3/To-Explain-or-to-Predict/10.1214/10-STS330.full