UNIVERSIDADE FEDERAL DE MINAS GERAIS

INSTITUTO DE CIÊNCIAS EXATAS

DEPARTAMENTO DE ESTATÍSTICA

ESTATÍSTICA

DANIELLY SANTOS SEVERINO

**PREVISÃO DE SÉRIES EPIDEMIOLÓGICAS INCORPORANDO ATRASO NA NOTIFICAÇÃO**

BELO HORIZONTE

2023

DANIELLY SANTOS SEVERINO

**PREVISÃO DE SÉRIES EPIDEMIOLÓGICAS INCORPORANDO ATRASO NA NOTIFICAÇÃO**

Monografia apresentada ao Departamento de Estatística da Universidade Federal de Minas Gerais (UFMG), como parte dos requisitos necessários à obtenção do título de bacharel em Estatística.

Orientador: Dani Gamerman

Coorientadora: Izabel Nolau

BELO HORIZONTE

2023

**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. 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 coorientadora 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.

**Resumo**

A gestão do risco de doenças infecciosas envolve a notificação e coleta de dados para monitorar a disseminação e identificar tendências das doenças. No entanto, os casos notificados muitas vezes sofrem atraso na notificação, ou seja, os casos são notificados após a incidência da doença. Os atrasos distorcem a relação entre a incidência relatada e a verdadeira incidência da doença. Isso pode ter consequências significativas para a saúde pública e para a eficácia de medidas de medidas de controle. Sem informações adequadas da gravidade da situação atual em relação à doença, as autoridades têm dificuldade em agir de maneira eficaz para identificar e responder adequadamente aos surtos. Sendo assim, 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-alvo.

Os modelos hierárquicos Bayesianos são utilizados para lidar com a incerteza em vários níveis dos dados, permitindo combinar informações de diferentes níveis para obter estimativas e previsões mais precisas e robustas. Ao lidar com doenças infecciosas é importante que os modelos sejam capazes de gerar estimativas que recuperem a informação existente com precisão e, especialmente, construir previsões robustas. As previsões podem ser feitas em tempo real (*nowcasting*) com base nos dados disponíveis ou para períodos futuros (*forecasting*) com base no histórico.

Neste trabalho, foram propostos modelos hierárquicos Bayesianos com duas estruturas diferentes para lidar com dados de contagem de incidência de doenças infecciosas e corrigir o atraso nas notificações. Esses modelos são uma forma flexível de corrigir os atrasos nas notificações e permitem gerar previsões e quantificar a incerteza associada. Os modelos propostos foram ilustrados usando dados semanais de incidência de dengue no Rio de Janeiro registrados entre janeiro de 2011 e abril de 2012. Além disso, a técnica de Monte Carlo via Cadeias de Markov (MCMC) foi utilizada para a extração de amostras para aproximar as densidades a posteriori das quantidades de interesse.

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

**Sumário**

**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, ou seja, casos que são relatados após a incidência da doença. 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, foram propostos modelos hierárquicos Bayesianos para dados de contagem de incidência das doenças infecciosas 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 do efeito do tempo e do efeito das defasagens.

**Estrutura típica dos dados com atraso na notificação**

Ao lidar com séries epidemiológicas, existe uma estrutura típica que é utilizada para lidar com dados de contagem da incidência de doenças com problema de atraso na notificação. 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 para contagens com atraso na notificação. As linhas da tabela correspondem ao tempo, enquanto as colunas indicam o período de atraso na notificação. é o tempo atual, é o atraso máximo relevante e é o horizonte máximo de previsão. Os eventos ocorridos no tempo com unidades de atraso são denotados por e, fixado qualquer intervalo de tempo (linha), a quantidade total de eventos é .

Os eventos ocorridos entre e foram integralmente registrados. Já os eventos ocorridos entre e são parcialmente registrados, pois existem eventos que ocorreram e não foram registrados devido ao atraso na notificação. Por fim, as contagens de até são os eventos futuros. Dessa forma, os valores em cinza claro na tabela são as contagens que já ocorreram, mas ainda não foram registradas e os valores em cinza escuro são as contagens futuras.

**Tabela 1 - Estrutura típica dos dados com atraso na notificação**

Tabela

Descrição gerada automaticamente

**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 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 de incidência de doenças 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 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.

**Estimações e previsões**

Sob o contexto da inferência Bayesiana, a estimação pontual é o processo de obter um único valor para estimar o vetor de parâmetros desconhecidos 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.

No cenário epidemiológico, a previsão de casos de incidência da doença é importante por diversas razões, dentre elas, o planejamento para alocação de recursos e a adoção de medidas de controle. 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.

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.

O fato de séries epidemiológicas apresentarem atrasos na notificação torna necessário a utilização de duas abordagens de previsão: *Nowcasting* e *Forecasting*.O Nowcasting abrange as previsões em tempo (quase) real, cujo foco principal é obter estimativas precisas e atualizadas para o momento atual baseadas nos dados e informações disponíveis. Esse procedimento é particularmente útil para o problema, uma vez que, permite o monitoramento em tempo real das doenças. Além disso, as informações em tempo real são fundamentais para a tomada de decisões imediatas. Por sua vez, o Forecasting refere-se as previsões para os períodos futuros a partir do histórico disponível. Logo, trata-se de uma tentativa de prever o que pode acontecer no futuro considerando padrões e tendências observadas no passado.

**Computação**

A forma analítica exata das distribuições a posteriori dos modelos propostos é desconhecida. Isso comumente ocorre quando o modelo estatístico é complexo e envolve múltiplos parâmetros e/ou possuí uma estrutura hierárquica. Portanto, é necessário aproximar as distribuições a posteriori.

A técnica de amostragem de *Markov Chain Monte Carlo* (MCMC) é utilizada para obter uma amostra de possíveis valores dos parâmetros. 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 (GAMERMAN, D.; LOPES, H.F., 2006) 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.

A convergência no MCMC é atingida quando o algoritmo é capaz de gerar amostras que se aproximam da distribuição alvo desejada. É essencial garantir que as amostras geradas sejam representações da distribuição alvo, sejam não viesadas e não apresentem um erro amostral significativo. O método utilizado para verificar a convergência no MCMC foi o monitoramento dos traços das cadeias. Os traços das cadeias MCMC são gráficos que ilustram os valores dos parâmetros amostrados em função do tempo ou iteração. Os gráficos indicam convergência quando as cadeias parecem estar “vagando” em torno de um valor estável, sem tendências significativas ou presença de comportamento errático.

**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.