Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Estatística

Brenda de Souza Ribeiro

CONSTRUÇÃO DE INDICADORES SOB A ABORDAGEM BAYESIANA

Brasília 2021

Brenda de Souza Ribeiro

CONSTRUÇÃO DE INDICADORES SOB A ABORDAGEM BAYESIANA

Relatório apresentado à disciplina TCC II do curso de graduação em Estatística, Departamento de Estatística, Instituto de Exatas, Universidade de Brasília, como parte dos requisitos necessários para o grau de Bacharel em Estatística.

Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Estatística

Orientador: Leandro Correia

Brasília
2021

Agradecimentos

O desenvolvimento desse trabalho de conclusão de curso contou com a ajuda de diversas pessoas, dentre as quais eu agradeço:

Primeiramente eu agradeço a Deus me cobriu com a sua graça me ajudando em todo o processo.

Ao meu orientador Leandro Correia que esteve ao meu lado em todo os momentos, me dando todo o auxilio necessário para elaboração deste trabalho.

Ao meu pai Arnaldo, meus irmãos Henrique e Lucas juntamente com minha cunhada Thayane e meu sobrinho Daniel, minha avó Cida e meu vô Alberto, sempre me apoiaram, e mesmo quando vinham ondas de desanimo eles me colocavam pra cima para não desistir.

Aos meus amigos, Karina, Davi, Claudinha, Kaio, Camila, Rayssa, Marcos sempre estiveram ao meu lado tanto fisicamente quando em orações.

CONSTRUÇÃO DE INDICADORES SOB A ABORDAGEM BAYESIANA

Brenda de Souza Ribeiro

2021

Resumo

O presente trabalho tem como objetivo, a analise dos indicadores sobre uma visão da estatística Bayesiana. Primeiro é apresentado conceitos sobre os indicadores e a importância do uso e do estudo mais detalhado deles. São mostradas as propriedades desejáveis dos indicadores. Nesse trabalho vai ser mais aprofundado as indicadores demográficos como a taxa de mortalidade e a expectativa de vida.

Como aplicação foi feita uma atualização do estudo de Schmertmann e Gonzaga (2018), "Bayesian estimation of age-specific mortality and life expectancy for smallareas with defective vital records", onde usam a inferência Bayesiana aplicando um modelo hierárquico Bayesiano para estimar as mortalidades e expectativas de vida em pequenas áreas do Brasil. Com a diferença que neste estudo são usados dados das Unidades da Federação, com projeções populacionais. Foram obtidos resultados significativos e inspiradores para o estudos futuros.

Palavras-chaves:Indicadores, Inferência Bayesiana, demografia, população, mortalidade, expectativa de vida.

Sumário

1 Introdução

Um Indicador Social é uma medida quantitativa dotada de um significado social fundamental, usado para substituir, quantificar ou operacionalizar um conceito social abstrato (Jannuzzi,2001). Os indicadores possuem uma grande relevância na atualidade, pois são cada vez mais utilizados para quantificar e explicar situações cotidianas. Em contextos acadêmicos, docentes e discentes usam para facilitar o aprendizado e também na formulação de pesquisas e artigos acadêmicos. Nas políticas públicas estabelecidas pelo governo utiliza-se desta ferramenta para análise dos dados e repasses das verbas públicas. Na saúde temos como exemplo, taxa de mortalidade, expectativa de vida, nascimentos, evolução ou declínio de algum tipo de doença.

Existe o interesse teórico sobre a composição dos indicadores que são comumente utilizados em pesquisas acadêmicas, e nelas, eles são um elo de ligação entre a teoria de conceitos sociais e a experiência vivida, por um conhecimento que é usado no dia a dia e não tem comprovações científicas sobre o assunto. Existe também o conhecimento de indicadores que se comportam como um instrumento operacional para monitoramento de uma realidade social, para a função de formulação e reformulação de políticas públicas e criação de estatísticas públicas.Vale destacar o impacto que os indicadores tem em políticas públicas, principalmente para tomada de decisões.

Na atualidade com o avanço da COVID-19, um dos temas mais comentados são indicadores de mortalidade e expectativa de vida. Com base nos dados do IBGE a expectativa de vida vinha tendo uma tendência de crescimento da taxa por anos consecutivos, em 2011, a esperança de vida do brasileiro era de 74,1 anos, já em 2019 passou a ser 76,6 anos. Provavelmente ocorrerá uma mudança nessa tendencia de crescimento, então esses indicadores são importantes para estudos futuros e a forma como se estuda os indicadores são relevantes para a sociedade.

Existem alguns trabalhos anteriores como a proposta abordada neste estudo. Chao (2017) usa os indicadores de taxa de mortalidade materna e taxa de mortalidade na infância (crianças menores de 5 anos) para fazer a estimação usando métodos Bayesianos. Li, Shuaibing(2017) propõe no estudo uma abordagem Bayesiana para determinar o índice probabilístico com relação à condição de transformadores de potência ou transformadores elétricos. Schmertmann e Gonzaga (2019), que é a grande referencia desse trabalho, propõe o estudo da estimação Bayesiana de taxas de mortalidade e expectativa de vida por sexo e idade em áreas menores com registros vitais incompletos.

O objetivo principal deste trabalho é estudar as aplicações da estatística Bayesiana no estudo de indicadores sociais. Como objetivos específicos de compreender, e definir os tipos de Indicadores e como que eles são melhor aplicados, compreender o uso de estatística Bayesiana, ajustar a teoria de inferência Bayesiana com o estudo dos indicadores, usar um

banco de dados com as aplicações de estatística Bayesiana. Como a aplicação atualizar o estudo "Bayesian estimation of age-specific mortality and life expectancy for small areas with defective vital records"feito pelo Marcos R. Gonzaga. Só que usando dados de Estados.

No Capitulo 2 é apresentada uma revisão bibliográfica sobre indicadores sociais e inferência Bayesiana, no Capitulo 3 é descrita a aplicação que foi realizada, descrevendo a estrutura do modelo, no Capitulo 4 mostramos os resultados encontrados, e por fim no Capitulo 5 as conclusões que foram apresentadas.

2 Revisão Bibliográfica

2.1 Indicadores Sociais

Um indicador social é uma medida em geral quantitativa dotada de significado social substantivo, usado para substituir, quantificar ou operacionalizar um conceito social abstrato, de interesse teórico (para pesquisa acadêmica) ou programático (para formulação de políticas). É um recurso metodológico, empiricamente referido, que informa algo sobre um aspecto da realidade social ou sobre mudanças que estão se processando na mesma.(JANNUZZI, 2001)

2.1.1 Indicador Social e Estatísticas Públicas

Segundo Jannuzzi (2001) estatísticas públicas correspondem ao dado social na sua forma bruta, não totalmente contextualizado em uma Teoria Social ou uma finalidade já programada, o dado é só parcialmente preparado para uso na interpretação empírica da realidade. Exemplos de estatísticas públicas são os dados censitários (IBGE), censo escolar (INEP - Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira), dados do Ministério da Saúde (DATASUS - Portal da Saúde), entre outras. As estatísticas públicas são extremamente úteis para a construção dos indicadores que permitem uma estimação mais contextualizada e comparativa da realidade social, exemplo disso são as taxas de mortalidade, taxa de evasão escolar, taxa de natalidade, etc.

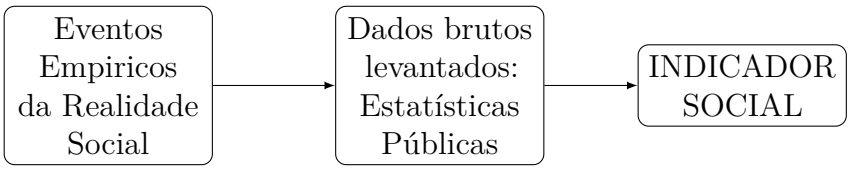

Figura 1 – Construção de Indicadores Sociais.

2.1.2 Classificação de Indicadores

Os indicadores podem ser divididos entre indicadores quantitativos e qualitativos. Indicadores quantitativos apresentam um acontecimento concreto da realidade social e são construídos a partir de estatísticas públicas, disponíveis. Os indicadores qualitativos são medidas construídas a partir de avaliações, levantadas em pesquisas de opinião, ou grupos de discussão.

Existem também outras classificações usuais para os indicadores:

- Indicador simples \ composto

- Indicador descritivo \ normativo
- Indicador quantitativo \ qualitativo
- Indicador objetivo \ subjetivo
- Indicador absoluto \ relativo

2.1.3 Propriedades Desejáveis dos Indicadores

No processo de construção dos indicadores sociais é possível identificar quais são as propriedades que os indicadores podem ter. Essas propriedades servem para que os indicadores sejam construídos de uma forma onde é possível saber onde estão os pontos fortes e fracos do que está sendo estudado.

2.1.3.1 Relevância social

A relevância social é um atributo fundamental para justificar a produção e legitimar o uso no processo de análise, formulação e implementação de políticas, então a relevância social do indicador se refere à sua pertinência para explicar a realidade em análise. A pertinência de produção do indicador é historicamente determinada, resultante de discussões de cada sociedade. Atualmente, há produção de indicadores mais específicos e geograficamente mais representativos, como forma de melhor entender cada realidade social. Por exemplo, problemas de exclusão e desigualdade social em países em desenvolvimento geram a necessidade de coleta de estatísticas e de construção de indicadores sobre intensidade de pobreza, níveis de carência e acesso a bens e serviços públicos. Ex: setores censitários; Indice de Desenv. Humano - IDH.

2.1.3.2 Validade

A validade do indicador corresponde ao grau de proximidade entre o conceito e a medida, isto é, a capacidade de refletir o conceito abstrato a que o indicador se propõe a substituir e operacionalizar. Diz respeito à proximidade entre indicador e indicando, propriedade fundamental para justificar o emprego e a denominação de uma medida quantitativa qualquer como um indicador social. Por exemplo, percentual de famílias com renda abaixo de um salário mínimo geralmente é um indicador mais adequado para retratar o nível de pobreza de uma população, do que a renda média per capita.

2.1.3.3 Confiabilidade

Confiabilidade diz respeito à qualidade do levantamento dos dados usados na estimação do indicador. Indicadores calculados por pesquisas amostrais realizadas por agências públicas são medidas confiáveis, porque os dados são coletados de forma padronizada, por

corpos técnicos qualificados, e seguindo uma metodologia de obtenção, registro e avaliação das informações. É preciso eliminar toda variação não aleatória na coleta e processamento dos dados para garantir que mudanças no indicador, ao longo do tempo, sejam analisadas de forma consistente.

2.1.3.4 Cobertura

É importante dispor de indicadores com boa cobertura espacial ou populacional de forma que sejam representativos da realidade empírica em análise. Os dados dos censos demográficos são importantes para o planejamento público justamente porque têm cobertura de todo o território nacional, além de possuir diversas variáveis para análise. Mesmo dados de órgãos públicos com cobertura parcial (tanto geograficamente, como conceitualmente) podem gerar importantes indicadores para a análise da realidade social.

2.1.3.5 Sensibilidade

Um indicador é sensível se for capaz de refletir mudanças significativas, em momentos que as condições que afetam a dimensão social em estudo se alterarem. Ao realizar a avaliação do impacto de um programa social, é preciso verificar qual indicador responde mais às mudanças implementadas na realidade social. Um indicador pode não apresentar mudanças estatisticamente significativas após a aplicação de políticas públicas, não somente porque não houve uma mudança nas condições de vida da população, mas talvez porque ele não possui sensibilidade suficiente para avaliação do tópico em estudo.

2.1.3.6 Especificidade

Um indicador é específico se tem a propriedade de refletir alterações ligadas somente às mudanças relacionadas à dimensão social em estudo. Se os indicadores constitutivos de indicadores compostos (índices sociais) têm baixa associação entre si, tais índices podem não ser específicos o suficiente para mostrar variações na direção esperada. Pode ser preferível utilizar um indicador parcial e limitado, mas que apresenta um significado claro de identificação com a realidade social.

2.1.3.7 Inteligibilidade de sua construção e Comunicabilidade

Inteligibilidade se refere à transparência da metodologia de construção do indicador. Um indicador também deve ser facilmente compreensível aos demais (comunicável). Isso é muito importante para indicadores voltados à formulação de políticas, já que a alocação de recursos públicos só pode se legitimar tecnicamente se os agentes envolvidos entenderem os critérios metodológicos utilizados, ainda que não concordem com os mesmos. A inteligibilidade e comunicabilidade são importantes para garantir a transparência no uso programático do indicador

2.1.3.8 Factibilidade para obtenção e Periodicidade na atualização

É preciso que o indicador possa ser factível de obtenção a custos acessíveis pelos órgãos de coleta ou pesquisadores. Um indicador se torna mais rico se há a possibilidade de coletar as estatísticas que o compõem com uma certa periodicidade. A regularidade com que as estatísticas sociais são coletadas indica se é factível a utilização do indicador em estudos específicos. O custo e tempo para obtenção do indicador têm que ser compatíveis com as necessidades e usos que se faz do mesmo.

2.1.3.9 Desagregabilidade

É importante que os indicadores se refiram aos grupos sociais de interesse (população-alvo) dos programas. Os indicadores sociais devem se referir aos espaços geográficos em análise (Estados, municípios, áreas de ponderação, setores censitários), a sub-grupos sociodemográficos (crianças, idosos, mulheres), ou grupos vulneráveis específicos (desempregados, analfabetos).

2.1.3.10 Historicidade

Historicidade de um indicador é a propriedade de se dispor de séries históricas extensas e comparáveis do mesmo. Dessa forma é possível comparar os níveis atuais com os do passado, estimar tendências e avaliar efeitos de políticas sociais implementadas. É importante que indicadores passados sejam compatíveis conceitualmente e tenham confiabilidade similar aos indicadores atuais.

2.2 Indicadores Demográficos

2.2.1 Taxa de Mortalidade

É o número total de óbitos, normalmente calculado por mil habitantes, na população residente em determinado espaço geográfico, no ano considerado.

$$TBM = \frac{R_x}{N_x} * 1000,$$

em que R_x é o número de óbitos registrados e N_x é a população exposta.

O indicador expressa a intensidade com a qual a mortalidade atua sobre uma determinada população. A taxa bruta de mortalidade é influenciada pela estrutura da população quanto à idade e ao sexo. Quando as taxas estão elevadas, podem estar associadas a baixas condições socioeconômicas ou refletir elevada proporção de pessoas idosas na população total. As taxas brutas de mortalidade padronizadas permitem a comparação temporal e entre regiões.

A Taxa Bruta de Mortalidade pode ser usada para analisar variações geográficas e temporais da mortalidade. É usada para viabilizar o cálculo do crescimento vegetativo ou natural da população, subtraindo-se, da taxa bruta de natalidade a taxa bruta de mortalidade. A TBM também contribui para estimar o componente migratório da variação demográfica, correlacionando-se o crescimento vegetativo com o crescimento total da população.

Esse indicador pode ter algumas limitações. O uso de dados de mortalidade, derivados de sistemas de registro contínuo, está condicionado a algumas correções, devido ao percentual não informado de óbitos, que acontece com frequência em áreas menos desenvolvidas e possíveis variações no número de óbitos, sobretudo em áreas com número reduzido de eventos. Recomendam o uso de médias usando três anos.

A base de dados demográficos utilizada para o cálculo do indicador pode apresentar dados sem muita precisão, decorrentes da coleta de dados ou da metodologia usada para elaborar as estimativas populacionais.

As projeções demográficas perdem precisão à medida que se distanciam dos anos de partida (ano censitário). Como a taxa é fortemente influenciada pela estrutura etária da população, a análise comparada entre populações de composição distinta exige padronização das estruturas etárias. As taxas padronizadas devem ser utilizadas apenas para análises comparativas.

A tábua de mortalidade, também conhecida como tábua de vida, é uma maneira de modelar a mortalidade de uma das idades especificas. A tábua de mortalidade é uma tabela que mostra informações sobre a mortalidade de uma coorte. Em sua forma clássica, a primeira coluna desta tabela representa a idade (em anos) de uma coorte e todas as outras colunas representam funções relacionadas à mortalidade, como o número de sobreviventes em determinadas idades, taxas de mortalidade por idade, mortes em intervalos de idade, expectativa de vida, entre outras. Uma das colunas de uma tábua de mortalidade é a probabilidade de morte à idade x e $x+n$, representada por ${}_n\mathrm{Q}_x$. É calculada através da relação entre o número de óbitos observados no intervalo de idade x e $x+n$, representado por ${}_n\mathrm{D}_x$ e o número de sobreviventes à idade exata x da população neste intervalo de idade, representado por l_x. Ou seja:

$$ {}_n\mathrm{Q}_x = \frac{{}_n\mathrm{D}_x}{\mathrm{l}_x}. $$

2.2.2 Expectativa de Vida

A expectativa de vida, que também pode ser chamada de esperança de vida ao nascer, consiste na estimativa do número de anos que se espera que um indivíduo possa viver, mantendo o padrão de mortalidade existente na população residente, no ano considerado. Esse dado é muito importante, visto que é um dos critérios utilizados pelo Programa

das Nações Unidas para o Desenvolvimento (Pnud), e também para calcular o Índice de Desenvolvimento Humano (IDH) de um determinado lugar.

Representa uma medida resumida da mortalidade, não estando afetada pelos efeitos da estrutura etária da população, como acontece com a taxa bruta de mortalidade. O aumento da esperança de vida ao nascer sugere melhoria das condições de vida e de saúde da população.

Usada para analisar variações geográficas e temporais a expectativa de vida da população, contribui para a avaliação dos níveis de vida e de saúde da população, subsidiar processos de planejamento, gestão e avaliação de políticas de saúde e de previdência social, entre outras, que estão relacionadas com o aumento da expectativa de vida ao nascer (oferta de serviços, atualização de metas, cálculos atuariais).

Esse indicador tem algumas limitações, como as imprecisões relacionadas a falhas na declaração da idade nos levantamentos estatísticos, ou à metodologia empregada para elaborar estimativas e projeções populacionais na base de dados. Utilizada para o cálculo do indicador, também para o cálculo da esperança de vida, são exigidas informações confiáveis de óbitos classificados por idade, e quando tem problema nos dados de mortalidade a expectativa de vida pode ficar prejudicada. Quando a precisão dos dados de sistemas de registro contínuo não é satisfatória, o cálculo deve basear-se em procedimentos demográficos indiretos, aplicáveis a áreas geográficas abrangentes.

Como citado anteriormente a tábua de mortalidade também possui a uma coluna com os valores da expectativa de vida por idade, que podem ser calculadas a partir da formula:

$$e_x^0 = \frac{T_x}{l_x},$$

em que T_x é o número de pessoas-ano vividos a partir da idade x, e l_x e o número de sobreviventes da população neste intervalo de idade.

2.3 Estatística Bayesiana

A estatística pode ser dividida em dois ramos principais a estatística descritiva e a estatística inferencial, o que vai ser abordado neste trabalho é a estatística inferencial, que busca tirar conclusões de um todo a partir de uma fração de informações, ou seja, quando é feita uma pesquisa, não é apenas uma descrição de dados e sim inferindo para uma população a partir de uma amostra o resultado da pesquisa. A estatística inferencial ainda pode ser dividida em duas vertentes a estatística clássica ou frequentista, e a estatística Bayesiana que é a que vai ser abordada.

É importante destacar que os indicadores são estatísticas descritivas, e a ideia é ter a junção com inferência Bayesiana para ter um indicador onde possa incorporar a incerteza, devido a má qualidade dos dados.

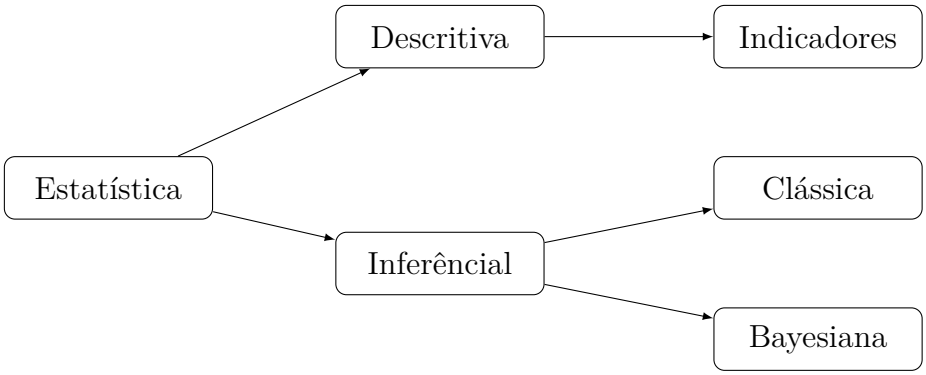

Figura 2 – Fluxograma mostrando as divisões da Estatística.

Na inferência Bayesiana além da amostra coletada, é adicionado ao modelo preditivo informações subjetivas, isto é, o pesquisador admite a existência de algum conhecimento que os dados coletados não dão conta de prever e que podem fazer diferença para a previsão. A inferência Bayesiana surge a partir do teorema de Bayes, que demonstra matematicamente como estas informações não contidas na amostra (chamadas de informações a priori) devem ser incorporadas no modelo preditivo.

A inferência Bayesiana tem como objetivo central estimar os parâmetros de um modelo através da distribuição de probabilidade, isto é, assumir uma distribuição de probabilidade atualizada com os dados da amostra. Além disso, a partir da inferência é possível gerar observações através da distribuição atualizada dos parâmetros.

Para a inferência Bayesiana o foco está na incerteza que se tem ao estimar o parâmetro do modelo probabilístico com o objetivo de levar em consideração o desconhecimento sobre o parâmetro. Com isso, esse parâmetro do modelo, na visão Bayesiana, é visto como uma variável aleatória e lhe é atribuído uma distribuição a priori, que representa a quantidade de informação conhecida sobre ele. A partir disso, é possível obter distribuições de probabilidades de uma dada variável, o que faz com que as probabilidades sejam atualizadas com a quantidade de informação obtida com os dados.

2.3.1 Teorema de Bayes

A base da inferência Bayesiana é um corolário da lei da probabilidade total o teorema de Bayes, nesse corolário é relacionada as probabilidades marginais e condicionais. O teorema de Bayes permite descobrir a probabilidade de um evento usando uma informação que ja ocorreu (priori), é dado por:

$$P(A \mid B) = \frac{P(A)P(B \mid A)}{P(B)}, \tag{2.1}$$

em que,

$P(A \mid B)$ é probabilidade condicional de A dado B

$P(A)$ é a probabilidade do evento conhecido A

$P(B \mid A)$ é probabilidade condicional de B dado A

$P(B)$ é a probabilidade marginal de B.

De maneira simplificada, a leitura de (3.1) é: probabilidade de A acontecer visto que B já ocorreu. Para isso, calcula-se a probabilidade a priori de A (conhecimento acumulado por eventos anteriores) multiplicada pela verossimilhança (dados coletados) dividida pela probabilidade B (evidência).

A equação (3.1) pode ser vista de outra forma e se dará por uma quantidade desconhecida θ e por H que representa o conjunto que contém a informação conhecida, a priori, sobre θ. Utilizando a forma condicional $p(\theta \mid H)$ sendo a informação sobre o parâmetro dado o conjunto de informações sobre ele e a atualização da informação através da amostra dos dados representado por $P(y \mid \theta, H)$ e por $p(y \mid H)$, tem a forma de atualização do parâmetro através da distribuição inicial e da informação dos dados dada por:

$$P(\theta \mid y, H) = \frac{P(\theta \mid H)P(y \mid \theta, H)}{P(y \mid H)},$$

onde,

$$P(y \mid H) = \int_{\theta} P(y, \theta \mid H)d\theta = \int_{\theta} P(y \mid \theta, H)P(\theta \mid H)d\theta.$$

$P(\theta \mid H)$ é a distribuição a priori de θ , ou seja, é a distribuição antes de se observar os dados y.

$P(y \mid \theta, H)$ é denominada a função de verossimilhança de θ, representa a função do parâmetro a partir das observações.

$P(\theta \mid y, H)$ é denominada distribuição a posteriori que representa o resultado após a observação dos dados.

Essa formula de Bayes pode ser simplificada por:

$$P(\theta \mid y) \propto P(y \mid \theta)P(\theta).$$

2.3.2 Prioris

Antes de definir o conceito de priori é importante definir o conceito de Permutabilidade, pois, tem papel de mostrar que a distribuição a priori é um elemento importante na composição junto com a função de verossimilhança. A partir do teorema da permutabilidade a variável pode ser considerada uma sequência permutável de observações. Assim, as sequências permutáveis são identicamente e independentemente distribuídas.

Como já foi falado anteriormente, prioris são distribuições que aclaram a distribuição inicial dos dados antes de observá-los. Como as prioris dependem, em geral, de um conhecimento prévio do pesquisador, a sua comprovação nem sempre é fácil e sua especificação errada pode trazer consequências posteriores. Existem alguns tipos de prioris como: prioris conjugadas, priori não informativa, e prioris especificadas por modelos hierárquicos.

2.3.2.1 Prioris conjugadas

A forma mais comum em especificar as informações a priori para os parâmetros, é utilizar distribuições de probabilidade que pertencem a famílias de distribuições conhecidas. Com o uso dessas famílias é necessário definir todos os seus parâmetros, deixando a priori bem especificada já que se trata de uma informação conhecida. Por exemplo a priori de um determinado parâmetro θ, referente a média da população que segue uma distribuição Normal, dessa forma é necessário definir uma média e uma variância para a priori, esses valores são conhecidos como hiperparâmetros. Uma priori é chamada de conjugada a um determinado modelo se a distribuição a priori e a posteriori pertencem a mesma classe de distribuições. Sendo assim a atualização do conhecimento que se tem de θ envolve apenas a mudanças de hiperparâmetros. Neste caso, o aspecto sequencial do método Bayesiano pode ser explorado definindo-se apenas a regra de atualização dos hiperparâmetros já que as distribuições permanecem as mesmas.

Definição. Se $F = \{p(x \mid \theta), \theta \in \Theta\}$ e uma classe de distribuições amostrais então uma classe de distribuições P é conjugada a F se

$$\forall p(x \mid \theta) \in F,$$

$$e$$

$$p(\theta) \in P \Rightarrow p(\theta \mid x) \in P.$$

2.3.2.2 Prioris não informativas

A falta de informações a priori a respeito do parâmetro desconhecido θ não impossibilita a inferência Bayesiana, nessas situações, são utilizadas prioris não informativas. Com isso, o objetivo é atribuir pouca influência na distribuição a priori para que ela não discorra significativamente na distribuição a posteriori. Então, com essa utilização, as posteriores serão praticamente formadas apenas com a verossimilhança, fazendo com que o resultado da posteriori usando uma priori não informativa tenha um resultado semelhante com a inferência clássica cuja estimação é feita somente a partir da verossimilhança.

Um exemplo clássico para quando não existe informação é o uso de uma distribuição Uniforme, dessa forma, todos os possíveis valores para o parâmetro são igualmente prováveis. Porem, se o intervalo de definição de θ for ilimitado vamos ter uma distribuição imprópria, já que sua distribuição acumulada deverá ter valor maior que 1.

2.3.3 Modelos Hierárquicos Bayesianos

Modelos Bayesianos hierárquicos são apropriados para modelar problemas com estruturas complexas de dependência. A ideia é dividir a especificação da distribuição priori em estágios. Com isso, a informação advinda da priori é constituída em duas etapas: a

primeira sendo a estrutural, que irá determinar a divisão dos estágios, e a segunda sendo a subjetiva, que irá especificar a distribuição para cada estágio, uma vantagem advinda da segunda etapa é que pode se utilizar mais de um estágio, hierarquia, e cada um deles pode apresentar uma distribuição diferente que melhor se adeque para a utilização desse tipo de priori. É natural esse tipo de abordagem em determinadas situações experimentais e também facilita nas especificações.

A distribuição a priori de θ depende dos hiperparâmetros ϕ, então podemos substituir $p(\theta)$ e escrever $p(\theta \mid \phi)$. Além disso, ao invés de fixar valores para os hiperparâmetros podemos especificar uma distribuição a priori $p(\phi)$ completando assim o segundo estágio na hierarquia. Dessa forma, a distribuição a priori conjunta é simplesmente

$$p(\theta, \phi) = p(\theta \mid \phi)p(\phi),$$

e a distribuição a priori marginal de θ pode ser obtida por integração como

$$p(\theta) = \int p(\theta, \phi)d\phi = \int p(\theta \mid \phi)p(\phi)d\phi,$$

A distribuição a posteriori conjunta fica

$$p(\theta, \phi \mid x) \propto p(x \mid \theta, \phi)p(\theta \mid \phi)p(\phi) \propto p(x \mid \theta)p(\theta \mid \phi)p(\phi).$$

Pois a distribuição dos dados depende somente de θ. Usando outras palavras, dado θ, x, e ϕ são independentes.

Um modelo hierárquico Bayesiano pode ser representado considerando, pelo menos 3 níveis de hierarquia:

- 1^o nível: modelo das observações: de X com os parâmetros;

- 2^o nível: modelo dos parâmetros e com seus hiperparâmetros;

-3^o nível: modelo dos hiperparâmetros.

Esses níveis de hierarquia podem ser explicados melhor, onde o 1^o nível é a especificação da verossimilhança, 2^o nível é as funções de ligação, e 3^o nível priori dos parâmetros

$$\left(\mu, \tau^2\right) \sim p(\mu, \tau^2),$$
$$\left(\theta_i \mid \mu, \tau^2\right) \sim N(\theta, \tau^2),$$
$$(y_i \mid \theta_i) \sim N(\theta_i, \sigma_i^2).$$

2.3.3.1 Monte Carlo via Cadeias de Markov

Um dos métodos mais conhecido e utilizados em inferência Bayesiana é a simulação de Monte Carlo via cadeias de Markov (MCMC). O objetivo é obter uma amostra da distribuição posteriori, sem saber exatamente qual é essa distribuição analiticamente, para isso utilizar a informação da priori e a verossimilhança, e calcular estimativas amostrais de características desta distribuição.

O MCMC é um método de simulação interativo, baseado em cadeias de Markov. Por conta dessa característica os valores não são independentes. Dessa forma, podem ser ultilizados alguns artifícios com o objetivo de obter valores independentes e identicamente distribuídos (iid)

Uma cadeia de Markov é um processo estocástico, no qual a distribuição de um X_t dados todos os valores anteriores, depende apenas do ultimo valor anterior (X_{t-1}). Os métodos de MCMC exigem que a cadeia de Marlov seja:

- Homogênea, isto é, as probabilidades de transição de um estado para outro não se alteram;

-Irredutível, ou seja, que cada estado pode ser atingido a partir de qualquer outro em um número finito de iterações;

- Aperiódica, que não haja estados absorventes;

Como resultado do MCMC tem-se uma amostra da posteriori, na qual é possível realizar inferências. Os métodos mais comuns são o Metropolis-Hastings e o Amostrador de Gibbs.

2.3.3.1.1 Metropolis-Hastings

Metropolis-Hastings é um algoritmo de que utiliza de uma distribuição auxiliar para obter um valor para a cadeia de Markov. Este mecanismo de correção garante a convergência da cadeia para a distribuição de equilíbrio, que neste caso é a distribuição a posteriori.

Suponha que a cadeia esteja no estado θ e um valor θ' é gerado de uma distribuição proposta $q(\cdot \mid \theta)$. Note que a distribuição em questão pode depender do estado atual da cadeia de Markov. O novo valor θ' é aceito com probabilidade

$$\alpha(\theta, \theta') = min\left(1, \frac{\pi(\theta')q(\theta \mid \theta')}{\pi(\theta)q(\theta' \mid \theta)}\right),$$

em que, π é a distribuição de interesse.

Em termos práticos, o algoritmo de Metropolis-Hastings pode ser especificado pelos seguintes passos:

1- Inicializar com o contador de iterações $t = 0$ e especificar um valor inicial $\theta^{(0)}$.

2- Gerar um novo valor θ' da distribuição $q(\cdot \mid \theta)$.

3- Calcular a probabilidade de aceitação (θ, θ') e gerar $u \sim U(0, 1)$.

4- Se $u \leq \alpha$ então deve-se aceitar o novo valor e fazer $\theta^{(t+1)} = \theta'$, caso contrário é preciso rejeitar e fazer $\theta^{(t+1)} = \theta$.

5- Incrementar o contador de t para $t + 1$ e voltar para o passo 2.

Embora a distribuição proposta possa ser escolhida de forma arbitraria na prática é necessário tomar alguns cuidados e garantir que o algoritmo funcione de forma mais eficiente. Em situações onde são usadas aplicações Bayesianas a distribuição de interesse é

a própria posteriori, i.e $\pi = p(\theta \mid x)$ e a probabilidade de aceitação assume uma forma particular,

$$\alpha(\theta, \theta') = min\left\{1, \frac{p(x \mid \theta')}{p(x \mid \theta)} \frac{p(\theta')}{p(\theta)} \frac{q(\theta \mid \theta')}{q(\theta' \mid \theta)}\right\}.$$

2.3.3.1.2 Amostrador de Gibbs

A ideia básica do amostrador de Gibbs é tomar um problema multivariado numa sequência de problemas univariados que vão se iterar para produzir uma cadeia de Markov. (Cassela e George, 1992) A distribuição de equilíbrio é a distribuição a posteriori desejada.

No amostrador de Gibbs a cadeia irá sempre se mover para um novo valor, i.e não existe mecanismo de aceitação nem de rejeição. As transições de um estado para outro são feitas de acordo com as distribuições condicionais completas $\pi(\theta_i \mid \mathbf{\Theta_{-i}})$, onde temos $\mathbf{\Theta_{-i}} = (\theta_1, \cdots, \theta_{i-1}, \theta_{i+1}, \cdots, \theta_d)$.

Em geral, cada uma das componentes θ_i pode ser uni ou multidimensional. Portanto, a distribuição condicional completa é a distribuição da i-ésima componente de $\mathbf{\Theta}$ condicionada em todas as outras componentes. Ela é obtida a partir da distribuição conjunta como,

$$\pi(\theta_i \mid \mathbf{\Theta_{-i}}) = \frac{\pi(\mathbf{\Theta})}{\int \pi(\mathbf{\Theta}) d\theta_i}.$$

Assim, para obter a distribuição condicional completa de x_i basta pegar os termos da distribuição conjunta que não dependem de x_i.

Em algumas situações, a geração de uma amostra diretamente de $\pi(\mathbf{\Theta})$ pode ser complicada ou simplesmente impossível. Mas se as distribuições condicionais completas forem conhecidas, então o amostrador de Gibbs é definido pelo seguintes passos,

1- Inicializar o contador de iterações da cadeia com $t = 0$;

2- Especificar valores iniciais $\mathbf{\Theta}^0 = (\theta_1^0, \cdots, \theta_d^0)'$

3- Obter um novo valor de Θ^t a partir de Θ^{t-1} através da sucessiva geração de valores.

$$\begin{gathered}
\theta_1^t \sim \pi(\theta_1 \mid \theta_2^{(t-1)}, \theta_3^{(t-1)}, \cdots, \theta_d^{(t-1)}), \\
\theta_2^t \sim \pi(\theta_2 \mid \theta_1^{(t)}, \theta_3^{(t-1)}, \cdots, \theta_d^{(t-1)}), \\
\vdots \\
\theta_d^t \sim \pi(\theta_d \mid \theta_1^{(t)}, \theta_2^{(t)}, \cdots, \theta_{d-1}^{(t)}).
\end{gathered}$$

4- Incrementar o contador de t para $t + 1$, logo depois retornar para o passo 2 e chegar numa convergência.

Cada iteração se completa após uma quantidade d de movimentos ao longo dos eixos coordenados das componentes de Θ. Após a convergência, os valores restantes formam uma amostra de $\pi(\Theta)$.

É interessante observar que o amostrador de Gibbs é um caso especifico do algoritmo de Metropolis-Hastings, no qual os elementos de Θ são atualizados um de cada vez ou

em blocos. Usando a distribuição condicional completa como proposta e probabilidade de aceitação igual a 1.

3 Metodologia

As estimativas de mortalidade podem ser um grande problema, principalmente porque uma das maiores questões na estimação de mortalidade e na expectativa de vida são os problemas nos registros vitais, pois os dados de mortalidade muitas vezes podem não estar completos. O Brasil sendo um país extenso, e com uma distinção grande, tanto numa visão regional quanto para os subgrupos populacionais sofre muito com esse problema sub notificação dos dados de estatísticas vitais.

Então as fontes de incerteza são:

- Erros de cobertura e declaração de idade (registro vitais e censo demográfico).
- Baixa exposição (óbitos/população) em pequenas áreas.

Mesmo com a qualidade dos dados sendo por muitas vezes insatisfatórias, exite uma grande melhoria na qualidade dos dados, levando em consideração aos anos, a qualidade de registro de óbitos de homens adultos passou de 83,2% no período 1980-1991 para 89,7% no período 2000–2010

Queiroz, B (2017) fez um estudo para avaliar a qualidade do registro de óbitos do Datasus, por sexo e estados brasileiros, e estimar as probabilidades de morte adulta, por sexo e estados, entre 1980 e 2010 e como resultado indicam uma melhoria considerável do grau de cobertura de óbitos no Brasil desde 1980.

O estudo do Gonzaga, M (2019) é uma analise da estimação Bayesiana de taxas de mortalidade e expectativa de vida por sexo e idade em áreas pequenas com registros vitais incompletos. O contexto em que o estudo está é uma acelerada transição da mortalidade, mudanças no perfil das causas de morte, a estrutura etária de mortalidade, e uma heterogeneidade regional para subgrupos populacionais.

O problema em questão é montar o modelo de como estimar taxas de mortalidade e expectativa de vida em pequenas áreas no Brasil incorporando, nas estimativas, duas fontes de incerteza: alta variabilidade nas taxas observadas e o erro de cobertura no registros de óbitos por sexo e idade. E como lidar com a dificuldade de identificar a baixa taxa de mortalidade em conjunto com o alto grau de cobertura dos óbitos. E observando as altas taxas de mortalidade levando em consideração um baixo grau de cobertura dos óbitos e adequar um modelo Bayesiano com grau de cobertura mais provável (uma priori).

Foi usado um modelo de regressão Bayesiana que aborda especificamente os problemas fundamentais na estimativa de mortalidade de pequenas áreas em países com potencial de registro defeituoso. O modelo suaviza as taxas de mortalidade especificas por idade, em pequenas amostras ao mesmo tempo que contabiliza a incerteza sore o nível de registro de óbitos.

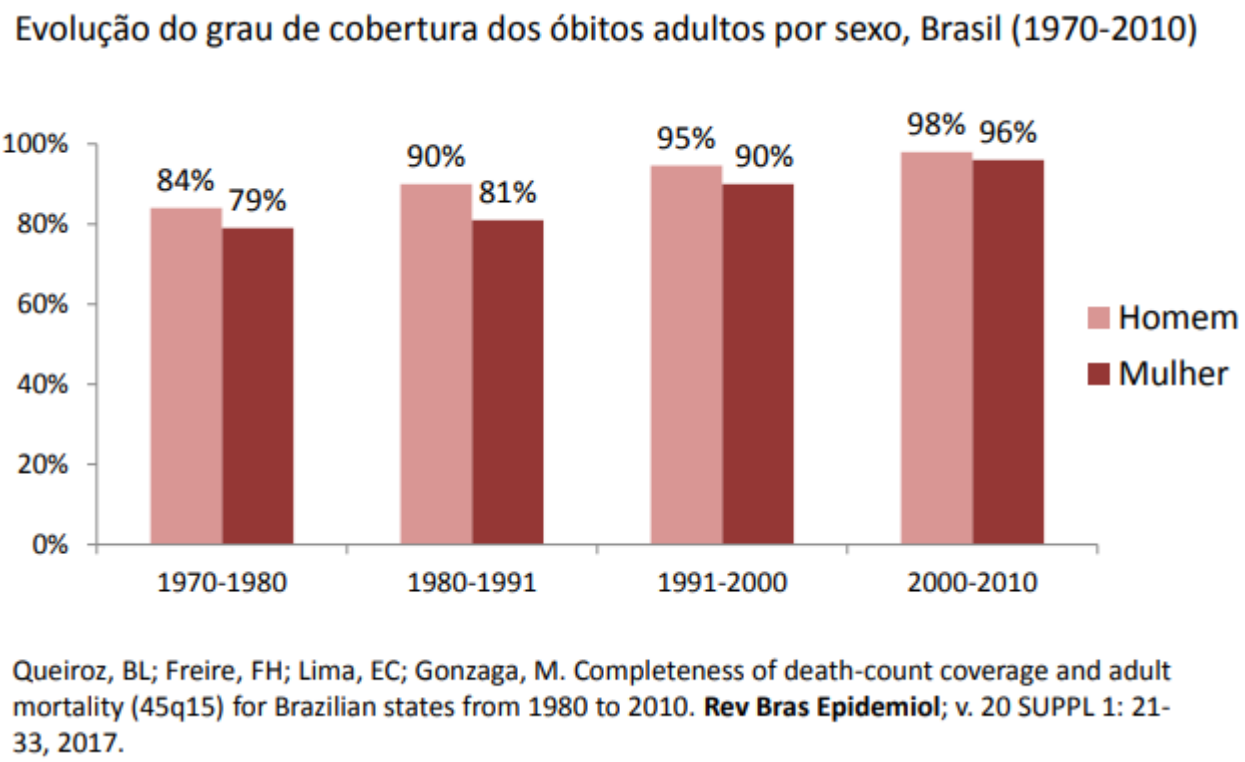

Figura 3 – Evolução do grau de cobertura dos óbitos adultos por sexo, Brasil (1970 - 2010)

A regressão Bayesiana produz estimativas de taxas de mortalidade e expectativa de vida. O modelo que Schmertmann e Gonzaga (2018) usam oferece várias vantagens em relação as abordagens não Bayesianas existentes. Além de incorporar as duas fontes de incerteza: a variabilidade de amostragem e a incerteza sobre cobertura de registro. Schmertmann e Gonzaga (2018) também usam uma nova abordagem de forma funcional para mortalidade que estabilizam estimativas de pequenas amostras sem exigir suposições fortes sobre os padrões de mortalidade específicos por idade. As taxas estimadas do modelo são funções continuas suaves, o que é o oposto de muitas correções existentes para sub notificação que utilizam de métodos diferentes para a mortalidade infantil e a mortalidade em adultos.

O objetivo de aplicação desse trabalho é atualizar os dados do estudo do Schmertmann e Gonzaga (2018) só que observando numa perspectiva estadual ao invés de uma visão municipal. Como o modelo que foi usado para pequenas regiões está bem definido, então usando o modelo para as Unidades Federativas é esperado que também seja aplicável.

3.1 Construção do Modelo

O modelo usado foi para óbitos registrados R_x. Em cada idade x de uma localidade, foi observado uma população exposta N_x e um número de óbitos registrados R_x.

O modelo hierárquico Bayesiano pode ser visto neste estudo onde o número verdadeiro

de óbitos é maior que o número de óbitos registrados que não é conhecido i. e $D_x \geq R_x$, então foi usado um modelo da forma expressa na Figura 4

Figura 4 – Modelo.

Assumindo que é verdade. As mortes têm distribuições de Poisson independentes em cada idade que dependem de taxas de mortalidade(μ_x)

$$D_x \sim Poisson(N_x \mu_x),$$

onde D_x são os Óbitos.

Além disso, assumindo que cada morte é registrada independentemente com uma idade específica de probabilidade (π_x), de forma que o número total de mortes registradas na idade x tenha uma distribuição binomial:

$$R_x \sim Binomial(D_x, \pi_x).$$

Schmertmann e Gonzaga (2018) mostram a prova da distribuição das mortes registradas (R_x). Isso implica $R_x \sim Poisson(N_x \mu_x \pi_x)$.

A distribuição posteriori completa combina a probabilidade R com a anterior para π, produzindo uma distribuição posteriori para μ que resume quais taxas de mortalidade são mais viáveis dado os dados observado

$$P(\mu \mid R, N) = \int_0^1 L(R \mid N, \mu, \pi) f_\pi(\pi) d\pi,$$

o resultado então é a distribuição posteriori das taxas de mortalidade locais

3.2 TOPALS

A mortalidade por cada idade especifica foi modelada com o modelo relacional TOPALS (Beer,2012). Em um modelo TOPALS, o modelo de log da mortalidade é a soma de duas funções: (1) um modelo constante (chamada de padrão) que incorpora as idades básicas padrões, e (2) uma função linear paramétrica composta de linhas retas com os nós representando as idades, que mostram as diferenças entre o esquema padrão e o log da mortalidade da população de interesse. O vetor de taxas de log ao longo das idades $x = 1, ..., 99$ e $\lambda \in R_{100}$ no modelo TOPALS é

$$\lambda = \lambda * + B\alpha,$$

onde, $\lambda* \in R_{100}$ é o padrão considerando o caso desse trabalho especifico são derivados de dados nacionais para o Brasil do censo de 2010. B é uma matriz 100 x 7 de funções com base linear B-spline fixas (Boor, 2001) e um vetor de parâmetro $\alpha \in R_7$.

Os sete parâmetros do modelo $\alpha = (\alpha_0, ..., \alpha_6)'$ são valores das funções das idades exatas 0, 1, 10, 20, 40, 70 e 100. Então a taxa de mortalidade na idade x em um modelo TOPALS é

$$\mu_x(\alpha) = exp(\lambda_x^* + b_x'\alpha),$$

sendo que, b_x' é o xth das linhas de B. De acordo com as premissas de distribuição descritas anteriormente o Log-Verossimilhança é

$$lnL(R \mid N, \alpha, \pi) = c - \sum_x [N_x \pi_x \cdot \mu_x(\alpha)] + \sum_x [R_x ln \mu_x(\alpha)],$$

em que, c é uma constante que não varia com α. Na prática as suposições de cobertura completa foram flexionadas e foram substituídas por distribuições a priori para π_x.

3.3 Resumo do Modelo

A Figura 5 resume bem a abordagem para estimar a mortalidade e a expectativa de vida. Como mostra no canto superior direito do organograma foi usada uma normal multivariada fraca antes de α, para estabilizar as estimativas em pequenas populações para que os valores dos Estados estejam certos.

São observadas exposições especificas para cada idade $(N_0, ..., N_{99})$ e mortes registradas $(R_0, ..., R_{99})$. O modelo combina as distribuições anteriores para os parâmetros de cobertura e mortalidade $(f_\pi e f_\alpha)$ e a Função de verossimilhança Poisson $L(R \mid N, \alpha, \pi) = \prod_x L(R_x \mid N_x, \mu_x(\alpha), \pi_x)$ para produzir uma distribuição posteriori marginal para $\alpha = (\alpha_0, ..., \alpha_6)'$

$$P(\alpha \mid R, N) = \int L(R \mid N, \alpha, \pi) f_\alpha(\alpha) f_\pi(\pi) d\pi.$$

Em que , $P(\alpha \mid R, N)$ responde a pergunta "Dadas as populações especificas de idade e mortos registrados, juntando com o conhecimento das probabilidades dos registros de óbitos locais. Quais os modelos para mortalidade que vão fazer mais sentido. Aplicando na prática a resposta dessa questão é aplicar essa equação aplicando Monte Carlo via cadeias de Markov (MCMC), fazendo um grande números interações aleatórias $\alpha_1^*, ..., \alpha_k^*$. Essa é a opção para fazer a mesma pergunta para expectativa de vida (e_0) fazendo uma distribuição posteriori simulada de $e_0(\alpha_1^*), ..., e_0(\alpha_k^*)$

$$P(e_0 < c \mid R, N) \approx \frac{1}{K} \sum_k I[e_0(\alpha_k^*) < c],$$

sendo que, I[] é uma função indicadora (0,1) igual a 1 se a condição entre colchetes for verdadeira.

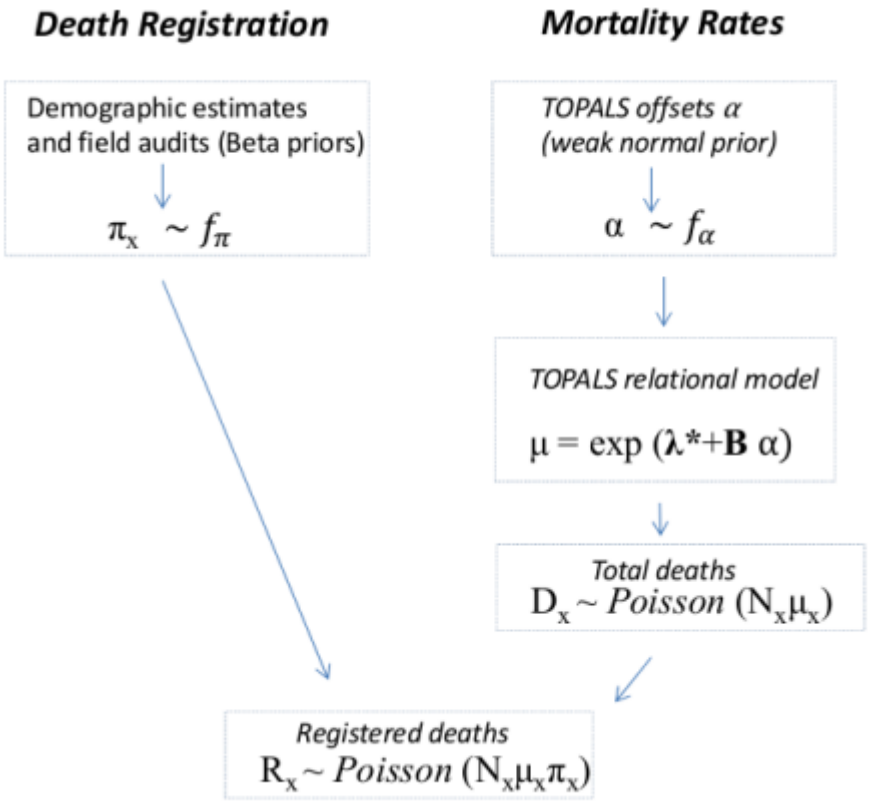

Figura 5 – Um modelo integrado de cobertura e mortalidade. Mortes registradas (R_x) e exposição (N_x) são observadas. α e π_x são parâmetros incertos (Schmertmann e Gonzaga, 2018)

3.4 Dados

Foi aplicado o modelo de estimativas para as Unidades da Federação. Os dados de mortalidade por sexo e por idade simples nos anos de 2017, 2018 e 2019 foram retirados da base do Datasus utilizando os microdados. Os óbitos registrados estão disponíveis para os 5.565 municípios brasileiros com idades de 0 a 99 anos dividido por sexo, ou seja, 1.113.000 combinações (município, idade, sexo) nos anos 2017-2019.

Já os dados de população usada foi a projeção da população por idade simples para o ano de 2019 retirados do IBGE (Instituto Brasileiro de Estatística e Geografia).

3.5 Prioris

As informações prévias sobre os níveis e os padrões de registros de óbitos são bastante importantes para o modelo. Para os hiperparâmetros da priori de α foram usados dados com base em correlações entre características de níveis municipais e nos níveis de cobertura de mortalidade, criando assim uma probabilidade para registro de óbitos para os 5.565 municípios brasileiros. Segundo Schmertmann e Gonzaga (2018) esses dados se mantém

constantes, então usamos as mesmas prioris para atualizar os dados.

Prioris para cobertura

Para usar os dados de cobertura, foi assumido que em cada área existe a probabilidade diferente para registro de óbitos para três grupos de idade. Esses intervalos são:

$$\pi_x = \begin{cases} \pi_0 \text{ if } x = 0 \\ \pi_1 \text{ if } x \in \{1, ..., 29\} \\ \pi_2 \text{ if } x \in \{30, ..., 99\} \end{cases},$$

que podem ser combinados para produzir uma probabilidade geral de registro:

$$\pi_{total} = w_0\pi_0 + w_1\pi_1 + w_3\pi_3,$$

onde w_x representa a proporção de mortes registradas que ocorreram em cada grupo de idade.

As prioris associadas então

$$\pi_0 \sim Beta(K_0\hat{\pi}_0, K_0[1-\hat{\pi}_0]), (K_0 - 5) \sim Exponencial(0.05),$$
$$\pi_{total} \sim Beta(K_{total}\hat{\pi}_{total}, K_{total}[1-\hat{\pi}_{total}]), (K_{total} - 5) \sim Exponencial(0.05),$$

onde K_0 e K_{todos} são hiperparâmetros que representam níveis (incertos) de precisão das estimativas.

Para as coberturas de mortalidade em idades acima de 30 anos (π_2) foi preciso usar informações retiradas do DDM que fornece as estimativas de π_2 a nível estadual retirando dados de 6 estudos e utilizando a média e variância das estimativas do DDM para estimar os parâmetros de uma distribuição beta pelo método de momentos. A prior obtida foi:

$$\pi_2 \sim Beta(K_2\Phi_2, K_2I[-\Phi_2]),$$

em que, Φ_2 é a média das seis estimativas DDM e K_2 é uma precisão estimada.

Por fim, usando as informação qualitativas, adicionando uma priori que exclui completamente as possibilidade de triplas probabilidades de cobertura que não correspondam a ordem assumida que

$$\pi_0 \leq \pi_2 \leq \pi_1.$$

Assumindo o fato que em todas as áreas a cobertura da mortalidade infantil não é maior que a cobertura da mortes nas idades de 1 - 29 anos que não pode ser inferior a de outras idades.

3.6 Computacional

Para a aplicação do modelo foi usado o software R, e para a implementação foi usado o pacote R-Stan (Carpenter. B, 2017), que se trata de uma linguagem que permite amostragem de MCMC de distribuições posterioris complexas. Foi aplicado o modelo feito pelo Schmertmann e Gonzaga (2018) com alterações para aplicar somente para UFs e com dados atualizados dos anos de 2017, 2018, e 2019 .Para ambos os sexos foram estimadas as distribuições posterioris de parâmetro de mortalidade (α), modelos completos para o log da mortalidade ($\lambda(\alpha) = \lambda^* + B\alpha$), e a expectativa de vida $e_0[\lambda(\alpha)]$ para as 27 Unidades da Federação brasileira isso foi definido como o modelo ajustado. Com o objetivo de entender sobre os efeitos das sub-notificações de óbitos, aplicando as mesmas distribuições posterioris sob uma suposição do registro incorreto de registro $100\%(\pi_0 = \pi_1 = \pi_2)$ essa versão sendo o modelo não ajustado.

4 Resultados

É de conhecimento de pesquisadores da área de saúde e epidemiologia no Brasil que o registro de óbitos registrado do país não é de 100%, isto é, o número de óbitos registrado é menor do que o número de óbitos que realmente aconteceu. Mesmo sabendo que essa cobertura melhorou bastante como o Queiroz, B (2017) mostrou. Esse problema de registro tem um impacto grande nas estimativas dos indicadores de expectativa de vida e as taxas de mortalidade.

A expectativa de vida é uma função não linear complexa em função do parâmetro α, que não se manifesta claramente, já a abordagem Bayesiana permite que esse calculo da incerteza em e_0 seja fácil tanto para a estimativa ajustada e a não ajustada, conforme utilizado na equação

$$P(e_0 < c \mid R, N) \approx \frac{1}{K} \sum_k I[e_0(\alpha_k^*) < c]$$

Na Figura 6 e 7 temos gráficos onde os pontos em azuis e rosas mostram o padrão encontrado no modelo proposto nesse trabalho. E os pontos onde tem o sinal "+"são os dados de mortalidade registrados. Esse gráfico mostra a qualidade dos registros de óbitos dos estados.

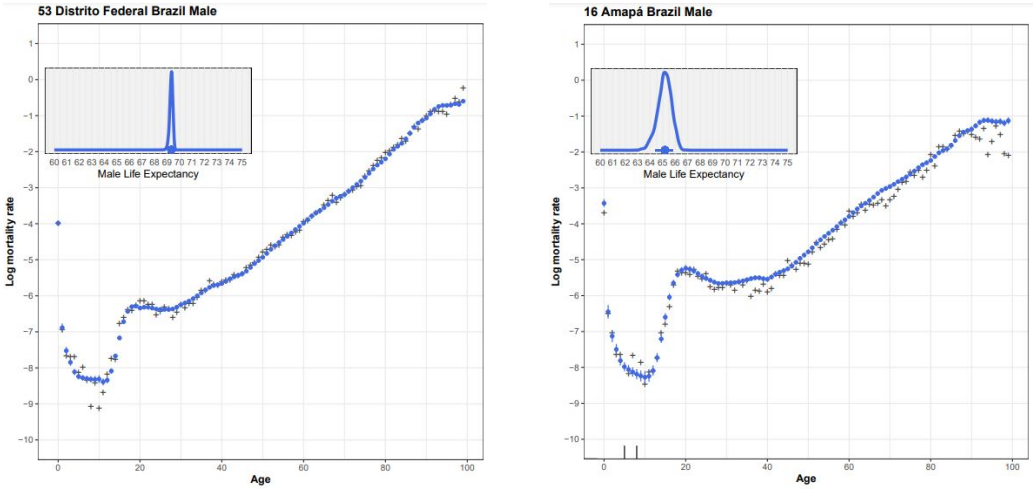

Figura 6 – Log Mortalidade e Expectativa de Vida do Distrito Federal e do Amapá para sexo masculino, no ano de 2019

Analisando os valores de expectativa de vida encontrados nesse trabalho e comparando com os valores de expectativa de vida que são divulgados oficialmente podemos ver uma diferença considerável. Nas Unidades da Federação a tendencia a ter os dados com problemas de registro como no Amapá essa incerteza sobre a expectativa de vida é bem maior em comparação com o Distrito Federal que os dados são registrados de forma bem

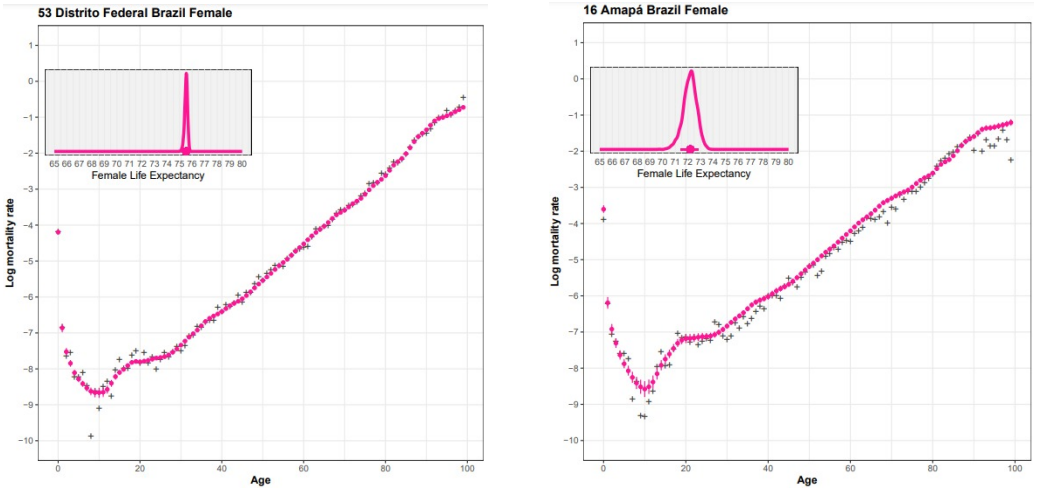

Figura 7 – Log Mortalidade e Expectativa de Vida do Distrito Federal e do Amapá para sexo feminino, no ano de 2019

mais confiável e completa, no entanto o Distrito Federal não necessariamente possui uma certeza sobre o assunto. Observando a Figura 6 e 7 onde mostram o Log da mortalidade dos dois estados, os desvios que tem no Amapá são bem mais visíveis em comparação com os do Distrito Federal, tanto os resultados masculinos quanto os femininos.

Olhando os valores da expectativas de vida oficiais do IBGE Figura 10, encontramos os valores que diferem bastante dos obtidos no nosso resultado que podem ser vistos nas Figuras 8 e 9. Isso se da por conta da instabilidade nos registros de determinadas UFs. São realizados grandes correções nos dados para valores menores nas expectativas de vida dos estados que possuem um número maior de casos de sub notificação de mortalidade normalmente que estão localizado nas regiões Norte e Nordeste. Os estados com registros mais estáveis são feitos pequenos ajustes como por exemplo Unidades da Federação localizados nas regiões Sul e Sudeste . E essa diferença nos valores encontrados e nos oficiais ocorre por consequência das taxas de correções utilizadas pelo IBGE. Então observando a Figura 8 e 9 podemos ver que são essas regiões que estão tanto no extremo de maior expectativa de vida como no extremo de menor.

Nas Figuras 8 e 9 mostram os intervalos de confiança de 80% para cada estado no sexo masculino e feminino respectivamente, o e_0 encontrados no modelo proposto. O gráfico mostra as barras horizontais que abrangem o intervalo entre 10º e 90º dos percentis da distribuição posteriori, os pontos sólidos indicam as medianas da posteriori e os estados estão classificados por ordem de medianas posteriori. Na Figura 9 onde temos os dados masculinos, Santa Catarina por exemplo tem a maior estimativa, seguido pelo Paraná assim por diante. Podemos ver estados com uma população grande como o Rio Grande do Sul e o Paraná que tem o registro vital é quase perfeito, e consequentemente a certeza sobre o e_0 é alta nesses estados. Em contraste temos os estados como o Roraima, Amapá e Amazonas que a cobertura de registros vitais são incompletos e não da pra saber ao certo

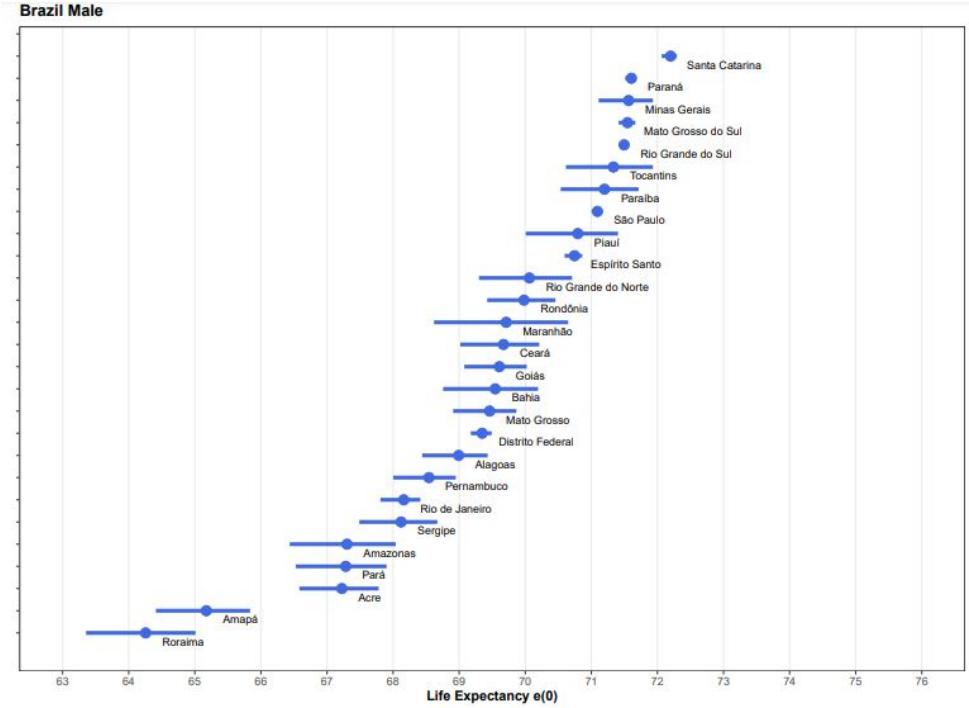

Figura 8 – Resultado do modelo - Valores de Expectativa de Vida nas Unidades da Federação, homens no Brasil em 2019

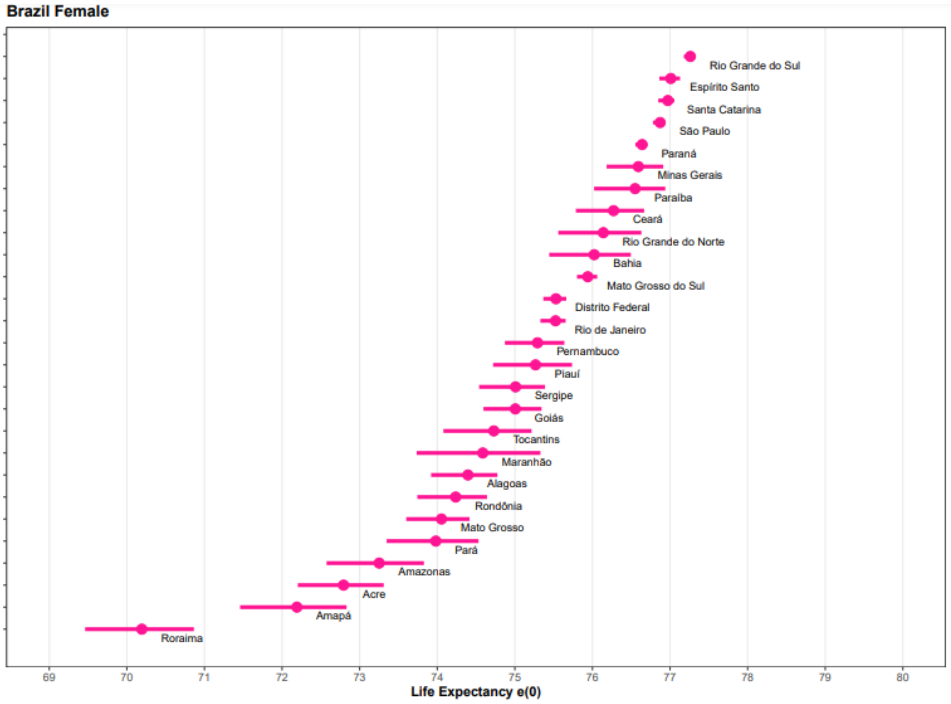

Figura 9 – Resultado do modelo - Valores de Expectativa de Vida nas Unidades da Federação, mulheres no Brasil em 2019

quem tem a maior expectativa de vida.

A tendencia é que as Unidades da Federação que tem melhor qualidade das informações são melhores como Santa Catarina, Distrito Federal, e São Paulo os valores do modelo em

comparação com os dados do IBGE estão próximos.

Nos resultados femininos podemos ver a mesma tendencia, onde os estados da região norte e nordeste mostram uma expectativa de vida menores e com problema na qualidade dos registros vitais. Já as maiores expectativas de vida estão nas regiões Sul e Sudeste, onde podemos destacar os estados do Rio Grande do Sul e Espirito Santo.

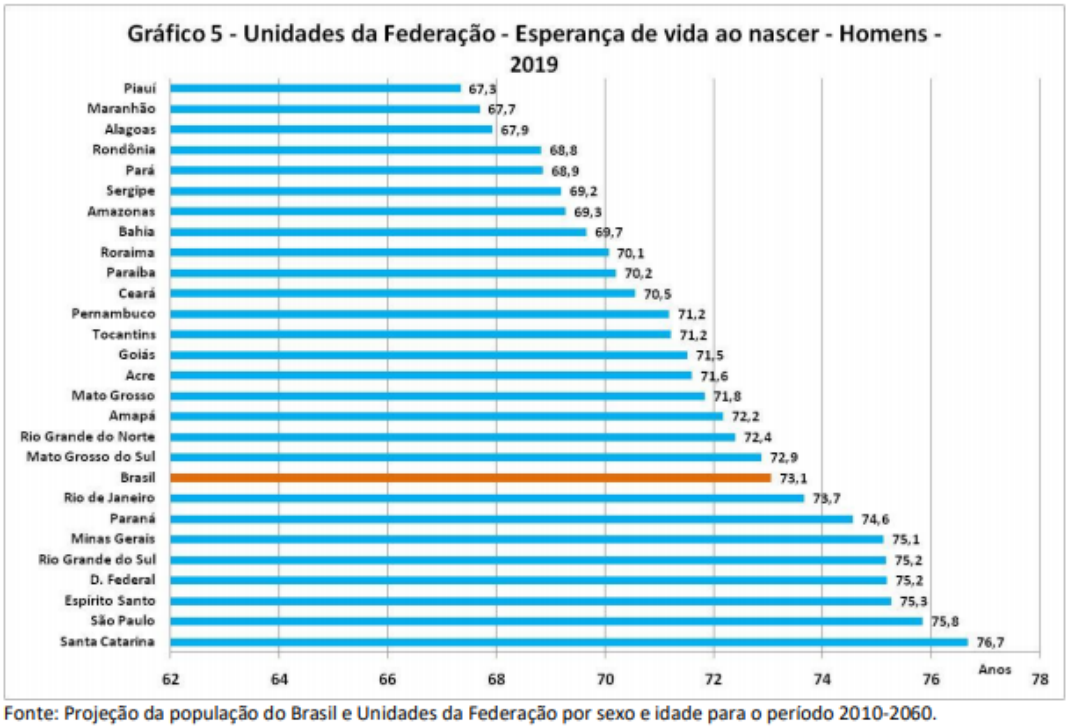

Figura 10 – Expectativa de vida oficial do IBGE, Fonte: IBGE

E importante destacar que como foram usadas projeções para os dados das populações estaduais, e nessas projeções o IBGE usa nos modelos complexos, e de varias etapas para corrigir o sub-registro da mortalidade, então quando é feito esse ajuste, e levado em consideração as sub notificações e por conta disso os valores das projeções podem aparecer diferenças entre as estimativas principalmente nessas Unidades da Federação em que os dados possuem mais correção, e por consequência do uso desses dados os nossos valores de expectativa de vida ficaram bem inferiores até mesmo considerando o estudo Schmertmann e Gonzaga (2019).

As comparações feitas entre os estados são importantes, considerando o motivo no qual os indicadores são usados, a mortalidade e expectativa de vida são indicadores muito usados em politicas publicas, saúde publica, em tomada de decisões, repasse de verbas. Mesmo com os dados sendo ajustados e também levando em consideração as sub notificações não é fácil estimar a expectativa de vida.

5 Conclusão

Um modelo Bayesiano é uma abordagem que deve ser mais explorada para o uso de indicadores, no caso da aplicação apresentada é o exemplo de como é um método que pode ajudar para estimar melhor os dados de mortalidade e de expectativa de vida, levando em consideração a incerteza atrelada a qualidade das informações.

Por conta dos sub registros e dos níveis de incerteza para estimar a expectativa em determinados estados ainda é considerado um desafio a ser vencido. Mesmo com bons métodos de estimar a mortalidade, a incerteza desses indicadores podem ser uma barreira.

Uma das grande contribuições que esse trabalho trouxe é a combinação entre os métodos estatísticos e a ferramentas de analise demográficas. Indicadores demográficos são de extrema importância para a sociedade e para as politicas publicas, pois é com eles que são feitas as principais tomadas de decisões do governo.

O estudo correto dos indicadores de mortalidade e de expectativa de vida tem um grande valor, com uma atenção voltada para a qualidade dos registros, pois com dados mais confiáveis é possível obter melhoras em vários cenários tanto politico, quanto social e demográficos. Também é importante destacar que indicadores com uma relevância tão grande precisam ser tratados com uma periodicidade maior, até para serem feitas melhores escolhas governamentais.

Importante destacar também que é ideal o uso de dados censitários para que os valores sejam mais fieis a realidade.

Contudo esse trabalho é um combustível para trabalhos futuros com esse tema. Propostas para achar uma priori mais acessível e que ajude a estimar esses indicadores da melhor forma.

Referências

Beer, J (2012) *Smoothing and projecting age-specific probabilities of death by TOPALS.* Demographic Research 27: 543–592. doi:10.4054/DemRes.2012.27.20. http://www.demographic-research.org/volumes/vol27/20/.

Boor, C. (2001) *A practical guide to splines. Applied mathematical sciences* Springer. https://books.google.com/books?id=m0QDJvBI_ecC.

CHAO, F (2017) *Bayesian Methods For Estimating Global Health Indicators.*, Public Health National University Of Singapore :14 August 2017

Departamento de Informática do Sistema Único de Saúde do Brasil (DATASUS). *disponível em: http://www2.datasus.gov.br/DATASUS/index.php/*

EHLERS. S.R (2007) *I*nferência Bayesiana, 5ª Ed. (2007)

GONZAGA, M (2019) *Estimação Bayesiana de taxas de mortalidade e expectativa de vida por sexo e idade em áreas menores com registros vitais incompletos* , Universidade de Brasília (UnB), 2019

GONZAGA, M (2016) *Estimating age-and sexspecific mortality rates for small areas with TOPALS regression: an application to Brazil in 2010*, Revista Brasileira de Estudos de População 33 (3): 629–652.

Instituto Brasileiro de Estatística e Geografia (IBGE) *disponível em: https://www.ibge.gov.br/*

Instituto Brasileiro de Pesquisa e Analise de Dados *O* que é estatística Bayesiana? , disponível em: https://www.ibpad.com.br/blog/o-que-e-estatistica-bayesiana/, 2020 .

JANNUZZI, P. M. I (2001) *Indicadores sociais no Brasil: conceitos, fonte de dados e aplicações.*, Campinas: Alínea, 2001.

Laboratório de Estimativas e Projeções Populacionais da UFRN (LEPP) *disponível em: https://demografiaufrn.net/laboratorios/lepp/*

LI, SHUAIBING (2017) *Bayesian Information Fusion for Probabilistic Health Index of Power Transformer.* , IET Generation, Transmission and Distribution. 12. 10.1049/iet-gtd.2017.0582.

PEDERIVA, B (2013) *Modelos Hierárquicos Bayesianos aplicados ao Marketing.*, Universidade Federal do Rio Grande do Sul, Porto Alegre 2013

R Development Core Team (2009) *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

REDE Interagencial de Informação para a Saúde *Indicadores básicos para a saúde no Brasil: conceitos e aplicações* Rede Interagencial de Informação para a Saúde - Ripsa. – 2. ed. – Brasília: Organização Pan-Americana da Saúde, 2008. 349 p.: il.

Schmertmann. C Gonzaga. M (2018) *Bayesian estimation of age-specific mortality and life expectancy for small areas with defective vital records*. https://osf.io/arn7d/