

UNIVERSIDADE FEDERAL DE MINAS GERAIS
Instituto de Ciências Exatas
Programa de Pós-graduação em Estatística

Thais Pacheco Menezes

Arquétipos Latentes: os Módulos dos Padrões espaciais complexos do câncer

Belo Horizonte
2020

Thais Pacheco Menezes

Arquétipos Latentes: os Módulos dos Padrões espaciais complexos do câncer

Dissertação apresentada ao Programa de Pós-graduação em Estatística da Universidade Federal de Minas Gerais, como requisito para a obtenção do título de Mestre em Estatística.

Orientador: Marcos Oliveira Prates

Coorientador: Renato Martins Assunção

Belo Horizonte
2020

Menezes, Thaís Pacheco.

M543a Arquétipos latentes: [recurso eletrônico] os módulos dos padrões espaciais complexos do câncer / Thaís Pacheco Menezes – 2020.

1 recurso online (43 f. il., color.): pdf.

Orientador: Marcos Oliveira Prates.

Coorientador: Renato Martins Assunção.

Dissertação (mestrado) - Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Estatística.

Referências: f. 42-43.

1. Estatística – Teses. 2. Análise espacial (Estatística) – Teses. 3. Decomposição em valores singulares – Teses. 3. Câncer – Brasil – Estatística – Teses. 4. Câncer – Estados Unidos – Estatística – Teses. I. Partes, Marcos Oliveira. II. Assunção, Renato Martins. III. Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Estatística. IV. Título.

CDU 519.2(043)



UNIVERSIDADE FEDERAL DE MINAS GERAIS

PROGRAMA DE PÓS-GRADUAÇÃO EM ESTATÍSTICA

UFMG

ATA DA DEFESA DA DISSERTAÇÃO DA ALUNA THAÍS PACHECO MENEZES

Realizou-se, no dia 28 de fevereiro de 2020, às 10:00 horas, 2040 ICEX, da Universidade Federal de Minas Gerais, a 253ª defesa de dissertação, intitulada *Arquétipos latentes: os módulos dos padrões espaciais complexos do câncer*, apresentada por THAÍS PACHECO MENEZES, número de registro 2018665060, graduada no curso de ESTATÍSTICA, como requisito parcial para a obtenção do grau de Mestre em ESTATÍSTICA, à seguinte Comissão Examinadora: Prof(a). Marcos Oliveira Prates - Orientador (DEST/UFMG), Prof(a). Renato Martins Assuncao - Coorientador (DCC/UFMG), Prof(a). Wagner Hugo Bonat (UFPR), Prof(a). Vinícius Diniz Mayrink (DEST/UFMG).

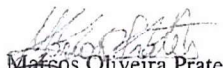
A Comissão considerou a dissertação:

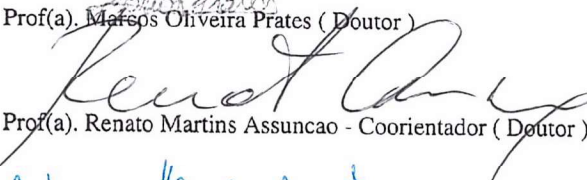
Aprovada


Reprovada

Finalizados os trabalhos, lavrei a presente ata que, lida e aprovada, vai assinada por mim e pelos membros da Comissão.

Belo Horizonte, 28 de fevereiro de 2020.


Prof(a). Marcos Oliveira Prates (Doutor)


Prof(a). Renato Martins Assuncao - Coorientador (Doutor)


Prof(a). Wagner Hugo Bonat (Doutor)


Prof(a). Vinícius Diniz Mayrink (Doutor)

Agradecimentos

Gostaria de agradecer ao suporte financeiro das agências de fomento CAPES, CNPq e FAPEMIG que auxiliam na infraestrutura do programa e facilitaram a realização dessa dissertação.

Resumo

Estudos sobre o comportamento epidemiológico de diversas doenças tem sido o foco de muitos pesquisadores que analisam os mapas de incidência e/ou de mortalidade em busca de padrões geográficos capazes de explicar o comportamento da doença em questão. Ao analisar regiões de maior taxa, é possível tirar conclusões que indiquem, por exemplo, que determinada doença está associada com alta poluição do ar devido à alta concentração de casos em áreas industriais. O problema nesse procedimento é que doenças como câncer possuem diversos tipos, o que torna seu estudo mais complicado pois, torna-se necessário a análise de diversos mapas individualmente. Com base nisso, o objetivo desse trabalho é propor um método capaz de reduzir a quantidade de mapas a serem analisados sem detrimento da análise realizada. A ideia base é usar a decomposição do valor singular para encontrar mapas latentes que são capazes de explicar os padrões geográficos de diversas doenças. Antes da aplicação do teorema da decomposição, serão feitos métodos de suavização Bayesiana além de que as taxas serão obtidas através da forma do SMR_i para retirar viés de faixa etária. Nesse trabalho, iremos analisar dados de mortalidade por câncer do Brasil e dos EUA e mostrar como essa redução de mapas pode ser feita, entendida e explorada. Como resultado, obtem-se uma eficácia de 90% na redução dos Estados Unidos e de 63,33% no caso brasileiro.

Palavras-chaves: Decomposição de Valores Singulares; estatísticas espaciais; fatoração de matrizes não negativas; mapas de câncer.

Abstract

Studies on the epidemiological behavior of several diseases have been the focus of many researchers who analyze incidence and/or mortality maps in search of geographic patterns capable of explaining the behavior of the disease in question. When analyzing regions with a higher rate, it is possible to draw conclusions that indicate, for example, that a certain disease is associated with high air pollution due to the high concentration of cases in industrial areas. The problem with this procedure is that diseases like cancer have different types, which makes their study more complicated because it is necessary to analyze several maps individually. Based on this, the objective of this work is to propose a method capable of reducing the amount of maps to be analyzed without the detriment of the analysis performed. The basic idea is to use the decomposition of the singular value to find latent maps that are able to explain the geographical patterns of different diseases. Before applying the decomposition theorem, Bayesian smoothing methods will be used and the rates will be obtained through the SMRi form to remove age-bias. In this work, we will analyze the cancer mortality data from Brazil and USA and show how this reduction of maps can be made, understood and explored. As a result, an efficiency of 90% is obtained in the reduction of the United States and of 63.33% in the Brazilian case.

Keywords: cancer maps; non-negative matrix factorization; Singular Value Decomposition; spatial statistics.

Lista de Tabelas

1	Quantidade de mapas bem explicados (percentil 90 do erro inferior ao ponto de corte considerado) para diferentes quantidades de mapas latentes utilizados - Dados EUA	24
2	Quantidade de mapas bem explicados para diferentes quantidades de mapas latentes utilizados	30
3	Quantidade de mapas que se tornaram bem explicados para diferentes quantidades de mapas latentes utilizados	36

Lista de Figuras

- 1 Esquema para visualização da aproximação via Decomposição do Valor Singular. A taxa da doença j de cada região do mapa observado é obtido como uma mistura do valor dessa mesma região em cada mapa latente, ponderada pelo peso do mapa e da doença em questão. Ressalta-se que o peso é igual para todas as regiões, mudando apenas quando considera outra doença. . . . 17
- 2 Gráfico de barras apresentando a morte total de cada um dos 30 cânceres de maior mortalidade dos EUA, ordenado do mais comum ao mais raro. . . . 20
- 3 Gráfico de barras apresentando a morte total de cada um dos 30 cânceres de maior mortalidade do Brasil, ordenado do mais comum ao mais raro. . . . 20
- 4 Curva de crescimento da quantidade de mapas bem explicados considerando diferentes níveis de cortes. O eixo vertical primário tem a quantidade bruta enquanto o secundário mostra a quantidade relativa. . . . 23
- 5 Percentil 90 do erro de aproximação dos mapas observados de cada câncer, ordenado do mais comum ao mais raro. . . . 24
- 6 Análise Espacial das regiões cujo o erro de aproximação foi superior ao ponto de corte definido 25
- 7 Figura da representação dos mapas latentes com seus respectivos pesos para cada câncer. As cores rosa representam os cânceres femininios enquanto as azuis os masculinos. Os gráficos estão ordenados pela sua importância dada a diagonal da matriz \mathbf{S} , aonde da esquerda para direita temos do mais importante ao menos importante. . . . 26
- 8 Comparação dos mapas reais e dos mapas aproximados considerando 3 mapas latentes. Nas representações, tem-se que a cor verde representa as menores taxas enquanto a vermelha indica maior risco e a amarela representa o risco intermediário. . . . 27
- 9 Curva de crescimento da quantidade de mapas bem explicados considerando diferentes níveis de cortes. O eixo vertical primário tem a quantidade bruta enquanto o secundário indica a quantidade relativa. . . . 29

10	Percentil 90 do erro de aproximação dos mapas observados de cada câncer, ordenado do mais comum ao mais raro.	30
11	Análise Espacial das regiões cujo o erro de aproximação foi superior ao ponto de corte definido	31
12	Figura da representação dos mapas latentes com seus respectivos pesos para cada câncer. As cores rosa representam os cânceres femininios enquanto as azuis os masculinos. Os gráficos estão ordenados pela sua importância dada a diagonal da matriz \mathbf{S} , aonde da esquerda para direita temos do mais importante ao menos importante.	32
13	Figura da representação dos mapas latentes com seus respectivos pesos para cada câncer. As cores rosa representam os cânceres femininios enquanto as azuis os masculinos. Os gráficos estão ordenados pela sua importância dada a diagonal da matriz \mathbf{S} , aonde da esquerda para direita temos do mais importante ao menos importante.	34
14	Comparação dos mapas reais e dos mapas aproximados considerando 11 mapas latentes. Nas representações, tem-se que as cor verde representa as menores taxas enquanto a vermelha indica maior risco e a amarela representa o risco intermediário.	35
15	Erro por microrregião na aproximação com 11 mapas latentes para os cânceres considerados mal aproximados.	37
16	Comparação dos mapas reais e dos mapas aproximados considerando 11 mapas latentes. Nas representações, tem-se que as cor verde representa as menores taxas enquanto a vermelha indica maior risco e a amarela representa o risco intermediário.	38

Sumário

1	Introdução	11
2	Metodologia	13
2.1	Preparação do Banco de Dados	13
2.1.1	Taxa de Mortalidade Padronizada (SMRi)	13
2.1.2	Taxa Bayesiana Empírica Global	14
2.2	Decomposição do Valor Singular	15
2.3	Medida de Similaridade Estatística	17
3	Dados Utilizados na Aplicação	19
4	Resultados	22
4.1	Dados dos Estados Unidos	22
4.2	Dados do Brasil	28
4.2.1	Análise dos mapas bem explicados	33
4.2.2	Análise dos mapas mal explicados	36
5	Conclusão	40
5.1	Trabalhos Futuros	41
	Referências	42

1 Introdução

Segundo definição do Instituto Nacional do Câncer (INCA), câncer é o nome dado a um conjunto de doenças cujo principal comportamento consiste no crescimento desordenado de células, que tendem a ser agressivas, incontroláveis e a invadir diversos sistemas, tecidos e órgãos do corpo humano. Em 2018, no Brasil, foram registrados 300.140 novos casos em homens e 282.450 em mulheres [INCA]. Em níveis mundiais, a Agência Internacional de Pesquisa de Câncer (IARC) apontou, em 2018, para o registro de 18,1 milhões de novos casos no mundo, além de 9,6 milhões de casos que resultaram em morte [IARC]. Por ser uma das principais causas de morte no mundo, o câncer em seus mais variados tipos têm sido alvo de estudos intensos. Outro ponto importante que contribui para o constante crescimento de estudos nesse tema é o fato diversos cânceres estarem associados a fatores de risco e, portanto, quanto maior a eficiência em se combater tais fatores, menor o surgimento de novos casos.

Uma grande dificuldade atual nesse cenário é que os mapas de incidência e/ou mortalidade de cada câncer, na maioria dos estudos, devem ser analisados individualmente. Para cada tipo, epidemiologistas precisam entender o padrão espacial e buscar explicações e métricas que sejam capazes de mostrar, por exemplo, o porquê que determinados cânceres são elevados em uma região enquanto outros possuem altas taxas em regiões completamente diferentes. Esse procedimento é demorado, trabalhoso e muitas vezes é inviável a comparação dos mapas, fatos que dificultam a proposição de medidas e programas eficazes no combate dessa doença além de atrapalhar a descoberta de possíveis relações entre os mais diversos tipos de câncer. Nesse contexto, metodologias de mineração de dados surgem como opções para esse processo de análise individual.

A abordagem multivariada em estudos de câncer é considerada em artigos científicos relevantes, que trabalham com modelos compartilhados cuja ideia base é: encontrar um efeito comum para diversas doenças. Knorr-Held e Best [7] introduziram o primeiro modelo de componentes compartilhados para a análise conjunta de 2 doenças. Futuramente, esse trabalho foi estendido para modelar conjuntamente a variação de diversas doenças [6]. Downing et al. [3] estudam a incidência de seis cânceres causados em decorrência do fumo em uma região da Inglaterra, sendo capazes de analisar os fatores de riscos comuns antes e depois de ajustar os

dados usando o *background* socioeconômico. Gómez-Rubio e Palmí-Perales [5] introduzem como utilizar a aproximação de Laplace integrada (INLA, Rue et al. 2009) em conjunto com passos de MCMC para fazer análises compartilhadas de diversas doenças. Cramb et al. [2] procuram por fatores comuns capazes de refletir os diversos riscos associados ao câncer de pulmão num contexto espaço-temporal.

O objetivo de tais modelos é analisar os dados de forma multivariada, porém ainda não são capazes de atender casos onde há muitas doenças, se limitando a uma quantidade inferior a 10 tipos diferentes. Tentando contornar essa limitação, Fernandes, Lopes e Assunção [4] iniciam um trabalho de análise multivariada com uma quantidade consideravelmente maior de doenças e é a base referencial principal desse trabalho.

A metodologia proposta nesse trabalho faz uso da Decomposição do Valor Singular [10] com o objetivo de encontrar mapas latentes que sejam capazes de explicar o comportamento de mais de uma doença simultaneamente. Assim, tem-se um avanço significativo nos estudos epidemiológicos, uma vez que o estudo individual dos mapas deixa de ser necessário e um estudo multivariado é apresentado, sendo capaz de focar a atenção em somente alguns poucos mapas para o entendimento de diversas doenças.

A decomposição do valor singular é um método que tem sido amplamente aplicado em trabalhos de *machine learning* [12, 1, 9]. Um problema típico de aprendizado de máquina pode ter várias centenas de variáveis, mas muitos algoritmos não são adequados se apresentados com mais do que algumas dúzias. Isto faz com que a decomposição do valor singular seja importante para fazer uma redução de dimensão nas variáveis a serem passadas aos algoritmos. Como mencionado, a proposta deste trabalho é encontrar mapas latentes bases que sejam capazes de explicar o comportamento espacial de diversas doenças simultaneamente.

Por fim, tem-se que o presente trabalho está organizado com a segunda seção apresentando a metodologia, capítulo que explora todos os passos de preparação do banco de dados, o teorema da decomposição e a estatística usada para medir a qualidade do ajuste. Em seguida, tem-se a terceira seção com os dados utilizados na aplicação e a quarta seção contendo os resultados. Ao final, tem-se a conclusão e as referências bibliográficas.

2 Metodologia

Nesta seção serão apresentados os passos de preparação do banco de dados, a metodologia de Decomposição do Valor Singular [10], além da medida de similaridade proposta para determinar a adequação da aproximação gerada.

2.1 Preparação do Banco de Dados

Antes de aplicar a metodologia proposta para encontrar mapas latentes capazes de resumir o comportamento de diversos cânceres simultaneamente é necessário pré-processar o banco de dados. Esse processo consiste em aplicar métodos capazes de suavizar a questão de alta variabilidade dos dados, aspecto capaz de dificultar a utilização dos métodos de redução de informação. Tais processos, ao analisarem taxas numericamente muito diferentes podem identificar como comportamentos completamente distintos, sendo que, na realidade, eles são geograficamente semelhantes e a diferença está, somente, em questões de escala.

2.1.1 Taxa de Mortalidade Padronizada (SMR_i)

No contexto epidemiológico, sabe-se que cada doença tem um comportamento diferenciado para cada faixa etária. Por exemplo, doenças como catapora e sarampo são mais comuns em crianças enquanto ataques cardíacos são consideravelmente maiores em adultos e idosos. Com isso, é importante levar em consideração a faixa etária na construção de mapas de doença para que as taxas calculadas reflitam o real risco da doença naquela área e não que espelhem uma realidade de maior mortalidade por ser, por exemplo, uma área onde há uma maior população afetada pela doença.

Nesse sentido, será realizado uma padronização indireta. Nesse processo, tem-se como pressuposto que o risco em cada faixa etária é constante no espaço, ou seja, independentemente da região considerada, o risco de uma determinada idade é o mesmo. Considerando a faixa etária, tem-se que a taxa de mortalidade da região i será definida como:

$$SMR_i = \frac{\text{Número de mortes na região } i}{\text{Número esperado de mortes na região } i} \quad (1)$$

onde o "Número esperado de mortes na região i " é calculado como um somatório do número

esperado de mortes de cada faixa etária para aquela região. Tal valor esperado, por sua vez, é dado por:

$$Esperado_{ij} = (\text{População da região } i \text{ na faixa } j) \times (TaxaNacional_j).$$

De modo que a taxa nacional é calculada por:

$$TaxaNacional_j = \frac{\text{Número de mortes na faixa } j}{\text{População na faixa } j}.$$

Por fim, tem-se que o SMR_i reflete quantas vezes o número de mortes observado foi maior (ou menor) do que o número esperado, sob risco constante. Sendo assim, a interpretação dessa taxa leva em conta que valores grandes indicam um risco maior naquela região visto que foram registradas uma quantidade de mortes consideravelmente maior do que o previsto. Como o valor esperado leva em consideração a faixa etária, tem-se que o risco refletido pelo SMR_i diz respeito sobre o risco real da doença naquela região, estando livre de efeitos da distribuição da população por faixa etária.

2.1.2 Taxa Bayesiana Empírica Global

O estimador Bayesiano Empírico Global [8] é usado para suavização das taxas observadas com o objetivo de diminuir a variabilidade dos dados para, assim, reduzir problemas de instabilidade estocástica. Em linhas gerais, supõe-se que a taxa real é uma variável aleatória de modo que a média é definida através da média global das taxas - que será usada para suavizar os valores registrados em cada região.

Seja E_i o número esperado de mortes da região i , O_i representando o número observado de mortes na região i e o SMR_i conforme definido em (1). Suponha também que $O_i|\theta_i$ segue uma distribuição Poisson com parâmetro $E_i\theta_i$. Nessa formulação, tem-se que θ_i representa o risco relativo à média global e é estimado pelo SMR_i . O problema é que, para pequenas populações, o SMR_i pode ter grande variação e, por isso, serão usados métodos de suavização. O estimador linear Bayesiano Empírico é dado por uma combinação linear entre a taxa

observada e a taxa média global:

$$\hat{\theta}_i = w_i \times SMR_i + (1 - w_i) \times RRG, \quad (2)$$

onde $RRG = \frac{\sum_{i=1}^n O_i}{\sum_{i=1}^n E_i}$ é o risco relativo global. O peso da combinação linear é dado por $w_i = \frac{Var(\theta_i)}{Var(\hat{\theta}_i)} = \frac{A}{A+RRG}$ em que $A = S^2 - \frac{RRG}{\sum_{i=1}^n \frac{E_i}{n}}$ e S^2 é a variância amostral.

Com esse processo, as regiões vão ter suas taxas re-estimadas através de uma média ponderada entre o valor médio global e as taxas observadas. É importante enfatizar que peso, w_i , é inversamente proporcional à população da região de modo que para o caso de populações reduzidas, a estimativa se aproxima mais da global. Realizando tal suavização, as taxas ficam menos sujeitas a flutuações causais não associadas com o risco, especialmente se o tamanho da população a ser considerado for pequeno.

2.2 Decomposição do Valor Singular

O teorema da decomposição do valor singular (SVD) [10] exprime uma maneira de realizar a decomposição de uma matriz. Definida a matriz \mathbf{X} de dimensão $n \times p$, onde n é a quantidade de regiões consideradas e p é o número de doenças, a SVD garante que tal matriz pode ser escrita em termos das matrizes \mathbf{U} e \mathbf{V} de dimensões $n \times p$ e $p \times p$, respectivamente. A relação entre tais matrizes é dada por:

$$\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^t \quad (3)$$

em que \mathbf{S} é uma matriz diagonal de dimensão $p \times p$. Nessa decomposição temos que a matriz \mathbf{U} agrupa os autovetores ortonormais da matriz $\mathbf{X}\mathbf{X}^t$ enquanto a matriz \mathbf{V} é responsável por reunir os autovetores ortonormais associados com os maiores autovalores da matriz $\mathbf{X}^t\mathbf{X}$. Por fim, tem-se que a raiz quadrada dos autovalores da matriz $\mathbf{X}\mathbf{X}^t$, que coincidem com os maiores autovalores de $\mathbf{X}^t\mathbf{X}$, se encontram agrupados de forma ordenada na diagonal da matriz \mathbf{S} . Em termos práticos, quando \mathbf{X} é a matriz de incidência por linha, para diversas doenças nas colunas, os mapas observados são reconstruídos como uma mistura dos mapas latentes agrupados na matriz \mathbf{U} com pesos dados pela multiplicação das matrizes \mathbf{S} e \mathbf{V}^t , que serão armazenados na matriz \mathbf{P} de dimensão $p \times p$. É importante ressaltar que, a ordenação

da matriz S implica que os mapas latentes agrupados na matriz U estão dispostos de acordo com sua importância onde a primeira coluna, por exemplo, contém o mapa latente de maior importância na composição dos valores presentes em X .

Quando a matriz S é composta por alguns poucos elementos relativamente grandes e os demais relativamente próximos a zero, é possível aproximar a matriz de dados X considerando somente partes das matrizes U e V . Seja k a quantidade de valores representativos da matriz S , definimos a matriz U_k como sendo a matriz de dimensão $n \times k$, a qual agrupa os k primeiros mapas latentes, e a matriz P_k de dimensão $k \times p$, que é obtida selecionando-se as k primeiras linhas da matriz P . Dessa forma, a matriz X pode ser aproximada por:

$$X \approx U_k \times P_k. \quad (4)$$

Com isso, tem-se que a taxa da doença j é dada por:

$$X^j \approx U_k \times P_k^j \quad (5)$$

onde P_k^j representa a coluna j da matriz P_k . Não usando a notação matricial, podemos escrever a aproximação como sendo:

$$X^j \approx P_1^j \times u_1 + \dots + P_k^j \times U_{jk}$$

Que é equivalente a:

$$X_j \approx \sigma_1 \times V_1^j \times U_1 + \dots + \sigma_k \times V_k^j \times U_k$$

onde σ_k é o k -ésimo elemento da diagonal da matriz S , V_k^j é a entrada k da coluna j de V e U_k é o k -ésimo mapa latente. Isto é, o padrão geográfico da doença j é aproximado através de uma combinação linear dos primeiros k padrões geográficos latentes. Em outras palavras, o mapa observado é aproximado como uma combinação de k mapas latentes de modo que, para cada mapa, consideramos a região i observada como sendo uma mistura dos valores na região i de cada mapa latente ponderados com diferentes pesos por mapa. A Figura 1 exemplifica o processo: considerando o mapa do Brasil divididos por estados, tem-se que a taxa observada de Minas Gerais será aproximada pelas taxas calculadas via decomposição do

valor singular nos k mapas latentes para Minas Gerais, ponderada por seus respectivos pesos gerados também via decomposição. Isso permite que os mapas observados sejam aproximados por k mapas onde $k \ll p$. Dessa forma, caso queria aproximar os valores do estado do Rio de Janeiro, deve-se misturar os valores dos mapas latentes relativos ao estado do Rio de Janeiro e podenderá-los pelo peso de cada mapa. Note, portanto, que o peso dado a cada mapa não depende da região, mas somente dos mapas latentes.

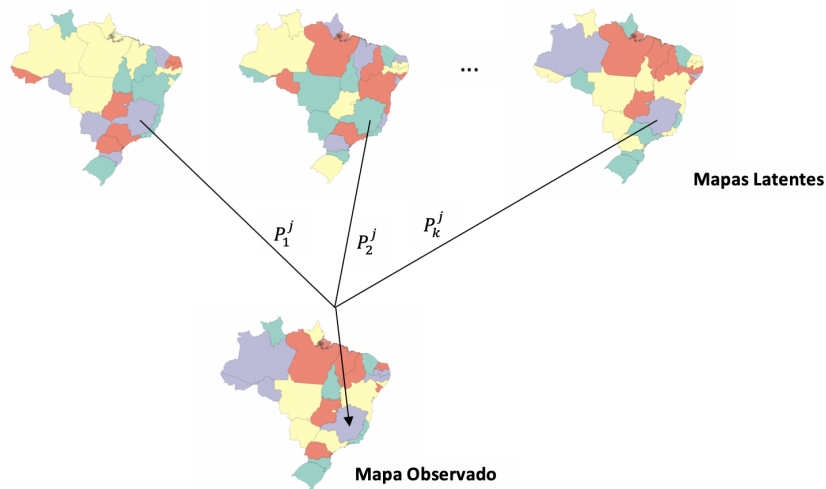


Figura 1: Esquema para visualização da aproximação via Decomposição do Valor Singular. A taxa da doença j de cada região do mapa observado é obtido como uma mistura do valor dessa mesma região em cada mapa latente, ponderada pelo peso do mapa e da doença em questão. Ressalta-se que o peso é igual para todas as regiões, mudando apenas quando considera outra doença.

2.3 Medida de Similaridade Estatística

É importante determinar uma medida que seja capaz de expressar o quão bem os mapas reais estão sendo aproximados pela método proposto em (3). Deseja-se que a aproximação das taxas feitas via combinação linear de k mapas latentes gere valores próximos dos valores reais observados. Sendo assim, considerando \mathbf{Y} como sendo a taxa aproximada e \mathbf{O} como sendo a taxa observada, espera-se que o modelo capaz de descrever a relação entre essas variáveis seja dado por $\mathbf{Y} = \mathbf{O}$ (cenário de aproximação perfeita).

Nesse sentido, definiu-se uma medida de erro capaz de mensurar o quão distante os

elementos de \mathbf{Y} se encontram de \mathbf{O} . Tal medida é calculada da seguinte forma: considera-se, para cada região, o valor absoluto da diferença entre o valor real e o aproximado e divide-se o resultado pelo valor real para que se fique em uma escala padronizada. Sendo assim, tem-se que para uma região i e uma doença j , o erro da aproximação é dado por:

$$e_i = \frac{|O_{ij} - Y_{ij}|}{O_{ij}}. \quad (6)$$

Considerando tal métrica, dizemos que um mapa é bem ajustado se 90% (percentil 90) de suas regiões possuírem um erro menor do que determinado ponto de corte. Sendo assim, definindo $E = (e_1, \dots, e_n)$ como o vetor de erros para cada região, onde os e_i são calculados como mostrado em (6), para a aproximação ser considerada adequada tem-se que:

$$P_{90}(E) \leq c$$

onde c é um ponto de corte a ser definido pelo especialista com base nos dados considerados no momento do ajuste.

3 Dados Utilizados na Aplicação

Em 2018, estudos da Agência Internacional de Pesquisa de Câncer (IARC) apontaram mais de 9 milhões de mortes causadas por câncer [IARC]. Essa doença, caracterizada pelo crescimento incontrolado e desordenado de células que se tornam agressivas e que se acoplam nas mais diversas partes do corpo humano, tem sido alvo de diversos estudos epidemiológicos. O objetivo é elaborar programas efetivos que freiem o expressivo crescimento das quantidades de novos casos e de mortalidade associados aos mais diversos tipos de cânceres.

Considerando a representatividade e importância desses estudos, a aplicação desse trabalho contará com duas bases de dados: os dados de mortalidade causada por cânceres no Brasil entre os anos 2000 e 2018 e nos Estados Unidos entre os anos de 1999 e 2017.

As informações dos dados brasileiros foram coletadas a nível de município e depois foram agrupadas em microrregiões seguindo as especificações adequadas. Os dados de mortalidade foram obtidos pela página do DataSUS (<ftp.datasus.gov.br/dissemin/publicos/SIHSUS>) enquanto que as quantidades populacionais foram encontradas no Censo do ano de 2010 do Instituto Brasileiro de Geografia e Estatística (IBGE) (<https://www.ibge.gov.br/estatisticas/sociais/populacao/9663-censo-demografico-2000.html?=&t=downloads>). No total, foram coletadas as informações de 452 cânceres para 557 microrregiões.

Em relação aos dados dos EUA, tem-se que a informação está a nível de condado e foram obtidas usando o "SEER Data & Software", ferramenta disponibilizada para acesso a informações epidemiológicas e também de quantidade populacional (<https://seer.cancer.gov/data-software/>). A base final é composta por 3107 condados e 82 cânceres.

Para ambos os casos, as taxas de mortalidade foram separadas por sexo de modo que, respeitando os cânceres exclusivos de mulheres e os exclusivos de homens, a base brasileira é composta por 862 cânceres separados entre sexo feminino e masculino enquanto a base norte-americana possui 151 doenças. Uma consequência de se trabalhar com essa elevada quantidade de cânceres é que muitos não terão um padrão espacial expressivo por se tratar de doenças com baixa taxa de mortalidade. Por esse motivo, optou-se por trabalhar, em ambas as bases, com os 30 cânceres mais comuns.

Antes da aplicação da metodologia proposta, uma análise descritiva dos cânceres mais

comuns de cada banco de dados é apresentada abaixo. Primeiramente, a Figura 2 contém o gráfico retratando a quantidade de mortes dos 30 cânceres mais comuns dos EUA.

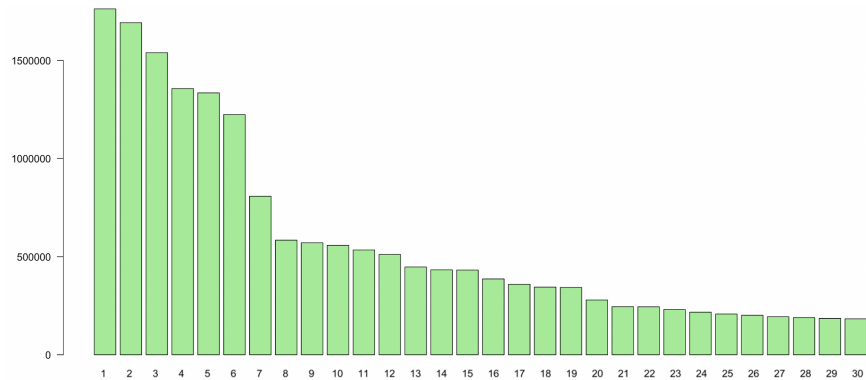


Figura 2: Gráfico de barras apresentando a morte total de cada um dos 30 cânceres de maior mortalidade dos EUA, ordenado do mais comum ao mais raro.

Como resultado, tem-se que a quantidade de mortalidade nos dados dos Estados Unidos estão entre 183.068 e 1.762.902 de casos. O "câncer do sistema respiratório em homens" é o mais comum da base dos USA e o "Linfoma Não-Hodgkin em mulheres" é o mais raro.

Do mesmo modo, a Figura 3 contém a representação visual da quantidade de mortalidade dos 30 cânceres considerados nos dados brasileiros.

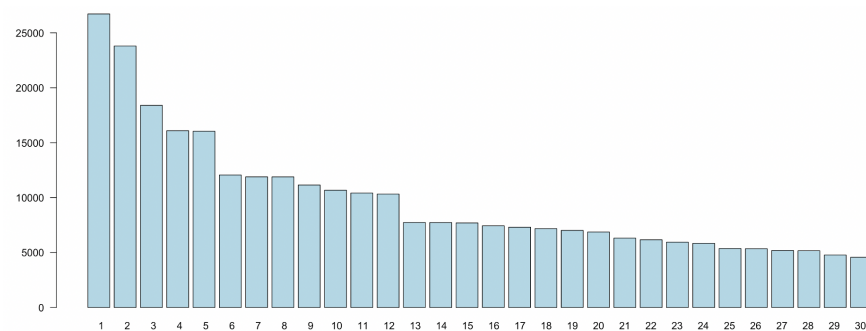


Figura 3: Gráfico de barras apresentando a morte total de cada um dos 30 cânceres de maior mortalidade do Brasil, ordenado do mais comum ao mais raro.

Analisando os dados, tem-se que as mortalidades para os dados brasileiros estão entre 4.571 e 26.718 casos sendo o "câncer dos brônquios ou pulmões, não especificado, em mulheres" o de maior mortalidade enquanto o "câncer de brônquios ou pulmões, com lesão invasiva, em homens" é o de menor.

Comparando as quantidades da mortalidade de cada um dos bancos de dados, observa-se que os dados dos Estados Unidos estão em uma escala consideravelmente maior do que os dados do Brasil. Tal resultado é explicado pela diferença na estruturação dos dados. Nos dados brasileiros, a descrição dos cânceres observados é feita de modo muito mais detalhado tendo assim, uma granularidade maior no banco de dados. Para exemplificar essa diferença, observa-se que o câncer de brônquios e de pulmão é dividido, no Brasil, em "Câncer do Brônquio Principal", "Câncer do Lobo Superior, Brônquio ou Pulmão", "Câncer do Lobo Médio, Brônquio ou Pulmão", "Câncer do Lobo Inferior, Brônquio ou Pulmão", "Câncer dos Brônquios e dos Pulmões com Lesão Invasiva" e "Câncer dos Brônquios ou Pulmões, não especificado", enquanto para a base norte americana, tal câncer é apresentado como "Câncer do Sistema Respiratório" ou como "Câncer de Pulmão e Brônquio". Para o presente trabalho, optou-se por manter essa diferença ao invés de atotar uma padronização nos dois bancos de dados. Entende-se que a estrutura de dados é desenhada por cada sistema de saúde do modo a melhor atender o seu país e, portanto, para manter os resultados representando como os cânceres são tratados em cada sistema, não se realizou nenhum tipo de agrupamento nas mortalidades observadas.

4 Resultados

Essa seção apresenta os resultados da aplicação da metodologia para os dados de mortalidade por câncer dos EUA e do Brasil.

4.1 Dados dos Estados Unidos

Os Estados Unidos, desconsiderando Alaska e Havaí, estão divididos em 3107 condados de modo que, para ajuste da metodologia, os dados foram estruturados como uma matriz de dimensão 3107×30 onde cada linha representa um condado e cada coluna um tipo de câncer. A padronização indireta foi feita e, portanto, as taxas passaram a ser representadas pelo SMRi descrito em (1). Além disso, quaisquer possíveis instabilidades estocásticas foram suavizadas usando a abordagem Bayesiana Empírica Global descrita em (2).

Aplicando a decomposição do valor singular apresentada em (3), foram encontradas as matrizes \mathbf{U} e \mathbf{P} . De posse dos mapas latentes e da matriz de peso, calculou-se a aproximação apresentada em (4) variando k de 1 até 30. Então, encontrou-se, então, o erro da aproximação conforme demonstrado em (6) e, para comparar o desempenho, considerou-se diferentes pontos de corte c para determinar se o mapa é considerado bem explicado. Cabe ressaltar que a aproximação é considerada boa se 90% das regiões possuem um erro menor do que o corte considerado de modo que quanto menor esse corte, mais restritivo é. A análise em questão foi feita para os diferentes cortes $c = 25\%, 20\%, 15\%$ e 10% e os resultados são apresentados na Figura 4.

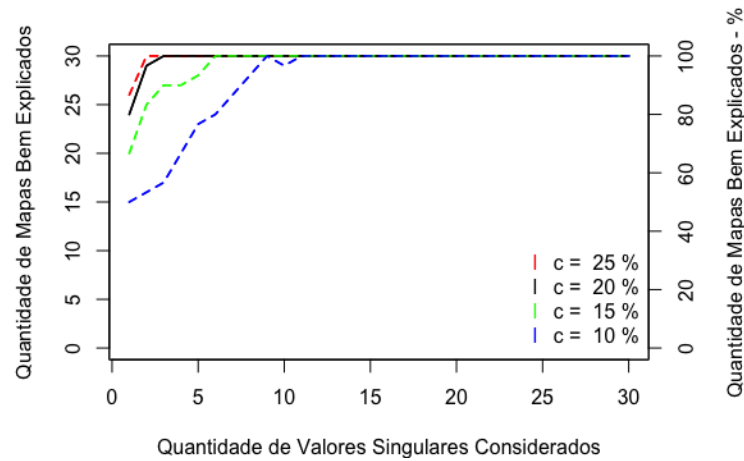


Figura 4: Curva de crescimento da quantidade de mapas bem explicados considerando diferentes níveis de cortes. O eixo vertical primário tem a quantidade bruta enquanto o secundário mostra a quantidade relativa.

A Figura 4 mostra que, mesmo considerando os níveis mais restritivos, a maioria dos cânceres ficam bem aproximados na presença de poucos mapas latentes, de modo que no cenário com o erro máximo tolerado sendo de 10% precisa-se de 11 mapas latentes para se ter todos os cânceres sendo considerados como bem explicados. Além disso, observa-se que, para todos os cortes, há uma quantidade considerável de mapas com um bom ajuste quando se tem somente 1 mapa latente.

Em uma análise menos restritiva, se uma aproximação é classificada como boa quando se tem 90% das regiões com um erro de no máximo 20%, são necessários 3 mapas latentes para que as 30 doenças fiquem bem explicadas, valor que representa uma redução de 90% na quantidade de mapas a serem estudados. No presente estudo, optou-se por trabalhar com esse corte, e assim, tem-se que cada doença vai ser considerada como bem explicada somente se no máximo 10% dos condados registrarem um erro de aproximação maior que 20%. Para esse cenário, a Tabela 1 mostra a quantidade de mapas bem explicados para diferentes quantidades de mapas latentes considerados.

Quantidade de Mapas Latentes	1	2	3
Quantidade de Mapas Bem Explicados	24	29	30
% de Mapas Bem Explicados	80.00	96.67	100.00

Tabela 1: Quantidade de mapas bem explicados (percentil 90 do erro inferior ao ponto de corte considerado) para diferentes quantidades de mapas latentes utilizados - Dados EUA

Como resultado da Tabela 1, tem-se que com apenas 1 mapa latente, 80% das doenças podem ser consideradas bem explicadas. A Figura 5 mostra o percentil 90 do erro de aproximação de cada câncer para diversas quantidades de mapas latentes usados na aproximação. No gráfico, o eixo x tem índice dos cânceres, onde o primeiro ponto é o câncer mais comum e o último o mais raro.

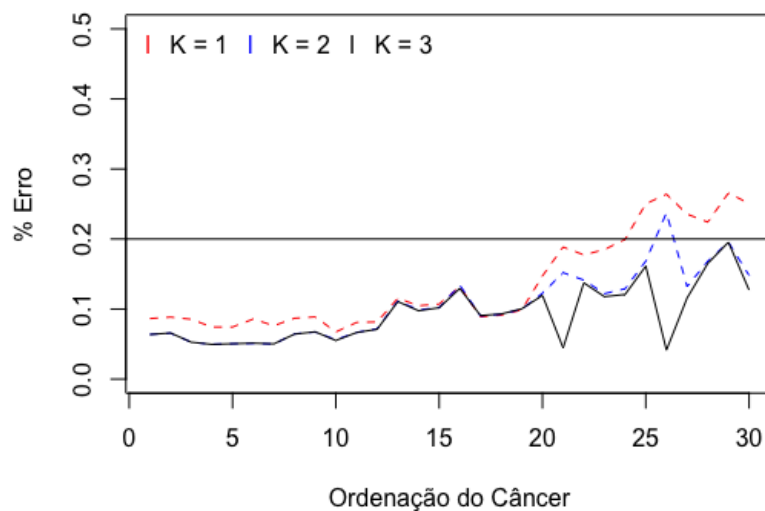


Figura 5: Percentil 90 do erro de aproximação dos mapas observados de cada câncer, ordenado do mais comum ao mais raro.

Observa-se que os cânceres mais raros desse banco de dados foram os que precisaram de mais mapas latentes para se ter 90% dos condados com erros de aproximação de no máximo 20%. Os cânceres mais comuns se enquadraram no cenário de boa aproximação quando se usou somente 1 mapa latente.

Os resultados apresentados mostram, então, que a Decomposição do Valor Singular foi capaz de gerar uma redução na quantidade de mapas a serem estudados, visto que 30 tipos

de cânceres estão sendo bem aproximados por 3 mapas latentes gerados via decomposição. Ressalta-se que considerar 3 mapas latentes quer dizer que, em (3), tem-se que $k = 3$, ou seja, os mapas observados serão aproximados através da multiplicação matricial dos 3 primeiros mapas latentes, que estão agrupados nas 3 primeiras colunas da matriz \mathbf{U} , com seus respectivos pesos, encontrados nas 3 primeiras linhas da matriz \mathbf{P} .

Antes de realizar uma análise visual dos mapas gerados e do resultado da aproximação, a Figura 6 mostra a dispersão das regiões cujo o erro de aproximação foi maior que o ponto de corte considerado para o câncer mais comum, um intermediário e o mais raro desse banco de dados. Cabe mencionar que a métrica de qualidade de ajuste permite que somente no máximo 10% dos condados se enquadrem nesse cenário. Na imagem, a cor vermelha indica que a região em questão ultrapassou o erro máximo aceitável.

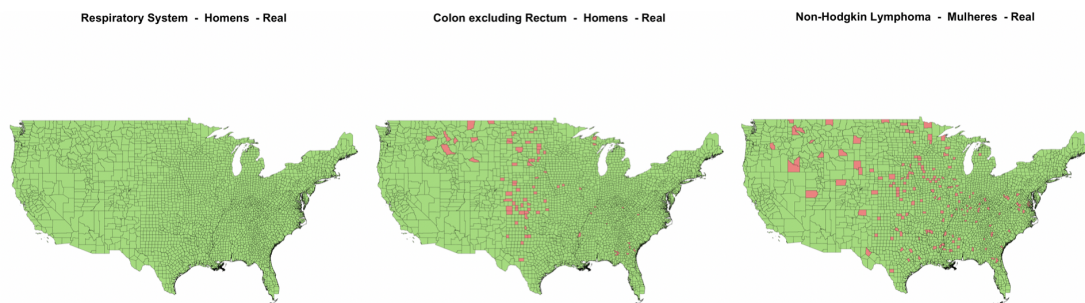


Figura 6: Análise Espacial das regiões cujo o erro de aproximação foi superior ao ponto de corte definido

Analisando os mapas da Figura 6, tem-se que o câncer mais comum não teve nenhum condado com um erro maior do que o aceitável. Para o câncer intermediário e para o mais raro, é possível observar poucas regiões marcadas de vermelho, mas não há nenhuma aglomeração específica ou padrão espacial que explique o pior ajuste de tais regiões.

Partindo desses resultados, a Figura 7 mostra a análise visual dos mapas latentes encontrados. Ao redor, tem-se o histograma circular do peso do mapa para cada um dos 30 cânceres considerados. Além disso, tem-se que todas as barras estão na mesma escala e as barras azuis representam os casos em homens enquanto a rosa indica o câncer registrado em mulheres.

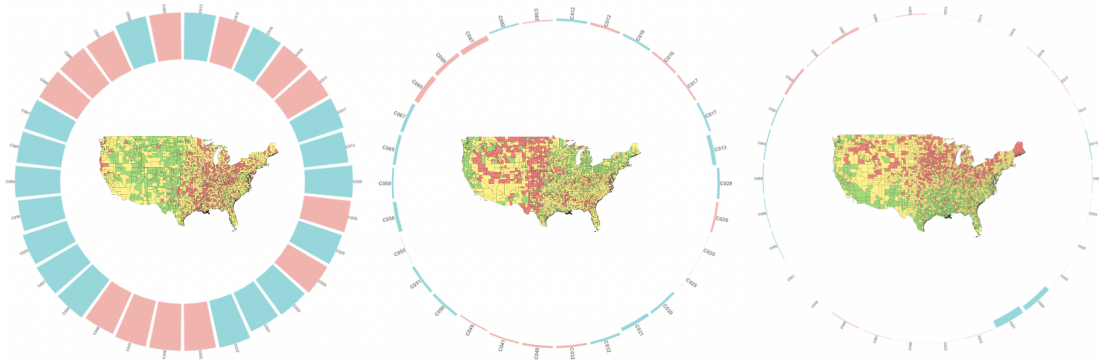


Figura 7: Figura da representação dos mapas latentes com seus respectivos pesos para cada câncer. As cores rosa representam os cânceres femininios enquanto as azuis os masculinos. Os gráficos estão ordenados pela sua importância dada a diagonal da matriz \mathbf{S} , aonde da esquerda para direita temos do mais importante ao menos importante.

Para analisar o resultado, é importante mencionar que um peso negativo indica que a direção das taxas é a oposta da representada no mapa latente, ou seja, para um peso negativo a área do mapa latente com maior taxa se tornará a de menor taxa assim como a região de menor taxa será a de maior concentração. Dito isso, tem-se que o primeiro mapa latente é de maior representatividade para todas as doenças de modo que os pesos associados a ele estão entre 52,18 e 58,75. Tal mapa latente retrata um cenário onde as menores taxas de mortalidade estão concentradas na metade da esquerda do mapa além de que na parte mais à direita observa-se uma mistura das taxas altas, médias e baixas, mas com maior concentração das maiores taxas.

O segundo e terceiro mapa latente, apesar de registrarem pesos consideravelmente menores em relação ao primeiro mapa, acabam apresentando uma representatividade maior para determinadas doenças. Para o segundo mapa, os pesos variam de 4,87 a 2,41 enquanto para o terceiro o intervalo é de $-7,05$ a 2,24. No que tange à análise espacial, o segundo mapa latente é marcado por um padrão de taxas mais altas na região central, seguido por uma concentração de baixas taxas na região Noroeste. Além disso, a costa à esquerda também é marcada pelo cenário de menores mortalidades. Já o terceiro mapa latente, tem como característica a predominância de taxas intermediárias e altas nas regiões centro-oeste e leste dos EUA. Na região sul, tem-se que a maior parte dos condados são marcados por menores taxas enquanto a região oeste tem prevalência de taxas intermediárias.

A análise dos mapas latentes fornece uma ideia base da caracterização da distribuição espacial da mortalidade por câncer nos Estados Unidos. Porém, uma análise comparativa do resultado da aproximação é necessária para ser possível concluir sobre a qualidade do ajuste. Nesse sentido, a Figura 8 mostra o confronto visual entre os mapas reais e os mapas aproximados para alguns cânceres. Na imagem, os gráficos à direita são os gráficos reais enquanto os da esquerda são os gerados via aproximação considerando $k = 3$. Além disso, o câncer na parte de cima é o mais comum, o da parte mais inferior é o mais raro e o do meio é um com taxas intermediárias.



Figura 8: Comparação dos mapas reais e dos mapas aproximados considerando 3 mapas latentes. Nas representações, tem-se que as cor verde representa as menores taxas enquanto a vermelha indica maior risco e a amarela representa o risco intermediário.

Na Figura 8, os confrontos dos mapas reais com os mapas aproximados confirmam a hipótese de que é possível aproximar 30 mapas de mortalidade utilizando poucos mapas latentes. Para o câncer do sistema respiratório em homens (câncer mais comum), tem-se que o mapa aproximado estimou taxas intermediárias para a região oeste do país onde, no real, são taxas menores. Além disso, alguns dos condados de alta mortalidade na região sul e na parte mais ao sul do centro-oeste também foram classificados como intermediários. Apesar de tais diferenças, o mapa aproximado entregou um bom ajuste. Já na análise do câncer mais intermediário, o câncer de colón excluindo o reto, para homens, tem-se que a quantidade de regiões com altas taxas que passaram a ser classificadas como intermediárias aumentou em relação à primeira comparação, de modo que no mapa aproximado há poucos condados na região sul classificados como uma área de risco maior. Apesar disso, o comportamento da região oeste ficou bem aproximado assim como na parte central do país. Por fim, o mapa de mortalidade do Linfoma Não Hodgkin apresentou boa aproximação visual. As taxas aproximadas das regiões nordeste e centro oeste estão semelhantes com os valores registrados no mapa real, além do meio do mapa também estar parecido. Na região Sul houve uma diferença no registro de alguns pontos de alta taxa, mas no geral o comportamento está próximo do real. A região oeste é a que teve maior diferença visto que no mapa aproximado a maior parte da região apresentou taxas intermediárias subestimando alguns pontos de alta incidência e superestimando alguns condados de menores taxas.

Com base em todos os resultados apresentados, tem-se que, considerando os 30 cânceres de maiores mortalidades nos Estados Unidos, a metodologia apresentada encontrou que 3 mapas latentes são capazes de explicar o comportamento de tais doenças. Essa descoberta gera uma redução de 90% na quantidade de mapas a serem estudados por epidemiologistas, de modo que, focando a atenção nos 3 mapas latentes bases é possível entender o comportamento dos 30 cânceres considerados.

4.2 Dados do Brasil

Para se testar a real eficiência da metodologia proposta, considerou-se também os dados de mortalidade de câncer no Brasil. Dividido por microrregião, os dados coletados foram agrupados em uma matriz \mathbf{X} de dimensão 557×30 de modo que cada linha é uma microrregião

e cada coluna é um tipo de câncer. Assim como nos dados dos EUA, a padronização indireta (1) e a abordagem Bayesiana Empírica Global (2) foram aplicadas para preparar o banco de dados para ajuste da Decomposição do Valor Singular (3).

Com o ajuste da decomposição, encontrou-se os mapas latentes e os pesos associados a eles para os dados brasileiros. Do mesmo modo que apresentado em (6), os erros de aproximação foram calculados para quando se usa de 1 até 30 mapas latentes na aproximação. Considerando como métrica de qualidade de ajuste um mapa sendo classificado como bem explicado se 90% das regiões possuem um erro menor que um ponto de corte, a Figura 9 mostra o desempenho da aproximação ao apresentar a curva de crescimento da quantidade de mapas bem explicados para os diferentes cortes apresentados na Seção 4.1.

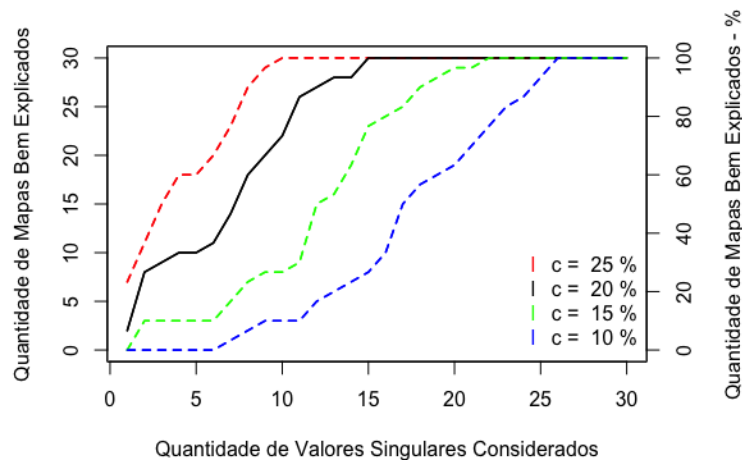


Figura 9: Curva de crescimento da quantidade de mapas bem explicados considerando diferentes níveis de cortes. O eixo vertical primário tem a quantidade bruta enquanto o secundário indica a quantidade relativa.

A Figura 9 mostra um crescimento mais lento na quantidade de mapas bem explicados, comportamento mais ressaltado nos cenários mais restritivos (10% e 15%). Para os cortes que aceitam um erro maior (25% e 20%), tem-se que 90% das doenças ficam bem ajustadas quando se considera 8 ou 11 mapas latentes. Mantendo o mesmo critério usado na análise dos resultados da base norte-americana, os mapas aproximados serão considerados próximos do real quando no máximo 10% das microrregiões possuírem um erro de aproximação superior a 20%. Para tal critério, a Tabela 2 mostra a quantidade de mapas bem explicados para

diferentes quantidades de mapas latentes considerados.

Quantidade de Mapas Latentes	1	2	5	8	11	15
Quantidade de Mapas Bem Explicados	2	8	10	18	26	30
% de Mapas Bem Explicados	6.67	26.67	33.33	60.00	90.00	100.00

Tabela 2: Quantidade de mapas bem explicados para diferentes quantidades de mapas latentes utilizados

A análise da Tabela 2 mostra que com 1 mapa latente somente 6.67% dos mapas podendo ser considerados bem explicados. Com 11 mapas latentes tem-se 90% com um bom ajuste, valor que gera uma redução de 63.33% na quantidade de mapas a serem estudados. Quando se tem 15 mapas latentes, todas as doenças podem ser consideradas bem aproximadas sendo que, essa quantidade representa uma redução de 50% no total de mapas que devem ser analisados. Apesar de precisar de mais mapas latentes, o cenário de redução na quantidade de mapas a serem estudados segue expressiva.

Uma outra análise no que diz respeito à qualidade do ajuste está apresentada na Figura 10, que apresenta o percentil 90 do erro para cada câncer, usando diferentes quantidades de mapas latentes na aproximação.

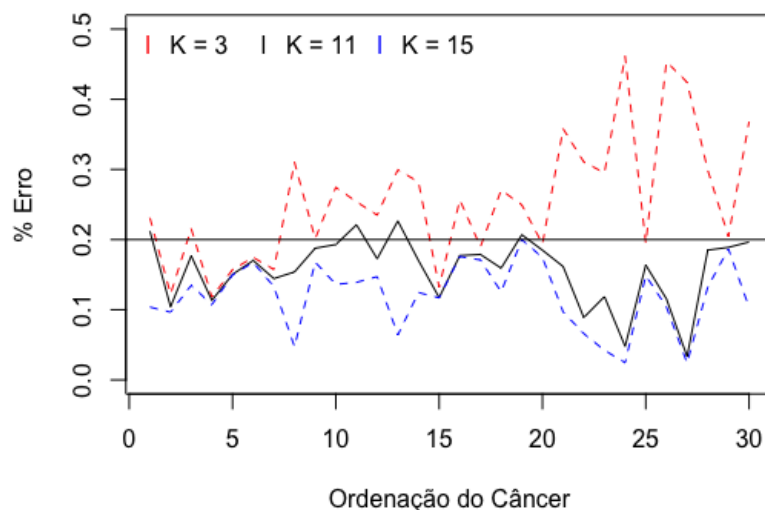


Figura 10: Percentil 90 do erro de aproximação dos mapas observados de cada câncer, ordenado do mais comum ao mais raro.

Como resultado, observa-se que os cânceres mais raros desse banco de dados apresentam erros mais elevados no cenário de menos mapas latentes, mas ficam bem ajustados com 11 mapas. Além disso, observa-se que o câncer mais comum desse banco de dados, juntamente com 3 cânceres com uma mortalidade mais intermediária, só passam a ser considerados bem explicados, nessa comparação, no cenário que considera 15 mapas latentes para realizar a aproximação.

Analisando os resultados até então apresentados, optou-se por trabalhar com o uso de 11 mapas latentes na realização da aproximação. Em tal cenário, há uma expressiva redução da quantidade de mapas a serem analisados e a maioria dos cânceres (90%) ficam bem explicados. Com esse valor, tem-se que apenas 4 mapas não se enquadram na classificação de boa aproximação.

Para esse cenário, a Figura 11 mostra a distribuição das microrregiões com erros superiores ao corte definido para três exemplos de cânceres do banco de dados (o câncer mais comum, um intermediário e o mais raro).

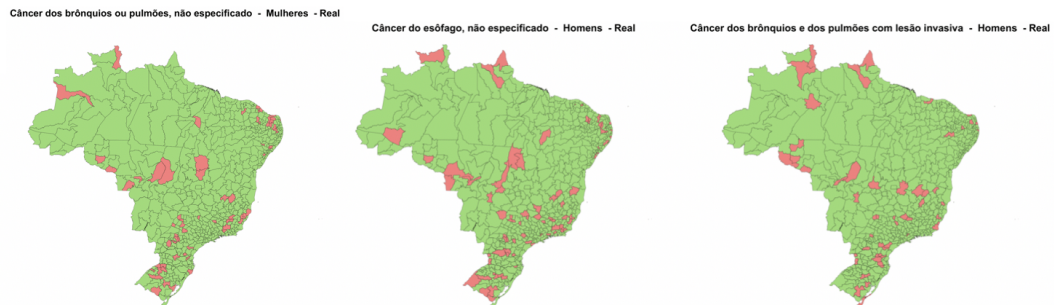


Figura 11: Análise Espacial das regiões cujo o erro de aproximação foi superior ao ponto de corte definido

Os gráficos mostram que não há um padrão espacial tão definido nas microrregiões com maior erro de aproximação. Além disso, observa-se um comportamento mais espalhado, de modo que os mapas não refletem uma possível clusterização das regiões mal ajustadas.

Uma análise visual desses mapas latentes é apresentada na Figura 12. Na imagem, há a representação circular do histograma do peso do referido mapa latente. Além disso, tem-se que as barras azuis se referem aos casos masculinos enquanto a rosa representa os femininos. Por fim, ressalta-se que todas as barras estão na mesma escala e tem-se que o cenário de

elevadas taxas do SMR_i é representado pela cor vermelha seguida da amarela até chegar na verde, que representa menores taxas.

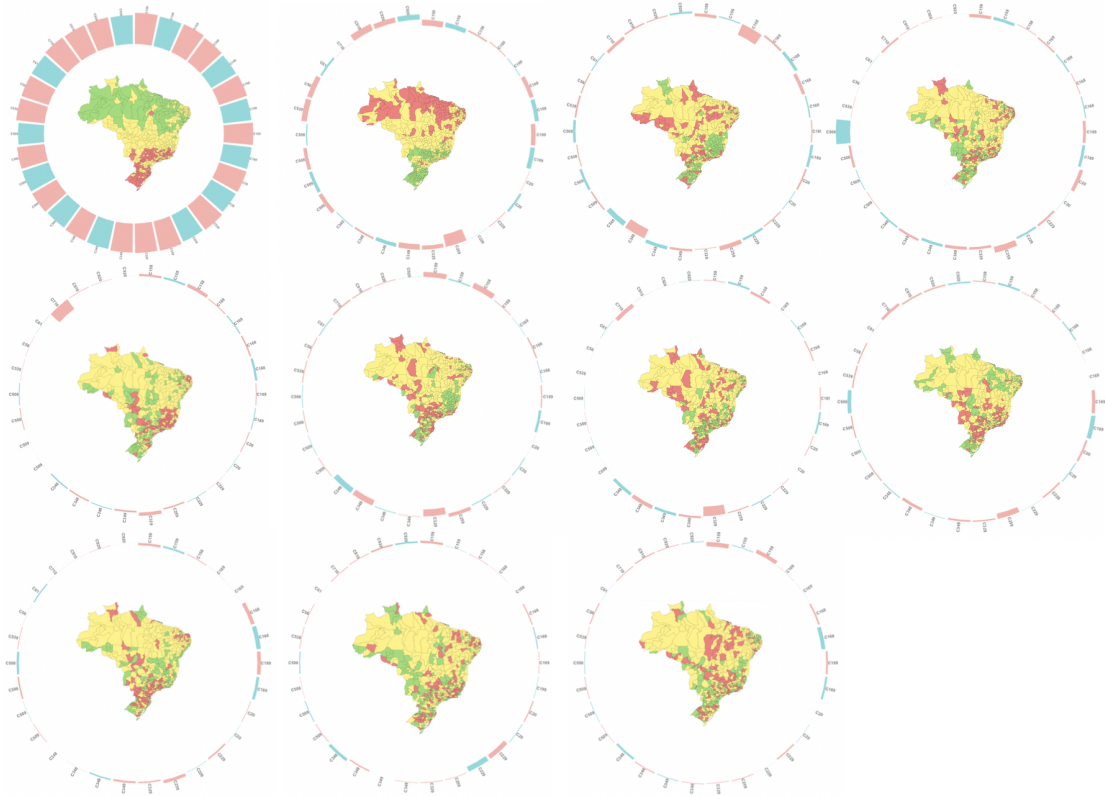


Figura 12: Figura da representação dos mapas latentes com seus respectivos pesos para cada câncer. As cores rosa representam os cânceres femininos enquanto as azuis os masculinos. Os gráficos estão ordenados pela sua importância dada a diagonal da matriz \mathbf{S} , aonde da esquerda para direita temos do mais importante ao menos importante.

Como resultado, os três mapas mais representativos foram analisados. Cabe ressaltar que a importância é ordenada pelo autovalor da matriz \mathbf{S} gerada em (3). Portanto, tem-se que o primeiro mapa latente é o de maior representatividade para todas as doenças onde os pesos para tal mapa estão entre 20,04 e 24,21. Tal mapa retrata um cenário de menor mortalidade na parte mais ao norte do país enquanto o Sul e o Sudeste apresentam predominância de maiores taxas de mortalidade. As regiões mais centrais registram valores intermediários.

Analisando o segundo mapa latente, tem-se que as altas concentrações estão agrupadas à direita da região Norte e mais ao norte dos estados do Pará, do Amapá e da Região Nordeste. As regiões Sul e Sudeste são caracterizadas por menores taxas. Por fim, tem-se que o terceiro mapa latente representa um padrão de menores taxas na Região Sul do país e no estado

de Minas Gerais. Além disso, tem-se que a maioria das microrregiões registraram taxas intermediárias de modo que as altas mortalidades estão mais dispersas em todo o país, com maior aglomeração no Acre e em Rondônia.

No que tange aos pesos, é possível ver que os mapas 2 e 3 possuem uma maior representatividade para algumas doenças, mas, de modo geral, esses mapas possuem pesos menores. O câncer de pâncreas sem demais especificações em mulheres, por exemplo, possui um peso maior no mapa latente 2 enquanto o mapa latente 3 tem uma relevância destacada para o câncer de esôfago com lesão invasiva em mulheres. Os demais mapas latentes também apresentam pesos menores do que o primeiro mapa, registrando um valor maior para uma ou outra doença.

Realizando essa aproximação com $k = 11$, tem-se que 26 mapas são considerados bem explicados e 4 se enquadram no cenário de mal explicados. Com isso, decidiu-se estudar os mapas cuja aproximação foi boa separadamente daqueles cujos resultados não foram tão satisfatórios.

4.2.1 Análise dos mapas bem explicados

Separando somente as 26 doenças que foram considerados bem explicadas seguindo os critérios estabelecidos, ajustou-se a decomposição do valor singular para tais casos separadamente. Nesse cenário, estudou-se os mapas latentes gerados pela decomposição que considera somente os mapas bem explicados. A Figura 13 apresenta a análise visual dos resultados. Do mesmo modo que no caso geral, tem-se que as barras no entorno representam o peso daquele mapa para cada uma das doenças. Além disso, a cor vermelha representa as maiores taxas enquanto a verde simboliza as menores.

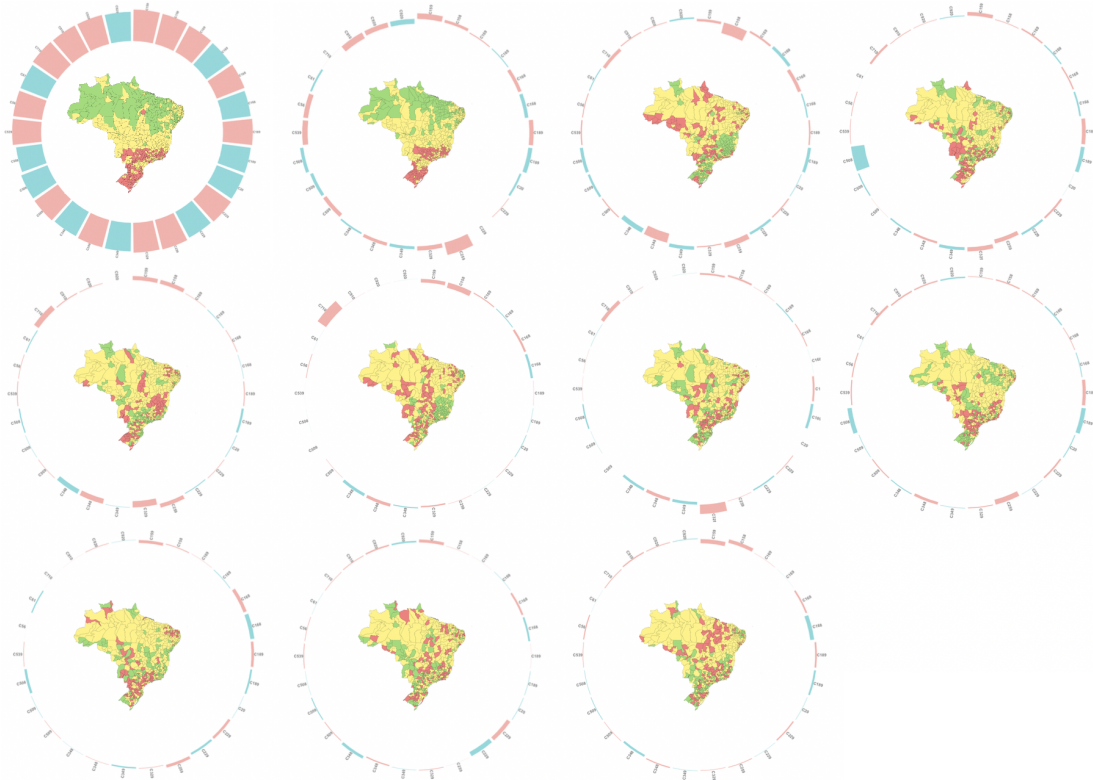


Figura 13: Figura da representação dos mapas latentes com seus respectivos pesos para cada câncer. As cores rosa representam os cânceres femininios enquanto as azuis os masculinos. Os gráficos estão ordenados pela sua importância dada a diagonal da matriz S , aonde da esquerda para direita temos do mais importante ao menos importante.

Analisando os pesos associados a cada um dos mapas latentes em questão, tem-se que o primeiro mapa latente apresenta pesos registrados entre 20,03 e 24,18. Já os demais mapas latentes, apresentaram pesos menores, com a maioria dos valores distribuídos entre -1 e 1 , mas com algumas doenças se destacando com pesos de 7.86 ou -8.03 , por exemplo.

Em questões visuais, observa-se que os mapas latentes apresentam distribuições espaciais semelhantes às geradas quando se considera todas as 30 doenças, tendo alteração na cor de algumas microrregiões. O primeiro mapa, por exemplo, tem o mesmo comportamento de baixas taxas na parte ao norte do país enquanto as regiões sul e sudeste registram maiores mortalidades. Já o segundo mapa, porém, inverteu o comportamento em relação ao mapa latente geral de modo que nas regiões Norte e Nordeste, houveram registros de baixa mortalidade enquanto a região Sul e o sul da região Sudeste foram marcadas com elevadas mortalidades. Por fim, o terceiro mapa, por exemplo, registra menores mortalidades concen-

tradas em Minas Gerais e ao Sul do País, enquanto nas demais regiões há registro de taxas intermediárias e algumas taxas altas espalhadas.

Para concluir as análises dos mapas que foram considerados bem explicados, vamos comparar os mapas gerados via aproximação com os reais observados. A Figura 14 contém na parte superior os resultados para o câncer mais comum desse grupo e na parte inferior o confronto para o mais raro. O gráfico no meio é uma doença de taxa de mortalidade intermediária.

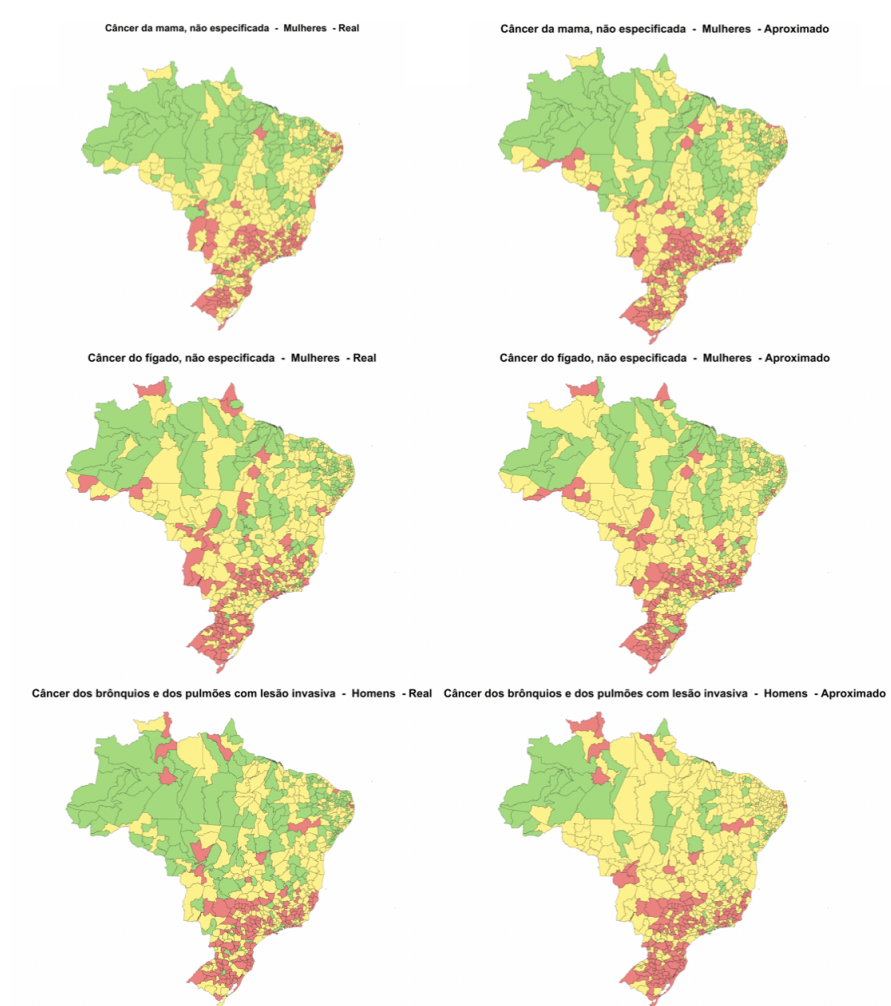


Figura 14: Comparação dos mapas reais e dos mapas aproximados considerando 11 mapas latentes. Nas representações, tem-se que as cor verde representa as menores taxas enquanto a vermelha indica maior risco e a amarela representa o risco intermediário.

Confrontando os mapas, pode-se observar que o mapa aproximado é similar ao mapa

real em todos os casos. Para o câncer mais comum, vemos que o padrão de altas taxas ao sul do país e ao sul da região Sudeste são captados pela aproximação, além da correta representação de menores taxas nas regiões Norte e Nordeste. Para esse câncer, o mapa real e o aproximado estão bastante parecidos de modo que a aproximação deixou de captar corretamente o padrão de poucas microrregiões. Para o câncer mais raro, é possível observar que houve uma diferença ligeiramente maior na aproximação. Na região Sul, o mapa aproximado captou o cenário de altas mortalidades, mas para as regiões Centro-Oeste, Sudeste e Nordeste enquanto o real tem predominância de taxas baixas, o mapa aproximado atribuiu taxas intermediárias. Cabe ressaltar que a escala de tal mapa é muito sensível e a mudança da taxa mais baixa para a intermediária é na ordem de 0.2. Por fim, o câncer intermediário também se enquadra no quadro de boa aproximação. É possível ver que há apenas algumas microrregiões de elevadas taxas na parte mais interna da região Centro-Oeste que não foram corretamente representadas na aproximação.

Com base em todas as considerações feitas, conclui-se que os mapas que se mostraram bem explicados de fato possuem um comportamento semelhante. Uma análise prática mostra que as altas mortalidades por câncer estão concentradas nas regiões sul e na parte sul das regiões sudestes e centro-oeste.

4.2.2 Análise dos mapas mal explicados

Considerando somente os 4 cânceres que ficaram mal explicados, optou-se por ajustar a decomposição do valor singular também para eles separadamente. Como resultado, a Tabela 3 mostra que, sozinhos, os mapas que foram considerados mal explicados precisam de 2 mapas latentes para se tornarem bem explicados.

Quantidade de Mapas Latentes	1	2
Quantidade de Mapas Bem Explicados	1	4
% de Mapas Bem Explicados	20	100

Tabela 3: Quantidade de mapas que se tornaram bem explicados para diferentes quantidades de mapas latentes utilizados

Um dos objetivos de estudar separadamente o grupo de mapas mal explicados é tentar descobrir "onde" a falta de ajuste está concentrada. Nesse sentido, considerando a decom-

posição feita com todos os 30 cânceres juntos, realizou-se a aproximação usando 11 mapas latentes e calculou-se o erro de cada microrregião, conforme apresentado em (6), para cada um dos 4 cânceres cujo ajuste não foi tão bom. O resultado dessa visão é apresentado na Figura 15.

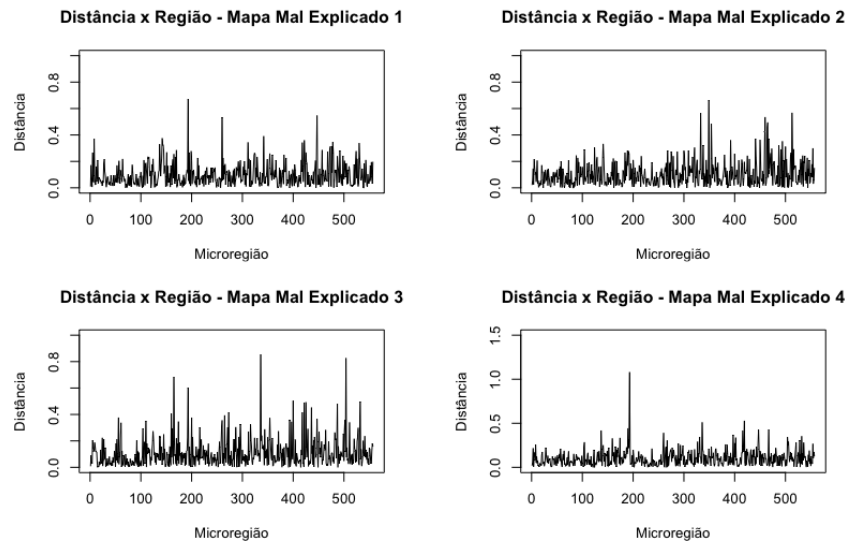


Figura 15: Erro por microrregião na aproximação com 11 mapas latentes para os cânceres considerados mal aproximados.

Analisando os gráficos apresentados na Figura 15, pode-se concluir que os erros não são tão grandes e estão entre 0 e 1. Em geral, conclui-se que o mal ajuste não está concentrado predominantemente em determinada microrregião, se fosse esse o caso, haveria um pico em todos os gráficos nos mesmos lugares, mas isso não acontece. Além disso, uma análise do percentil 90 dessas doenças mostram que as mesmas não se enquadraram no cenário considerado bem explicado por valores na ordem de 0.03 visto que o maior percentil entre elas foi de 0.226, fato que indica que a aproximação para tais doenças também é eficiente e que as mesmas foram classificadas como "ruins" por estarem fora da métrica de qualidade de ajuste.

Por fim, um modo de visualizar esse ajuste é comparar o comportamento desses mapas mal ajustados. A Figura 16 mostra o confronto do mapa real com o mapa aproximado.

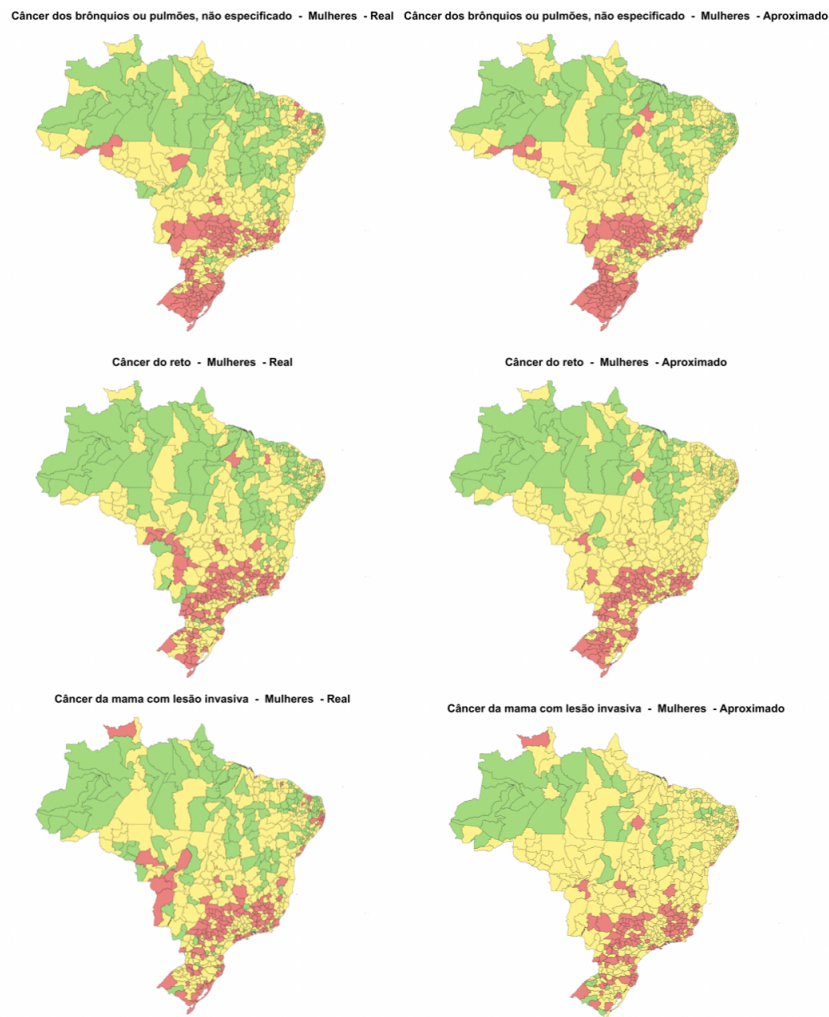


Figura 16: Comparação dos mapas reais e dos mapas aproximados considerando 11 mapas latentes. Nas representações, tem-se que as cor verde representa as menores taxas enquanto a vermelha indica maior risco e a amarela representa o risco intermediário.

Conforme o esperado, os mapas aproximados são capazes de captar bem o padrão dos mapas reais de modo que, percebe-se que a classificação como "mal ajustado" aconteceu em um limite muito tênue. Para o câncer mais comum, apresentado no topo da imagem, o mapa aproximado captou bem o padrão de altas mortalidades na região Sul, na parte mais ao Sul de Minas e nos estados de São Paulo, Rio de Janeiro e Espírito Santo. Além disso, na região Norte predomina o registro de taxas mais baixas, enquanto na região Sudeste há o registro de taxas intermediárias. Nesses pontos, o mapa aproximado falhou atribuindo valores intermediários a regiões mais baixas, assim como atribui taxas mais elevadas no Acre e no

norte de Rondônia. Nessa comparação, observa-se que o mapa da doença mais rara desse grupo foi o que mais apresentou diferenças entre o aproximado e o real de modo que, na aproximação, registra-se o predomínio de taxas intermediárias enquanto no real as mesmas se mesclam com taxas baixas.

Assim, entre as doenças mal explicadas, constata-se que a aproximação usando 11 mapas latentes é boa. Os mapas aproximados são capazes de captar o padrão espacial mais geral dessas doenças, acontecendo registros diferentes em algumas localidades. Esses erros, porém, foram ligeiramente maiores do que os registrados nas doenças consideradas bem explicadas, de modo que não atenderam à medida de qualidade por uma diferença pequena.

5 Conclusão

Nesse trabalho, introduzimos uma metodologia na qual o objetivo é representar os padrões e comportamentos das incidências e/ou mortalidades em determinados mapas por uma pequena quantidade de padrões espaciais que, uma vez entendidos, viabilize a proposição de medidas e programas para redução do número de casos. O problema atual nesse processo é que atualmente existem muitos tipos de cânceres, o que implica que o processo de estudo dos mapas de incidência ou de mortalidade, individualmente, é extremamente difícil e trabalhoso o que dificulta determinar medidas que sejam eficazes para um grupo de doenças.

A teoria da Decomposição do Valor Singular foi utilizada para decompor os mapas observados em mapas latentes. Como resultado para dados norte-americanos, obteve-se que, em um cenário de 30 cânceres, são necessários apenas 3 mapas latentes para que a todos os mapas dos Estados Unidos possam ser considerados bem explicados pela aproximação gerada via SVD. No Brasil, os resultados apontam para o uso de 11 mapas latentes para se ter a maioria com um bom ajuste. Comparando os resultados, tem-se que a diferença na quantidade de mapas a serem usados se deve a um padrão espacial mais bem definido para os casos dos Estados Unidos cujos números de morte são consideravelmente maiores do que os registrados no Brasil.

Além disso, um confronto dos mapas aproximados com os mapas reais aponta para um cenário onde usar poucos mapas latentes geram aproximações coerentes com a realidade. Ressalta-se, porém, que, para os dados brasileiros, há doenças cuja aproximação não atendeu aos critérios de qualidade estabelecidos (percentil 90 dos erros de no máximo 0.2), mas as mesmas apresentaram bons resultados, não sendo classificadas como bem explicadas por uma diferença do corte estabelecido da ordem de 0.03.

Por fim, podemos concluir que o estudo dos mapas de câncer individualmente deixa de existir e passa-se a ter um processo onde o foco é alguns poucos mapas latentes. Para os EUA, a redução gerada foi de 90% enquanto para o Brasil foi de 63,33%.

5.1 Trabalhos Futuros

Como próximos passos desse trabalho, tem-se a possibilidade de inclusão de componentes estocásticos para melhorar mais as aproximações e também oferecer um entendimento da variabilidade apresentada pelos dados. Para facilitar a interpretação do mapa com os pesos, métodos de decomposição que retornem só valores positivos são de interesse e devem ser investigados. Além disso, melhorias no código para deixá-lo mais automatizado e estruturado para a futura criação de um pacote.

Referências

- [1] Daniel Billsus, Michael J Pazzani et al. “Learning Collaborative Information Filters.” Em: *Icml*. Vol. 98. 1998, pp. 46–54.
- [2] Susanna M Cramb et al. “Inferring lung cancer risk factor patterns through joint Bayesian spatio-temporal analysis”. Em: *Cancer epidemiology* 39.3 (2015), pp. 430–439.
- [3] Amy Downing et al. “Joint disease mapping using six cancers in the Yorkshire region of England”. Em: *International Journal of Health Geographics* 7.1 (2008), p. 41.
- [4] Mauricio Fernandes, Ramon Lopes e Renato Assunção. “Quantos mapas existem de fato? Minerando o atlas de incidência de câncer”. Em: (2012).
- [5] Virgilio Gómez-Rubio e Francisco Palmí-Perales. “Multivariate posterior inference for spatial models with the integrated nested Laplace approximation”. Em: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 68.1 (2019), pp. 199–215.
- [6] Leonhard Held et al. “Towards joint disease mapping”. Em: *Statistical methods in medical research* 14.1 (2005), pp. 61–82.
- [7] Leonhard Knorr-Held e Nicola G Best. “A shared component model for detecting joint and selective clustering of two diseases”. Em: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 164.1 (2001), pp. 73–85.
- [8] Alastair H Leyland e Carolyn A Davies. “Empirical Bayes methods for disease mapping”. Em: *Statistical Methods in Medical Research* 14.1 (2005), pp. 17–34.
- [9] Arkadiusz Paterek. “Improving regularized singular value decomposition for collaborative filtering”. Em: *Proceedings of KDD cup and workshop*. Vol. 2007. 2007, pp. 5–8.
- [10] David Poole. *Linear algebra: A modern introduction*. Cengage Learning, 2014.
- [11] Havard Rue et al. “INLA: functions which allow to perform a full Bayesian analysis of structured additive models using Integrated Nested Laplace Approximation”. Em: *R package version 0.0* (2009).

- [12] Yongchang Wang e Ligu Zhu. “Research and implementation of SVD in machine learning”. Em: *2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS)*. IEEE. 2017, pp. 471–475.