Universidade de São Paulo

Faculdade de Saúde Pública

Aglomerados espaciais de alta mortalidade por câncer no Brasil: uma abordagem de *machine learning*

Bruno Casaes Teixeira

Dissertação apresentada à Faculdade de Saúde Pública da Universidade de São Paulo, para obtenção do Título de Mestre em Saúde Pública.

Área de Concentração: Saúde Pública

Orientador: Prof. Dr. Alexandre Dias Porto Chiavegatto Filho

São Paulo

2020

Aglomerados espaciais de alta mortalidade por câncer no Brasil: uma abordagem de *machine learning*

Bruno Casaes Teixeira

Dissertação apresentada à Faculdade de Saúde Pública da Universidade de São Paulo, para obtenção do Título de Mestre em Saúde Pública.

Área de Concentração: Saúde Pública

Orientador: Prof. Dr. Alexandre Dias Porto Chiavegatto Filho

Versão revisada

São Paulo

2020

Catalogação da Publicação
Ficha elaborada pelo Sistema de Geração Automática a partir de dados fornecidos pelo(a) autor(a)
Bibliotecária da FSP/USP: Maria do Carmo Alvarez - CRB-8/4359

Teixeira, Bruno Casaes
Aglomerados espaciais de alta mortalidade por câncer no Brasil: uma abordagem de Machine Learning / Bruno Casaes Teixeira; orientador Alexandre Dias Porto Chiavegatto Filho. -- São Paulo, 2021.
88 p.

Dissertação (Mestrado) -- Faculdade de Saúde Pública da Universidade de São Paulo, 2021.

1. Machine Learning. 2. Epidemiologia Espacial. 3. Câncer. 4. Modelagem Epidemiológica. I. Chiavegatto-Filho, Alexandre Dias Porto, orient. II. Título.

À três mães:

À mãe de minha mãe, que me ensinou dignidade.

À minha mãe, que me ensinou a ser forte.

À mãe de meu filho, que me ensinou a perseverar.

Agradecimentos

Meu primeiro agradecimento não poderia deixar de ser para a instituição da Faculdade de Saúde Pública da Universidade de São Paulo e tudo aquilo que ela representa como um centro de excelência em saúde coletiva, mas também como um baluarte de sanidade, palavra que adquire um significado especial no momento de politização e irracionalidade da saúde em que nos encontramos no último ano.

Ao Prof. Dr. Alexandre Dias Porto Chiavegatto Filho pelo apoio, orientação e inúmeras boas ideias que ajudaram a lapidar este projeto.

Ao Prof. Dr. Francisco Chiaravalloti-Neto de quem aprendi os conceitos fundamentais que compõem os alicerces deste projeto.

Ao Prof. Dr. Fredi Alexander Diaz Quijano que fez de suas aulas de análises de dados epidemiológicos os momentos mais inspiradores e decisivos em minha formação em epidemiologia.

À Teresa Lemmer, minha esposa, por toda motivação, apoio e cobrança durante o processo de desenvolvimento deste projeto.

Todas as coisas estão relacionadas com todas as outras, mas coisas próximas estão mais relacionadas do que coisas distantes.

Primeira Lei da Geografia

Waldo Tobler

Resumo

Título: Aglomerados espaciais de alta mortalidade por câncer no Brasil: uma abordagem de machine learning

Objetivos: Este estudo teve como objetivo avaliar se a mortalidade por câncer e seus subtipos pode ser prevista utilizando modelos de *machine learning* e dados socioeconômicos, demográficos e de cobertura de saúde como variáveis independentes. Adicionalmente buscou-se avaliar a associação geográfica dos resíduos destes modelos, ou seja, a porção de mortalidade não explicada por variáveis sociodemográficas e de saúde. **Metodologia:** Dados de mortalidade foram extraídos para os anos de 2008 a 2016 utilizando o Sistema de Informações de Mortalidade (SIM) e ajustados por idade utilizando a população padrão da Organização Mundial da Saúde (OMS). Variáveis sociodemográficas e de cobertura de saúde foram obtidas do Censo 2010 e do Ministério da Saúde do Brasil, respectivamente. Foram selecionados os algoritmos mais populares de *machine learning* para dados estruturados*: random forest, extreme gradient boosting, polynomial support vectors machines* e regressão lasso, treinados com 80% dos dados para predizer a taxa ajustada de mortalidade por câncer no nível municipal e sua performance foi testada com os restantes 20% das cidades. À partir dos resíduos, foram identificados os municípios com as taxas de mortalidade acima da esperada. Os aglomerados espaciais foram identificados utilizando a estatística de Kulldorff. Os testes foram repetidos para os dez tipos de câncer com maior mortalidade no Brasil no período avaliado. **Resultados:** Em geral, o algoritmo com maior R^2 foi o *gradient boosting trees* (R^2=0,66). Para o consolidado de todos os cânceres, todos os algoritmos apontaram a existência de um aglomerado espacial na região entre Bagé e Rio Grande (excesso de mortalidade de 27%) e três algoritmos identificaram aglomerados na região da cidade de Porto Velho (excesso entre 27% e 40%). Para câncer de esôfago, na região oeste do estado do Rio Grande do Sul foram identificados importantes aglomerados parcialmente sobrepostos por todos os algoritmos (excessos entre 48% e 96%), sendo que outros aglomerados importantes foram identificados no sul do Paraná, norte de Minas Gerais e Espírito Santo. Para câncer de estômago foi identificado um importante cluster na região de Macapá (excesso de 82%) e na região de Porto Velho (excesso de 85%). As variáveis com maior impacto na predição da mortalidade para todos os cânceres foram percentual de população branca, com uma contribuição positiva e linear, e percentual de casas com computador, com uma contribuição positiva e não linear. **Conclusão:** Algumas regiões geográficas brasileiras mostram taxas significativamente acima do esperado para mortalidade

por câncer, independentemente de variáveis sociodemográficas. Análises adicionais poderão explorar a causalidade dessas diferenças geográficas.

Descritores: Machine Learning, Epidemiologia Espacial, Câncer, Modelagem Epidemiológica

Abstract

Title: Spatial clusters of cancer mortality in Brazil: a machine learning modelling approach

Objectives: This study aimed to assess whether cancer mortality and its subtypes can be predicted using *machine learning* models and socioeconomic, demographic and health coverage as independent variables. Additionally, we sought to evaluate the geographical association of the residuals of these models; in other words, the portion of mortality not explained by sociodemographic and health variables. **Methodology:** Mortality data were extracted for the years 2008 to 2016 using the Mortality Information System (SIM) and adjusted for age using the standard population of the World Health Organization (WHO). Sociodemographic and health coverage variables were obtained from the 2010 Census and the Ministry of Health of Brazil, respectively. We selected some of the most popular *machine learning* algorithms for structured data: random forest, extreme gradient boosting, polynomial support vectors machines and lasso regression, trained with 80% of the data to predict the adjusted cancer mortality rate at the municipal level and their performance was tested with the other 20% of cities. From the residuals, municipalities with higher-than-expected mortality rates were identified. Spatial clusters were identified using Kulldorff statistics. The tests were repeated for the ten cancer types with the highest mortality in Brazil in the evaluated period. **Results:** In general, the algorithm with the highest R^2 was the gradient boosting trees (R^2 = 0.66). For the all cancers group, all algorithms pointed to the existence of a spatial cluster in the region between Bagé and Rio Grande (27% of excess mortality) and three algorithms identified clusters in the region of Porto Velho city (excess between 27% and 40%). For esophageal cancer, in the western region of the state of Rio Grande do Sul, important clusters were partially overlapped by all algorithms (excesses between 48% and 96%). Other important clusters were identified in southern Paraná, northern Minas Gerais, and Espírito Santo. For stomach cancer, an important cluster was identified in the Macapá region (82% excess) and in the Porto Velho region (85% excess). The variables with the greatest impact on the mortality prediction for the all cancers group were the percentage of the white population, with a positive and linear contribution, and the percentage of houses with computers, with a positive and non-linear contribution. **Conclusion:** Some Brazilian geographic regions show significantly higher than expected rates for cancer mortality, regardless of sociodemographic variables. Additional analyzes may explore the causality of these geographical differences.

Key words: Machine Leearning, Spatial Epidemiology, Cancer, Epidemiological Modeling

Anexo à esta dissertação encontra-se o artigo submetido para publicação em uma revista científica internacional

SUMÁRIO

1. INTRODUÇÃO

1.1. CÂNCER

Estima-se que em 2018 tenha havido 18 milhões de novos casos de câncer no mundo e um total de 9,6 milhões de mortos. Cerca de 20% dos homens e 17% das mulheres terão câncer em algum momento de suas vidas e 12% dos homens e 9% das mulheres morrerão dessa doença (IARC, 2018).

O termo câncer refere-se a uma grande quantidade de doenças cuja característica comum é a multiplicação descontrolada de células do corpo (Sondik, 1990). A incidência dessas doenças varia consideravelmente em relação a fatores como idade, sexo, status socioeconômico, geografia e genética.

A estratégia mais comum de classificar os diferentes tipos de câncer é pelo tecido do qual se iniciou a doença, um processo conhecido por oncogênese. São denominados carcinomas os tumores que se iniciam de tecidos epiteliais como a pele e mucosas. Aqueles que se iniciam de tecidos conjuntivos como cartilagens, ossos e músculos são conhecidos como sarcomas.

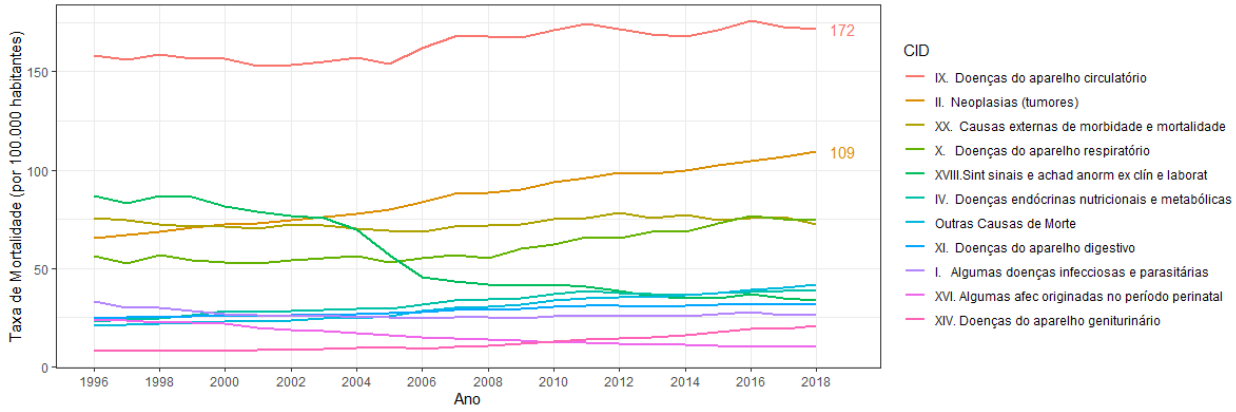
Outro importante sistema de classificação de tumores é pela sua capacidade de invadir diferentes tipos de tecido, sendo o tumor localizado aquele que não invadiu tecidos adjacentes e os tumores metastáticos aqueles que invadem órgãos distantes (Loda et al., 2016).

A compreensão dos elementos que influenciam a oncogênese é de extrema importância para o desenvolvimento de estratégias de prevenção, controle e tratamento do câncer. Como uma forma de complementar o crescente desenvolvimento do conhecimento dos fatores genéticos que levam ao câncer, amplamente conhecido como “genômica”, busca-se também a compreensão dos fatores ambientais com os quais cada organismo interagiu. Entender, portanto, a “exposômica”, ou seja, os fatores de exposição e risco, é uma área crescente de interesse científico (Wild, 2005).

No Brasil, de acordo com dados do Sistema de Informações de Mortalidade (SIM), em 2017 o câncer foi responsável por 221.821 mortes e é a segunda causa mais frequente de morte, depois das doenças cardiovasculares com 358.882 pessoas. Em termos relativos, o câncer possui uma taxa de mortalidade de 107

mortes por 100.000 habitantes, que vem crescendo desde 1996 com uma taxa média anual de 2,4%, a mais alta taxa de crescimento de todas as causas analisadas na Figura 1.

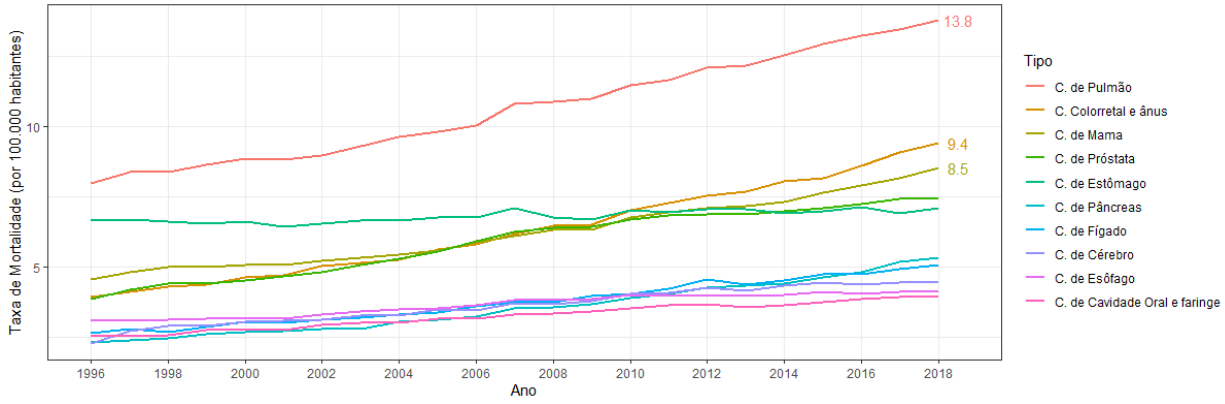
Figura 1 - Taxa de Mortalidade por Capítulo do CID-10 (Brasil 1996 - 2018).



Fonte: Dados do Sistemas de Informação de Mortalidade (SIM), do autor.

Entre os tipos de câncer, os que mais causam mortes são aqueles relacionados aos órgãos digestivos (estômago, pâncreas, cólon e reto), seguidos de cânceres relacionados ao aparelho respiratório (pulmão e brônquios). Mais detalhes podem ser observados na Figura 2.

Figura 2 – Taxa de mortalidade pelos tipos de Tumor mais comuns (Brasil 2017).



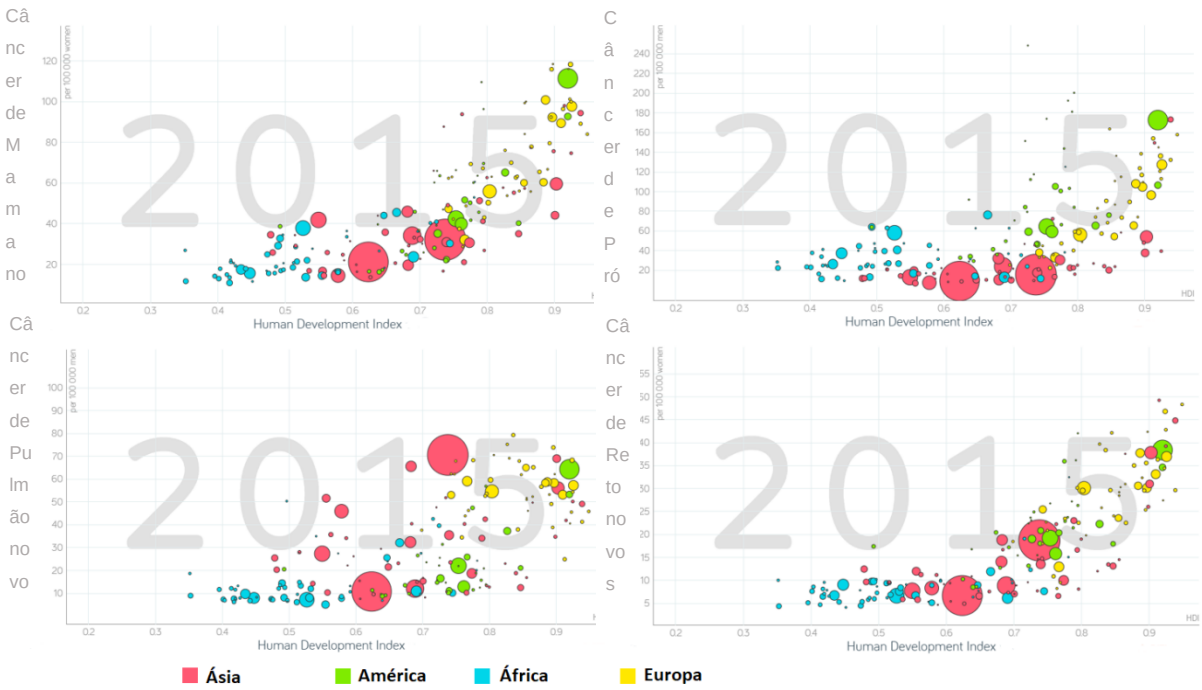
Fonte: Dados do Sistemas de Informação de Mortalidade (SIM), do autor.

Em relação à incidência, a estimativa mais atual do Instituto Nacional do Câncer (INCA) é de 685.960 casos novos de câncer no Brasil para o ano 2020. Os cânceres mais comuns são os de próstata e de mama (Ministério da Saúde do Brasil, 2019). Estes, entretanto possuem uma taxa de letalidade baixa em relação a outros tumores, fazendo com que não sejam as principais causas de morte de acordo com a Figura 2.

1.2. O CÂNCER E A TEORIA DA TRANSIÇÃO EPIDEMIOLÓGICA

A relação do câncer com a riqueza dos países é conhecida na literatura. Como pode ser observado na Figura 3, a comparação do Índice de Desenvolvimento Humano (IDH) com a taxa de incidência (ajustada por idade) dos 4 diferentes tipos de câncer em diversos países, mostra que quanto maior o IDH, uma medição composta de indicadores de expectativa de vida, educação e renda *per capita* utilizada pela Organização das Nações Unidas em seu programa de Desenvolvimento (Anand and Sen, 1994), maior a incidência de alguns tipos de Câncer.

Figura 3 - Taxas de Incidências em Homens e Mulheres de 4 tipos de Câncer (Mama, Prostata, Pulmão Colorretal) em função do Índice de Desenvolvimento Humano (2015) de Diferentes Países. Diâmetro do círculo representando a população do país.



Fonte: Gapminder (Gapminder Tools, 2020).

Uma vez que as doenças oncológicas afetam as pessoas de maneira diferente com relação a idade, é necessário ajustar o dado para a diferença da composição etária dos países para a correta comparação da sua incidência entre diferentes regiões (Milyo and Mellor, 2003). A razão dessa discrepância está relacionada à Teoria da Transição Epidemiológica.

1.2.1. Teoria da Transição Epidemiológica

A Teoria da transição epidemiológica foi proposta em 1971 por Abdel Omran (Omran, 1971). Nela, Omram argumenta que o perfil epidemiológico das sociedades varia de maneira complexa devido às suas interações com elementos demográficos, econômicos, determinantes sociológicos e suas consequências. Assim, o padrão de saúde-doença muda com os avanços sociais e sanitários, promovendo melhores probabilidades de uma vida mais longa, afetando consequentemente os índices de fertilidade. Dessa forma, devido ao aumento da expectativa de vida e da menor mortalidade infantil, ocorre uma transição das doenças infecciosas para doenças crônicas. Essa transição, de acordo com a proposta de Omran, ocorre em três estágios:

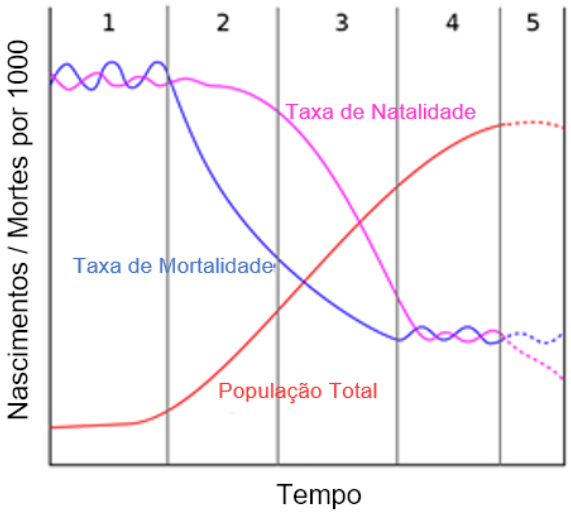
1. **A era da peste e da fome:** Neste estágio há uma taxa alta e flutuante de mortalidade associada a uma alta taxa de natalidade, o que impede o crescimento populacional. Nesta etapa há uma baixa expectativa de vida ao nascer, variando entre 20 e 40 anos com um elevado número de mortes infantis e de mulheres em idade fértil. A vida das pessoas neste estágio é marcada por doenças infecciosas e parasitárias, mortalidade no parto, pela fome e por guerras.
2. **A era das pandemias em queda:** Neste estágio, com o avanço tecnológico e sanitário, os picos epidêmicos se tornam menos frequentes ou mesmo desaparecem, e assim a taxa de mortalidade cai significativamente e de forma abrupta. A expectativa de vida ao nascer passa rapidamente de 30 anos para 50 anos. A taxa de natalidade mantém-se estável e por essa razão a população cresce de forma exponencial.
3. **A era das doenças degenerativas e causadas pelo homem:** Neste estágio as taxas de mortalidade continuam caindo até se estabilizarem em níveis baixos. Nesta etapa, as doenças degenerativas ganham grande importância juntamente com doenças cardiovasculares, câncer e violência. O principal fator afetando o crescimento populacional é a taxa de natalidade, que começa a cair neste estágio.

À proposta inicial de Omran, foi proposta por Barrett (1998) a adição de mais dois estágios em que mudanças na alimentação e estilo de vida e acesso à medicina moderna diminuem a mortalidade por doenças cardiovasculares, o que aumenta a taxa de envelhecimento da população.

4. **A era da mortalidade descendente por doenças cardiovasculares:** Neste estágio, os avanços na tecnologia médica estabilizam a mortalidade e a taxa de natalidade se estabiliza em níveis baixos. Devido a mudanças no estilo de vida e medicina, há redução da mortalidade por doenças cardiovasculares. A crescente resistência a antibióticos, novos patógenos, reascendem a mortalidade por agentes infecciosos.
5. **A era da qualidade de vida ansiada com desigualdades persistentes:** Neste estágio, há um declínio nas taxas de natalidade e a expectativa de vida mantém o seu crescimento em média.

Os cinco estágios da teoria da transição epidemiológica permitem compreender o efeito do desenvolvimento econômico, social e tecnológico nos comportamentos da dinâmica demográfica e populacional de acordo com a Figura 4.

Figura 4 - Evolução demográfica e das taxas de Nascimento e Mortalidade nos Estágios da Teoria da Transição epidemiológica de Omran.



Fonte: Adaptado pelo autor de (Health Knowledge, n.d.)

Ainda que presente desde a antiguidade com menções em papiros egípcios (Mukherjee, 2012), o câncer se tornou um problema de saúde pública mais recentemente com o envelhecimento da população crescendo principalmente nas fases três e quatro da transição epidemiológica, quando doenças degenerativas começam a ganhar importância relativa.

Com o avanço da medicina e a redução da mortalidade por doenças cardiovasculares, o câncer se tornou uma causa de morte muito mais frequente em países desenvolvidos que, capazes de tratar mais amplamente a população, reduziram significativamente a mortalidade por doenças cardiovasculares, aumentando ainda mais o envelhecimento populacional, além da adoção de comportamentos de risco como tabagismo e alto consumo calórico, fatores de risco importantes para o câncer (Gersten and Wilmoth, 2002).

1.3. A DESIGUALDADE DO BRASIL COMO UM MODELO PARA MODELAGEM SOCIOECONÔMICA DE DOENÇAS

O Brasil é um país com marcantes desigualdades. A desigualdade pode ser observada por meio da análise de fatores econômicos (Azzoni, 2001), características de saúde (Albuquerque et al., 2017; Barbosa et al., 2016; Chiavegatto Filho et al., 2013) e indicadores educacionais (Rios-Neto and Guimarães, 2010).

As raízes da desigualdade brasileira refletem movimentos históricos marcados pela exploração de recursos naturais com distribuição heterogênea no território nacional, bem como a movimentações demográficas que priorizaram a ocupação da região costeira do Brasil(Rigotti, 2008).

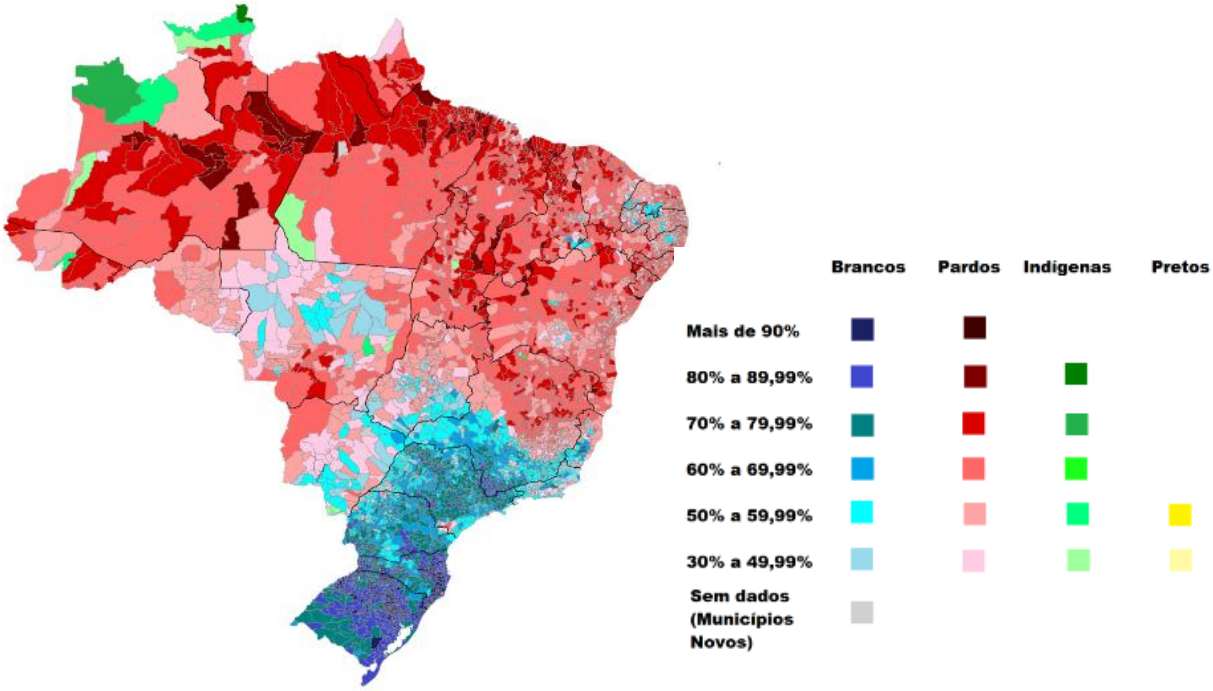
Existe também no país uma importante diversidade na distribuição étnica dos brasileiros que, como pode ser observado na

, possuem uma predominância europeia concentrada no sul do Brasil, e de pardos, pretos e indígenas na região Norte e Nordeste do país.

Um importante fator de confusão em modelos estatísticos globais de doença e variáveis socioeconômicas é o fato de que cada país possui um sistema de saúde distinto, incluindo um fator de confusão para modelos inferenciais que

precisa ser transposto. No Brasil, com o Sistema Único de Saúde, tem-se um ambiente relativamente mais homogêneo de políticas públicas de saúde, com importantes variações no acesso a elas devido às características socioeconômicas de cada região.

Figura 5 - Distribuição étnica do Brasil



Fonte: (Brasil, 2019)

Dessa forma, com um ambiente étnico, ambiental e socioeconômico diverso, porém com um acesso relativamente homogêneo a políticas públicas de saúde, o Brasil possui um ambiente favorável à modelagem estatística para questões epidemiológicas.

Entretanto, há poucos trabalhos que exploram essas características. Chiavegatto Filho e colaboradores (2018) analisaram a diversidade econômica do Brasil para identificar as práticas em saúde associadas a um maior impacto no aumento ou diminuição da expectativa de vida dos municípios brasileiros. Karagiannis-Voules e colaboradores (2013) exploraram as diferenças demográficas, na flora e climáticas para avaliar a relação dessas com as taxas de incidência de leishmaniose cutânea e visceral no Brasil.

1.4. FONTES DE DADOS NO BRASIL

Além da diversidade étnica e socioeconômica e da homogeneidade do sistema de saúde, colaboram com um ambiente favorável para modelagens estatísticas no Brasil o fato de o país possuir bons sistemas de dados que capturam de forma sistemática variáveis associadas à assistência de saúde (DATASUS), mortalidade (SIM), doenças de notificação obrigatória (SINAN), socioeconômicas (Censo), entre outras.

Esses dados estão amplamente disponíveis de acordo com a Lei de Acesso à Informação (Presidência da República, 2011), que garante a obrigatoriedade dos órgãos de governo em contribuírem para a sua disponibilização.

Uma importante vantagem dessas bases de dados é a sua cobertura incluir todo o território nacional com diferentes níveis de granularidade, podendo chegar até mesmo ao nível de poucas dezenas de metros, como é o caso dos setores censitários do Censo de grandes centros urbanos.

A qualidade dessas informações foi amplamente avaliada e considerada de qualidade pela maioria dos estudos da literatura (Correia et al., 2014; Machado et al., 2016; Queiroz et al., 2017), ainda que observadas algumas diferenças regionais e temporais significativas.

1.4.1. DATASUS

O Sistema de Informações de Saúde do SUS (DATASUS) compreende diversas plataformas e sistemas que objetivam centralizar todos os dados de saúde produzidos pelo Sistema Único de Saúde (SUS). Entre os mais importantes estão o Sistema de Informações Ambulatoriais (SIA-SUS) e o Sistema de Informações Hospitalares (SIH-SUS) que tratam de toda a produção ambulatorial e hospitalar realizada no SUS (Lima et al., 2009). Essas bases de dados estão relacionadas às transações financeiras entre o SUS e os hospitais, clínicas e secretarias governamentais que atendem ao SUS. Dados de faturamento desse tipo já são utilizados para pesquisa científica em diversas áreas da epidemiologia, como em septicemia (Baine et al., 2001), doenças raras (Baine et al., 2001; Cosmatos et al., 2013), doença renal (Saran et al., 2017), entre outras.

1.4.2. SIM

Além do SIA-SUS e do SIH-SUS, outro importante sistema para a avaliação epidemiológica é o Sistema de Informações de Mortalidade (SIM) que possibilita identificar as principais causas de mortalidade no Brasil, seus estados e municípios.

Com o objetivo de se criar um sistema básico de vigilância epidemiológica, o SIM foi implementado pelo Ministério da Saúde em 1975 com a integração e unificação de sistemas estaduais que já coletavam esta informação. Para isso, foi estipulada a padronização da declaração de óbito (DO) e da declaração de óbito fetal e o amplo treinamento de codificadores de causa básica feito pelo Departamento de Epidemiologia da Faculdade de Saúde Pública (Fundação Nacional de Saúde (FUNASA), 2001).

A qualidade dos dados do SIM foi avaliada como crescente por Queiroz (2017), com um grau de cobertura significativo, porém com algumas variações regionais importantes, especialmente nas Regiões Norte e Nordeste do país.

1.4.3. Censo IBGE

Em 2010, o IBGE realizou o Censo Demográfico em que foram visitados 67,6 milhões de domicílios em 5.565 municípios Brasileiros (IBGE, 2010). As informações são disponibilizadas em diversos formatos e níveis de granularidade, porém as de maior interesse para este estudo são os dados de Resultados do Universo agregados por Setor Censitário (IBGE, 2012), as quais, para as análises realizadas, foram consolidadas por unidades municipais.

1.5. MACHINE LEARNING

O termo *machine learning* foi introduzido em 1959 por Arthur Samuel em um trabalho sobre o jogo de damas enquanto trabalhava para a IBM. Nesse trabalho, Samuel já previa a ampla aplicação destas metodologias em diversas situações além do jogo de damas (Samuel, 1959). Desde então, *machine learning* desenvolveu-se para, hoje, ser a ferramenta de escolha para a identificação de padrões complexos em dados.

Breiman (2001) demonstrou que os algoritmos de *machine learning* (como árvores de decisão e redes neurais) conseguem muito frequentemente superar a performance preditiva de metodologias estatísticas historicamente utilizadas para

explicar a relação entre as variáveis dependente e independentes, como as regressões linear e logística. Isso é especialmente verdadeiro quando se lida com casos em que o número de variáveis independentes supera o número de observações.

No nível do paciente, métodos de *machine learning* são comumente aplicados em oncologia (Kourou et al., 2015). Estudos têm demonstrado que é possível prever a sobrevida de pacientes com câncer com uma significativa taxa de sucesso (Kim et al., 2014; Park et al., 2013), além das taxas de resposta a quimioterapia (Ma et al., 2006; Mucaki et al., 2019), estimação de dados faltantes (Jerez et al., 2010), classificação de tumores (Wang et al., 2005) e predição de recaídas (Cirkovic et al., 2015). Apesar de sua ampla utilização em oncologia, modelos de *machine learning* ainda têm sido pouco explorados em análises ecológicas.

Algoritmos de *machine learning* são divididos em **não supervisionados** e **supervisionados,** sendo que estes últimos podem ser utilizados em problemas de **regressão** e **classificação**.

1.5.1. Métodos Não Supervisionados

Métodos não supervisionados se referem a algoritmos que buscam agrupar uma série de observações em grupos baseando-se na semelhança de uma série de variáveis independentes.

1.5.2. Métodos Supervisionados

Métodos supervisionados se referem a algoritmos que utilizam uma variável dependente (que será predita) e variáveis independentes (que serão utilizadas como preditores). Nesse caso, o dado é separado em dois grupos de treino e teste. O grupo de treino será utilizado para treinar o algoritmo até que alguma métrica de performance seja atingida ou maximizada, e então o algoritmo passa pela base de teste para que sua performance seja medida de maneira independente em novos dados.

Os problemas supervisionados, por sua vez, podem ainda ser classificados em dois grupos distintos:

1.5.3. Problemas de Regressão

Neste tipo de problema, a variável dependente é uma variável contínua. Uma forma comum de mensurar a qualidade de um modelo de regressão é utilizando o RMSE (Root Mean Square Error – Erro médio quadrático) (Barnston, 1992).

$$RMSE = \sqrt[2]{\frac{1}{N} * \sum_{i=1}^{N} (predito_i - observado_i)^2}$$

Sendo **N** o número de observações avaliadas.

Uma segunda forma de avaliar problemas de regressão é utilizando o R quadrado ou R^2, que é obtido pela seguinte fórmula:

$$R^2 = \frac{\sum_i (y_i - \hat{y_i})^2}{\sum_i (y_i - \bar{y})^2}$$

Sendo $\hat{y}$ o valor estimado, $\bar{y}$ a média dos valores observados e y o valor observado para cada observação i.

1.5.4. Problemas de Classificação

Neste tipo de problema, a variável dependente é uma variável categórica que pode ser positiva/negativa ou assumir mais do que apenas duas categorias. Existem diversas formas de se avaliar esse tipo de problema sendo a mais comum a área abaixo da curva ROC (Fawcett, 2006). Com essa análise, é frequente avaliar ao mesmo tempo a sensibilidade e a especificidade do modelo.

1.5.5. Dilema Previsibilidade versus Interpretabilidade

O uso de técnicas de *machine learning* tem sido cada vez mais popular pela sua alta eficiência preditiva. Essa eficiência está associada à crescente complexidade dos modelos construídos que podem incluir interações não lineares, com a criação de padrões de decisão cada vez mais específicos e, principalmente, pela capacidade de modelar a interação complexa de variáveis independentes. Toda essa complexidade tem como consequência uma menor capacidade de interpretar diretamente essas relações.

Essa relação inversa entre a capacidade preditiva e a interpretabilidade ganha especial relevância em aplicações em saúde. Um exemplo importante foi

trazido Mucaki e colaboradores (2019) ao criar um modelo preditivo da resposta à quimioterapia utilizando as complexas relações entre diferentes genes. Nesse caso, ao buscar pelo melhor modelo preditivo abriu-se mão da capacidade de compreensão do porquê o modelo funcionar. Por um lado, para a prática clínica, é importante compreender por que se deve prescrever um tipo ou outro de quimioterápico, por outro, a certeza de que por trás desta “caixa-preta”, há um modelo capaz de indicar aquela droga com maior probabilidade de sucesso para aquele perfil genético específico do paciente. Técnicas mais recentes permitem compreender melhor a relação das variáveis com a predição e por consequência “abrir a caixa preta” da interpretabilidade (Lundberg and Lee, 2017).

Por outro lado, algoritmos de *machine learning* podem refletir vieses atuais da prática clínica. Algoritmos desenvolvidos em contextos não-médicos já apontaram para o importância das fontes de informações que foram utilizadas para construí-los, como é o caso de um modelo construído para auxiliar juízes a predizer a probabilidade de reincidência que apresentava uma significativa propensão à discriminação racial (Angwin et al., 2016).

Esse dilema técnico transforma-se, portanto, em um dilema ético quando o resultado indicado pelo modelo diverge daquele de melhor resultado de acordo com a experiência clínica do médico e este resiste em utilizá-lo por não compreender as razões dessa divergência, optando assim por um tratamento menos eficaz, porém com um mecanismo conhecido e plausível.

1.5.6. Explicação Aditiva de *Shapley* (SHAP)

Com o objetivo de aproximar modelos de alta performance, e por consequência complexos, da interpretabilidade possibilitada por modelos simplificados sem que para isso se abra mão da capacidade preditiva proporcionadas por eles, foi desenvolvida recentemente a técnica SHAP (*SHapley Additive exPlanation*)(Lundberg and Lee, 2017).

De forma similar à resolução de um problema conceitual da teoria de jogos, Shapley (1951) introduziu o que ficou posteriormente conhecido como o Valor de Shapley, um método matemático que busca uma forma justa de compensar jogadores por sua contribuição para o resultado global em jogos de cooperação.

Considera-se, neste caso, um modelo complexo de *machine learning* como um jogo de cooperação, onde os jogadores são representados pelas variáveis independentes e a contribuição desses jogadores como o impacto que essas variáveis tiveram no resultado de cada instância.

O método de SHAP é uma implementação do valor de *Shapley* com incrementos aditivos para algoritmos de combinação de árvores de decisão como é o caso de Florestas Aleatórias e *XGBoost* (Lundberg et al., 2018). Para tal, o método SHAP trata cada contribuição de forma cumulativa no que resulta em um modelo linear especificado por:

$$g(z') = \emptyset_0 + \sum_{j=1}^{M} \emptyset_j z'_j$$

Sendo g o modelo de explicação, ***z'*** representa a presença ou a ausência de cada variável dentro da coalizão, aceitando valores zero e um para cada variável, $\emptyset_j$ representando o Valor *Shapley* atribuído a cada variável e *M* representando o número máximo de coalizões. O cálculo do Valor *Shapley* é feito pela interação entre os estados presentes e ausentes de uma amostra de cada variável aplicada ao modelo original com o impacto no resultado ajustado em um modelo linear sendo $\emptyset_j$ o coeficiente deste modelo para cada variável *j*.

Cabe destacar que o método de SHAP não permite uma interpretação de causalidade entre a variável e o objeto da previsão. O resultado SHAP mostra a contribuição individual para a previsão final de cada variável no modelo independente de outras variáveis avaliadas, podendo estas estarem sujeitas a influência de outras variáveis não mensuradas.

1.6. ESTATÍSTICAS ESPACIAIS

1.6.1. Moran

Para avaliar a distribuição e concentração geográfica de algum atributo, o método mais utilizado é o Índice de Moran Local ou Indicador Local de Associação Espacial (LISA) (Anselin, 2010). Nesse método, é construída uma matriz de proximidade utilizando o mapa da região a ser analisada e é calculado para cada região *i* um índice local de associação *I* utilizando a seguinte fórmula:

$$I_i = z_i \sum_{j=1}^{n} w_{ji} z_j \Big/ \sum_{j=1}^{n} z_j^2$$

Sendo w_{ij} o valor de cada elemento da matriz normalizada de vizinhança, n o número de áreas, e z_i valores normalizados do atributo para a região i e seus vizinhos j. Dessa forma, ao analisar incidência, por exemplo, buscar áreas com índices z e I altos equivale a encontrar áreas que possuam altas incidências e que por sua vez também estejam cercadas de áreas com alta incidência.

1.6.2. Estatística de varredura espacial de Kulldorff

Com o objetivo de identificar as regiões com maior incidência de determinado evento sem incorrer em viés de seleção, foi proposto por Kulldorff (1997) o método que foi posteriormente denominado estatística de varredura espacial de Kulldorff.

Neste método, cria-se uma janela circular de tamanho variado centralizada em um ponto (no caso, o centro geográfico de uma cidade). Essa janela circular inclui no agregado as cidades vizinhas e busca aqueles agregados que possuem a maior incidência de determinado evento ajustando para o tamanho da população de cada agregado.

O valor do agregado é comparado com uma distribuição obtida por simulações de Monte Carlo para obter o valor *p* referente à probabilidade hipótese nula, ou seja, a de que a alta incidência daquele cluster poderia ocorrer pela combinação aleatória de cidades.

A estatística de varredura de Kulldorff é um método amplamente utilizado não apenas em epidemiologia. Romero Canal e colaboradores (2017), utilizou essa metodologia para encontrar aglomerados regionais de casos de dengue na cidade de São José do Rio Preto. Vieira e Cançado (2013) utilizaram essa mesma metodologia para identificar os principais aglomerados de acidentes aéreos no Brasil. Em oncologia, Kulldorff e colaboradores (1997) aplicaram a metodologia de varredura espacial para identificar os aglomerados de casos de câncer de mama da região Nordeste dos Estados Unidos da América.

2. OBJETIVOS

Os objetivos deste trabalho foram:

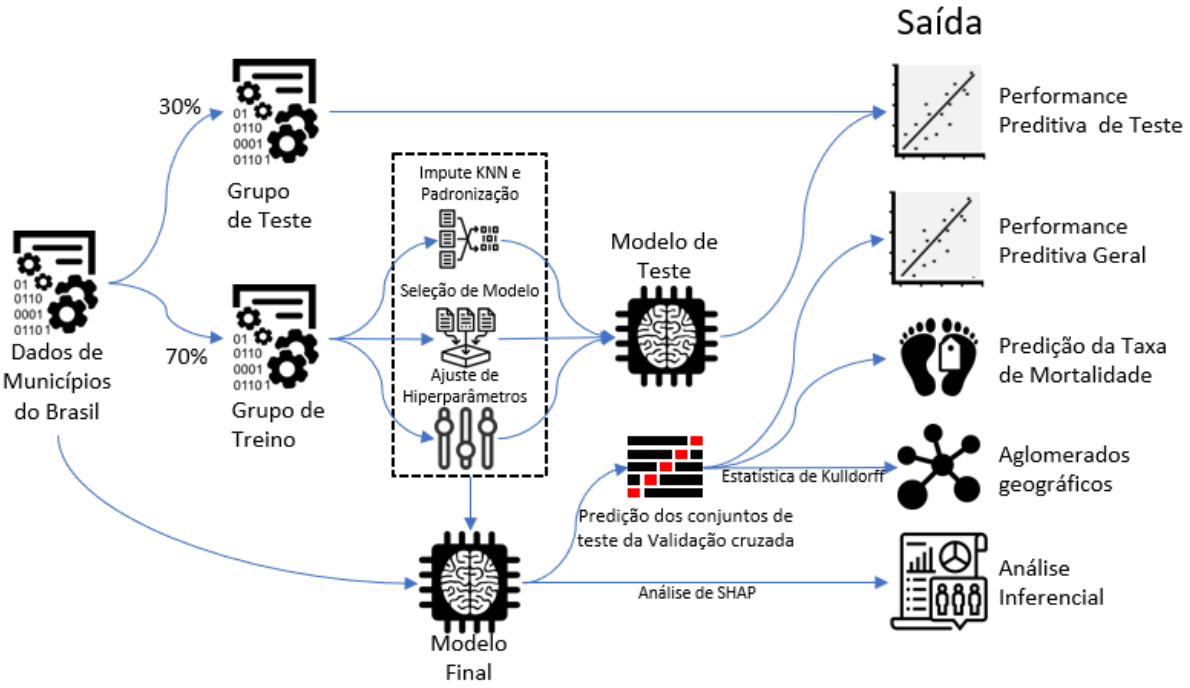
Objetivos

1- Analisar se a mortalidade ajustada por idade nos municípios do Brasil pode ser predita utilizando modelos de *machine learning*.
2- Avaliar a distribuição espacial do resíduo desses modelos, ou seja, a diferença entre o predito e o real, em busca de agregados espaciais de alta taxa de mortalidade não explicada por variáveis socioeconômicas.
3- Analisar a relação das variáveis independentes com cada variável dependente.

3. MÉTODOS

Um diagrama com os passos para a obtenção de resultados pode ser observado na Figura 6.

Figura 6 - Diagrama Metodológico



3.1. FONTE DE DADOS

Dados foram coletados de forma agregada no nível municipal para todo o Brasil.

3.1.1. Variáveis Independentes

Um total de 40 variáveis sociodemográficas relacionadas à renda, patrimônio, demografia e urbanização foram coletadas no último censo (IBGE, 2010). O percentual de cobertura privada de saúde foi obtido do portal da ANS (Agencia Nacional de Saúde Suplementar - ANS, 2019). Os detalhes de todas as variáveis da análise estão listados na Tabela 2. As coordenadas geográficas dos municípios para as análises espaciais foram obtidas no Instituto Brasileiro de Geografia e Estatística (IBGE, 2018).

3.1.2. Dados de Mortalidade

Os dados de Mortalidade foram obtidos do Serviço de Informações de Mortalidade SIM e foram extraídos utilizando o Software de Análises Estatísticas R (R Core Team, 2018).

Os dados foram filtrados com os seguintes critérios

- Mortes ocorridas entre os anos de 2007 e 2016.
- Segundo causa básica utilizando os códigos relacionados a câncer no Código de Identificação de Doenças versão 10 (CID-10) e aos 10 subtipos que mais causaram mortes no período, relacionados na Tabela 1.

Tabela 1 - Tipos de câncer considerados neste estudo e número de mortes contabilizadas no Período entre 2008 e 2016 (Brasil)

Tipo de Câncer	Número de Mortes Contabilizadas
C34 – Câncer de Pulmão	212.379
C18-C21 – Câncer Colorretal	132.802
C50 – Câncer de Mama	125.008
C16 – Câncer de Estômago	123.188
C61 – Câncer de Próstata	120.999
C22 – Câncer de Fígado	76.704
C25 – Câncer de Pâncreas	74.123
C15 – Câncer de Esôfago	70.386
C71 – Câncer Cerebral	64.741
C80 – Câncer de Localização não especificada	56.340

3.2. AJUSTES DE VARIÁVEIS

3.2.1. Padronização por Idade

As taxas de mortalidade ajustadas por idade para cada município foram calculadas usando a população padrão da Organização Mundial da Saúde de 2000 a 2025 (National Cancer Institute (NCI), 2013).

3.2.2. Padronização de Variáveis

Utilizando o pacote Caret (Kuhn et al., 2018), as variáveis independentes foram colocadas dentro de um intervalo entre 0 e 1. Dessa forma, a diferença de amplitude de medição das variáveis não afeta diretamente o modelo final (ex. PIB per capita em milhares de reais e percentual de geladeiras com valores entre 0 e 1).

3.2.3. Tratamento de Variáveis Faltantes com *K-Nearest Neighbors* (KNN)

Valores ausentes nas variáveis foram tratados com o método de KNN, que adiciona um valor médio para aquela variável com base em outras cidades com características semelhantes. Este tipo de conduta se mostrou valiosa em problemas clássicos de predição (Zhang, 2008).

3.3. MODELO PREDITIVO

3.3.1. Comparação de modelos

Os municípios foram separados de forma aleatória em grupo de treino (80% do total) e grupo de teste (20% do total). No conjunto de treino foram realizados testes com nove algoritmos populares de *machine learning* para regressão supervisionada. Essa análise foi realizada utilizando validação cruzada de 10 vezes com três repetições. Os algoritmos testados foram:

- Regressão Linear
- Regressão com LASSO
- Regressão com RIDGE
- Florestas Aleatórias (RF)
- Extreme Gradient Boosting (XGB)
- Máquina de Vetor de Suporte Linear
- Máquina de Vetor de Suporte Polinomial (pSVM)
- Modelo de Árvore de Inferência Condicional
- Árvores de Decisão

A performance dos modelos foi testada por meio do R quadrado com intervalo de confiança de 95% (IC 95%) e os quatro melhores algoritmos foram selecionados para a próxima fase (XGB, RF, pSVM e LASSO).

3.3.2. Seleção de Hiperparâmetros

Para cada algoritmo é necessário escolher parâmetros relacionados à performance e ao aprendizado, estes são genericamente denominados hiperparâmetros. Para cada algoritmo, foi feita uma busca dos hiperparâmetros com melhor performance utilizando validação cruzada de 10 vezes com três repetições e variação aleatória de hiperparâmetros do pacote caret (Kuhn et al., 2018).

Métodos

3.3.3. Teste dos Modelos

Foi analisada a performance dos modelos selecionados com seus hiperparâmetros no conjunto de teste.

3.3.4. Previsão

Após selecionar a combinação dos hiperparâmetros dos algoritmos com melhor desempenho, cada um foi treinado em todo o conjunto com validação cruzada de 10 vezes e os resultados dos conjuntos de testes de cada grupo de validação cruzada foram os valores preditivos para as próximas etapas da análise, a fim de garantir que cada município tenha um resultado.

3.4. MODELO INFERENCIAL

3.4.1. SHAP

A importância de variáveis e sua relação com a predição, foram avaliadas usando os valores SHAP (Explicação Aditiva de Shappley) (Lundberg and Lee, 2017).

3.5. ESTATÍSTICA ESPACIAL

3.5.1. Moran

Para obter a estatística de Moran (Anselin, 2010) foi primeiro obtida a matriz de vizinhança tipo Queen para todos os município do Brasil utilizando o arquivo *shape* do repositório online do Instituto Brasileiro de Geografia e Estatística (IBGE, 2018). Existem dois municípios no Brasil cujos territórios estão localizados em ilhas (Ilha Bela e Fernando de Noronha); para esses, foi utilizado *K-Near Neighbors* (KNN) para completar a matriz de vizinhança que, no tipo Queen, retornou nula para esses municípios.

Para a análise gráfica foram selecionados os municípios que tinham o resíduo superior à média mais um desvio padrão, além de um p-valor para a estatística de Moran Local inferior a 0,05 e com seus vizinhos com valores de resíduo superior à média.

3.5.2. Kulldorff

Os resíduos da predição dos algoritmos de aprendizado de máquina foi usado para identificar grupos geográficos de taxas de mortalidade por câncer mais altas do que o esperado com as estatísticas de varredura de Kulldorff (Kulldorff,

1997). O único parâmetro na estatística de varredura de Kulldorff é o tamanho máximo do cluster, a ser determinado pelo território espacial ou pela população em risco. Kulldorff relatou que uma janela de varredura de até 50% da população em risco é a regra ideal para evitar a detecção de agrupamento negativo (Kulldorff and Nagarwalla, 1995). Devido ao baixo valor e alta variância da densidade populacional no Brasil, um tamanho de cluster de 0,3% do total da população brasileira foi definido a priori para capturar uma extensão territorial relevante para o objetivo deste estudo.

4. RESULTADOS

4.1. ANÁLISE DESCRITIVA DE VARIÁVEIS

O resumo das variáveis analisadas pode ser observado na Tabela 2.

Tabela 2 - Variáveis Utilizadas no Modelo e Distribuição

		Total (N=5565)	
Variável	**Origem**	**Média (Desvio Padrão)**	**Min. - Máx.**
Expectativa de Vida (anos)	Censo 2010	73,0 (2,7)	65,3 - 78,6
Residentes	Censo 2010	34.277,8 (203.112,6)	805 – 11.253.503
Mediana de Idade (anos)	Censo 2010	29,0 (4,4)	13,6 - 46,7
Proporção de Idosos (%)	Censo 2010	12,1 (3,3)	2,6 - 29,2
Razão de Dependência (dependentes * 100/ativos)	Censo 2010	60,3 (9,0)	15,9 - 123,0
Proporção de Mulheres (%)	Censo 2010	49,5 (1,6)	18,9 - 54,2
Nascimentos per capita	Censo 2010	13,9 (3,6)	3,4 - 47,5
Proporção de Casados (%)	Censo 2010	36,5 (8,7)	6,6 - 64,0
Proporção de Evangélicos (%)	Censo 2010	17,1 (9,4)	0,4 - 85,8
Índice de Deficiência	Censo 2010	24,6 (4,7)	2,5 - 44,3
Densidade Municipal (Habitantes / km^2)	Censo 2010	108,2 (572,4)	0,1 - 13.024,6
Proporção de Área urbana (%)	Censo 2010	63,8 (22,0)	4,2 - 100
Proporção de casas com geladeira (%)	Censo 2010	88,6 (11,6)	16,7 - 100
Proporção de casas com computadores (%)	Censo 2010	21,3 (14,0)	0,4 - 72,7
Proporção de casas com automóveis (%)	Censo 2010	31,9 (19,9)	0,0 - 90,7
Densidade de Casas (Casas / km^2)	Censo 2010	3,4 (0,4)	2,6 - 6,9
Proporção de residentes em favelas (%)	Censo 2010	0,1 (0,4)	0,0 - 9,9
Proporção de casas com eletricidade (%)	Censo 2010	97,0 (5,8)	29,5 - 100,0
Proporção de áreas verdes (%)	Censo 2010	43,3 (24,5)	0,0 – 98,0
Proporção de Ruas Pavimentadas (%)	Censo 2010	46,5 (23,3)	0,0 – 99,0
Proporção de Brancos (%)	Censo 2010	46,95 (24,0)	0,86 - 99,2
Proporção de Alfabetizados (%)	Censo 2010	85,3 (8,9)	58,4 - 99,1
Proporção de Universitários (%)	Censo 2010	5,5 (3,3)	0,3 - 33,8
Proporção de Concluintes do Ensino Médio (%)	Censo 2010	16,2 (6,0)	1,9 - 47,5
Proporção de Migrantes (%)	Censo 2010	5,5 (4,5)	0,0 - 45,8

Proporção de Estrangeiros (%)	Censo 2010	0,1 (0,5)	0,0 - 37,7
Renda Mediana (R$)	Censo 2010	566,1 (265,7)	128,7 - 2210,7
Taxa de Desemprego (%)	Censo 2010	3,7 (2,0)	0,0 - 17,0
Trabalho Infantil (%)	Censo 2010	13,0 (8,3)	0,0 - 72,1
Proporção de Aposentados (%)	Censo 2010	16,7 (4,7)	2,3 - 40,2
Índice de Horas Extras (%)	Censo 2010	28,6 (10,9)	0,9 - 73,1
Proporção de Crianças Pobres (%)	Censo 2010	59,9 (22,5)	2,4 - 95,5
Renda Per Capita (R$)	Censo 2010	12.587,9 (14.676,8)	2.258,0 – 31.2257,3
Coeficiente de Gini	Censo 2010	0,5 (0,1)	0,3 - 0,8
Cobertura de Bolsa Família (%)	Censo 2010	75,9 (18,7)	0,0 - 100,0
Residentes trabalhando fora da cidade (%)	Censo 2010	10,8 (9,8)	0,0 - 69,3
Proporção de Residentes cobertos por planos de Saúde (%)	ANS 2010	9,1 (11,8)	0,1 – 100,0
Taxa de mamografias por 100 mulheres	MS 2010	0,1 (0,1)	0,0 - 2,5
Estratégia de Cobertura de Saúde Bucal (%)	MS 2010	65,4 (37,2)	0,0 - 100,0
Cobertura de Saúde Primária (%)	MS 2010	79,6 (27,2)	0,0 - 100,0
Cobertura Vacinal (doses / população alvo)	MS 2010	79,1 (10,7)	0,0 - 182,4
Taxa de nascidos abaixo do peso (%)	MS 2010	7,7 (3,5)	0,0 - 40,0
Leitos Hospitalares por 10.000 Habitantes	MS 2010	1,5 (1,8)	0,0 - 24,5
Times de estratégia de Saúde Familiar por 10.000 Habitantes	MS 2010	0,3 (0,1)	0,0 - 1,2
Máquinas de Ultrassom por 10.000 Habitantes	MS 2010	4,6 (8,3)	0,0 - 125,0
Proporção de Cesarianas (%)	MS 2010	50,9 (18,4)	3,3 - 100,0
Máquinas de raio X por 10.000 Habitantes	MS 2010	0,1 (0,1)	0,0 - 1,4
Equipamentos de Suporte à vida por 10.000 Habitantes	MS 2010	0,3 (0,4)	0,0 - 3,2
Proporção de Óbitos mal definidos (%)	SIM 2007-2016	3,1 (3,2)	0,0 - 45,2

4.2. RESULTADOS DO MODELO PREDITIVO

4.2.1. Seleção do Modelo

Na fase de seleção dos modelos, os algoritmos com melhor desempenho foram: RF (R^2 0,651, IC 95% 0,640-0,662), XGB (R^2 0,626, IC 95% 0,615-0,637), pSVM (R^2 0,599, IC 95% 0,576-0,622) e LASSO (R^2 0,588, 95% CI 0,578-0,598). Os resultados para cada algoritmo na fase de seleção são apresentados na Figura 7.

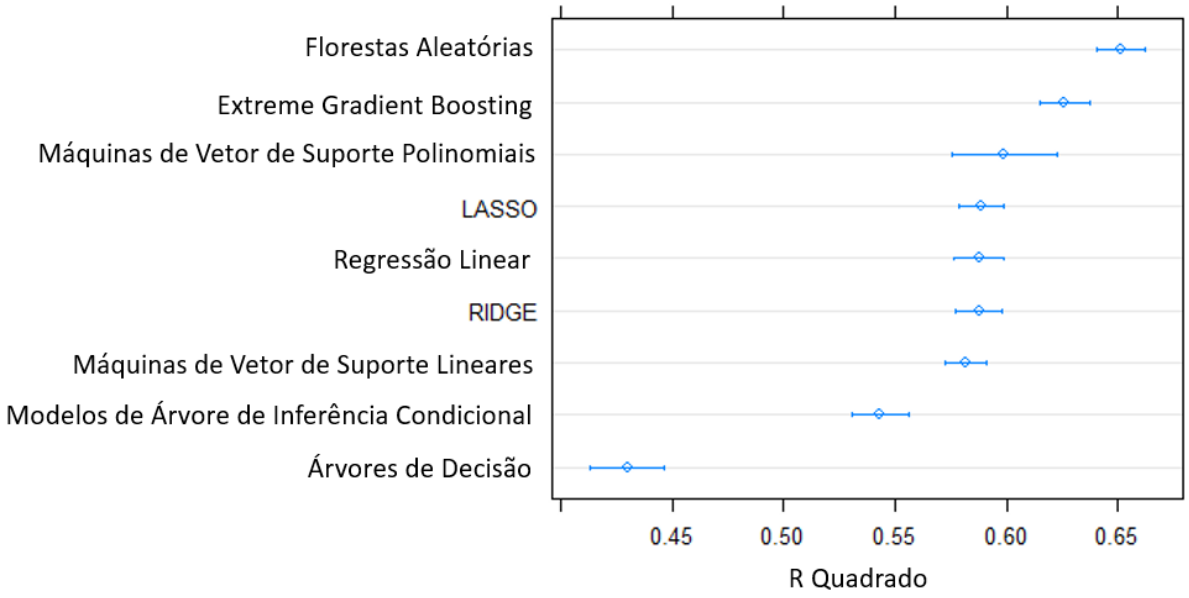
4.2.2. Avaliação Geral

Considerando-se todos os tipos de câncer, o gráfico de dispersão para os diferentes modelos de *machine learning* podem ser observados na Figura 8 e os valores de R^2 podem ser observados na Tabela 5.

Considerando-se todos os tumores, o algoritmo *eXtreme Gradient Boosting* (XGB) foi o que apresentou melhor resultado na avaliação do modelo final obtendo um R^2 no valor de 0,66, seguido do algoritmo Florestas Aleatórias com um R^2 de 0,65.

Considerando tumores específicos, o melhor algoritmo para câncer de pulmão foi o *random forests* (RF) com R^2 de 0,53 na avaliação do modelo final. O algoritmo de RF foi também o melhor para câncer colorretal (R^2 0,43), esôfago (R^2 0,32) e pâncreas (R^2 0,27).

Figura 7 - Resultado da Seleção Primária de Modelos no conjunto de treino com intervalos de confiança de 95%.

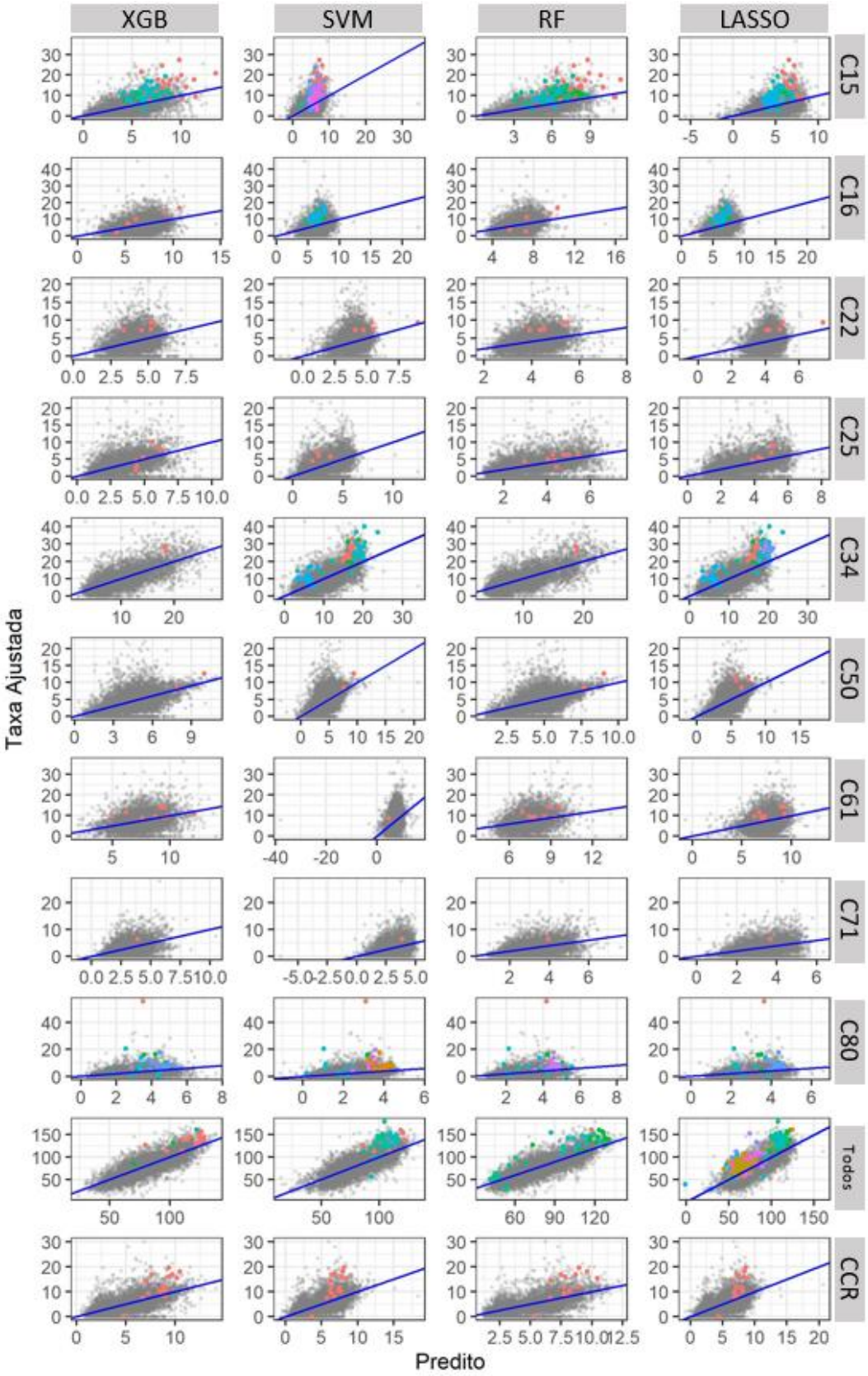


4.3. RESULTADOS DA EXPLICAÇÃO ADITIVA DE SHAPLEY (SHAP)

A análise da explicação aditiva de *Shapley* (SHAP) encontrou que a variável mais importante para o resultado preditivo foi a proporção de brancos (Figura 9) que obteve uma contribuição positiva, crescente e linear por toda a distribuição da variável (Figura 10).

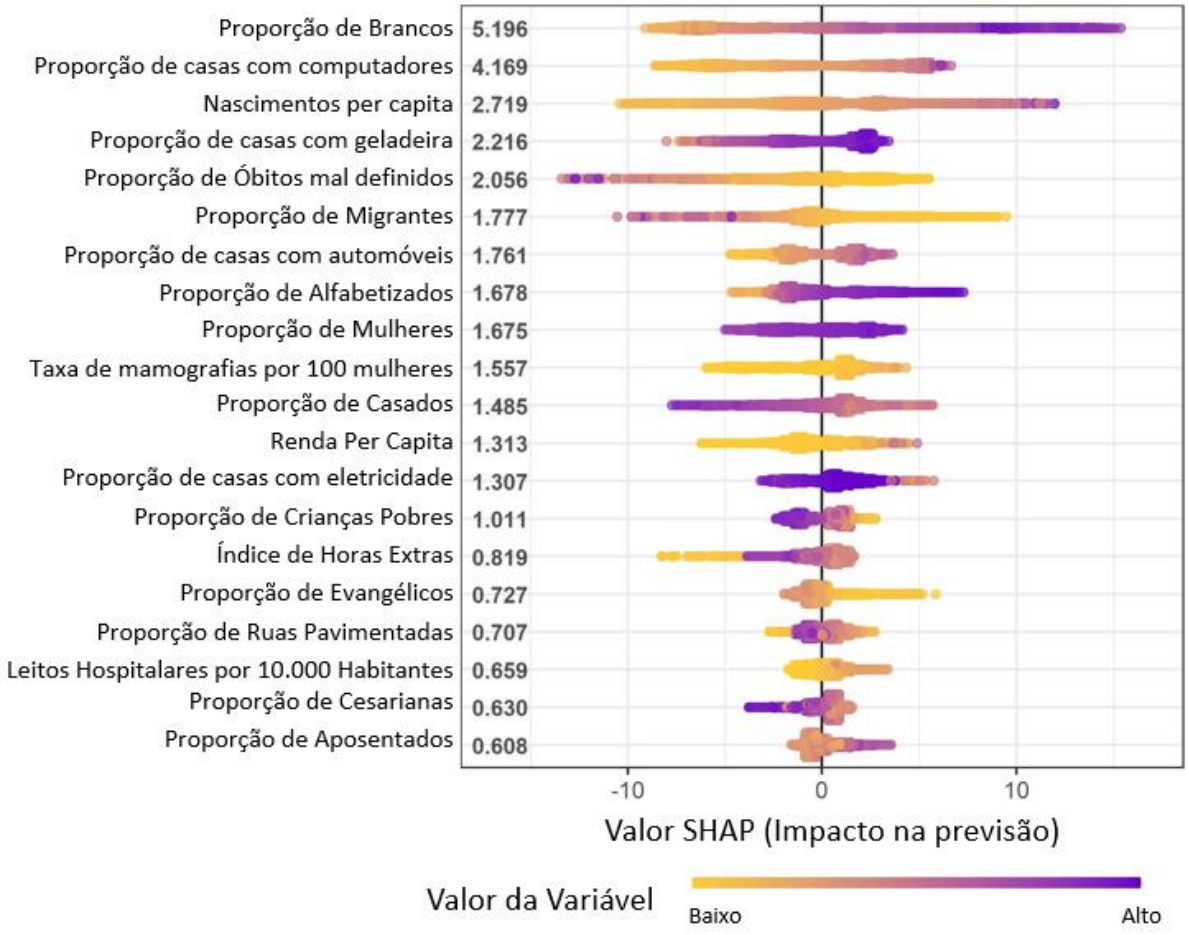
A segunda variável mais importante foi a proporção de casas com computadores, que apresentou uma relação não-linear positiva, mantendo-se linear e crescente até por volta de 30%, estabilizando a contribuição em 5 mortes adicionais por 100.000 acima deste patamar.

Figura 8 - Gráfico de dispersão da taxa ajustada de mortalidade e o valor predito pelo modelo por município do Brasil, tipo de câncer e algoritmo de Machine Learning. Cores identificando conjuntos de aglomerados espaciais.



A taxa de nascimentos per capita foi a terceira variável de mais importante contribuição, de forma positiva, não linear e crescente até por volta de 25 nascimentos por mil habitantes quando a contribuição adquiriu comportamento assintótico por volta de 10 mortes adicionais por 100.000 habitantes.

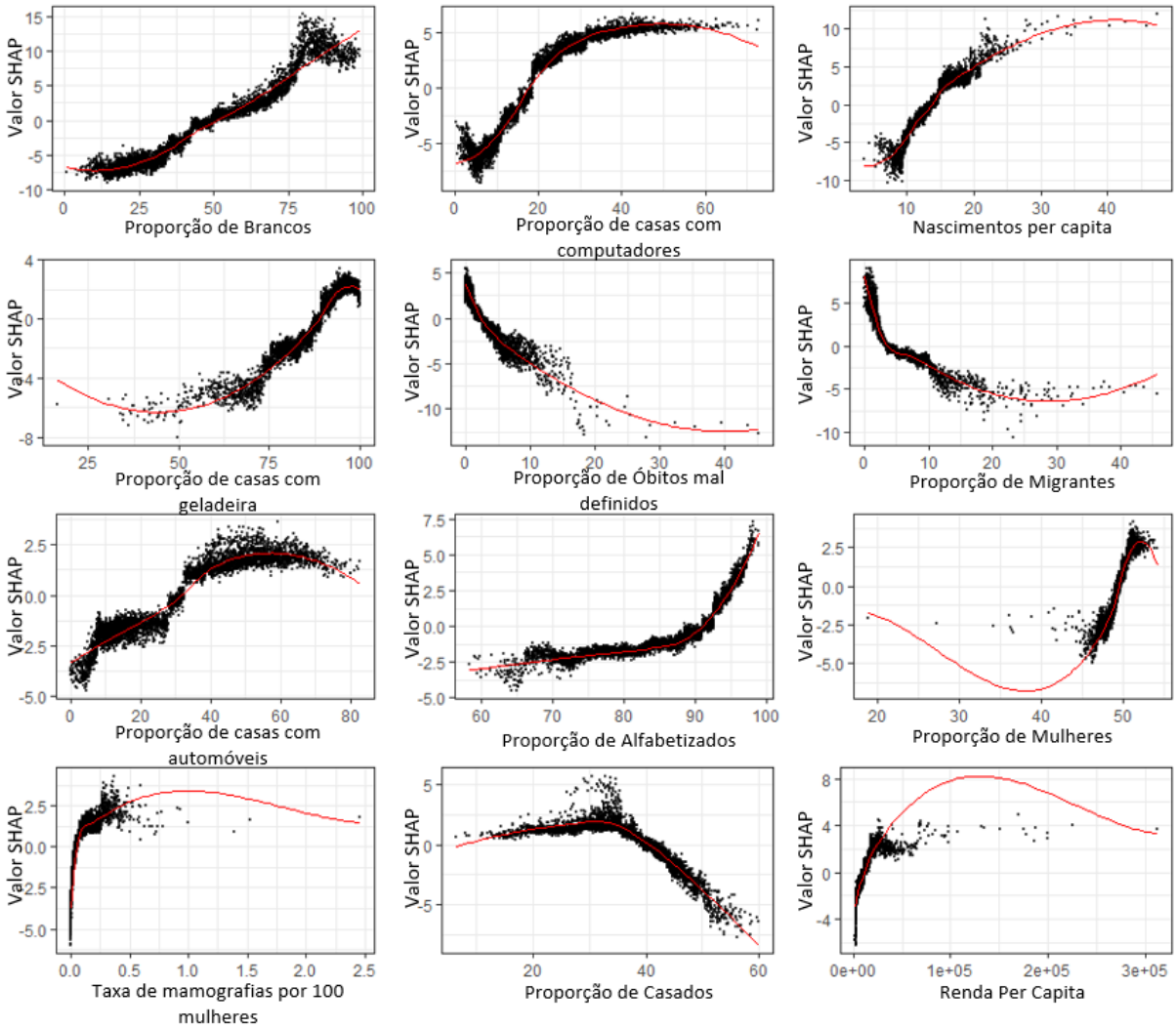
Figura 9 - Gráfico de importância de variáveis com as 20 mais importantes variáveis para o modelo eXtreme Gradient Boosting (XGB) para o agregado de todos os tumores.



A proporção de casas com geladeira também mostrou uma relação positiva e crescente com a predição. Abaixo de 60% essa variável apresentou um comportamento assintótico ao redor de -6 mortes adicionais por 100.000 habitantes.

A proporção de óbitos mal definidos foi a quinta variável mais importante com uma contribuição negativa em toda a sua extensão, tendo a mínima contribuição em -10 mortes por 100.000 habitantes em cidades com até 40% das mortes mal definidas.

Figura 10 - Gráfico detalhando o comportamento das 9 principais variáveis para o modelo de eXtreme Gradient Boosting (XGB) para o agregado de todos os tumores em relação à contribuição para o modelo (Valor SHAP).



4.4. RESULTADOS DA ANÁLISE DE RESÍDUOS

A Figura 11 mostra o mapa do Brasil com os resíduos da taxa de mortalidade para o agregado de todos os cânceres obtido a partir do modelo XGB. A Figura 12 e a Tabela 3 apresentam o resultado agregado desses resíduos por estado.

No nível estadual, o R^2 entre a taxa de mortalidade ajustada real e a obtida pelo modelo foi de 0,84. Os estados da Região Sul do Brasil tiveram as maiores taxas, tanto previstas quanto reais. O estado com o maior desvio positivo foi o Amazonas com um desvio de 21,5% acima do predito e o negativo foi o estado do Alagoas com um desvio de 12,3% abaixo do predito para esse estado.

A Figura 13 mostra o excesso de mortalidade para o modelo de XGB para o consolidado de todos cânceres, mostrando a cidade de Manaus com a maior taxa de mortalidade não explicada pelo modelo.

Figura 11 - Gráfico de resíduos do modelo de eXtreme Gradient Boosting (XGB) para o agregado de todos os cânceres.

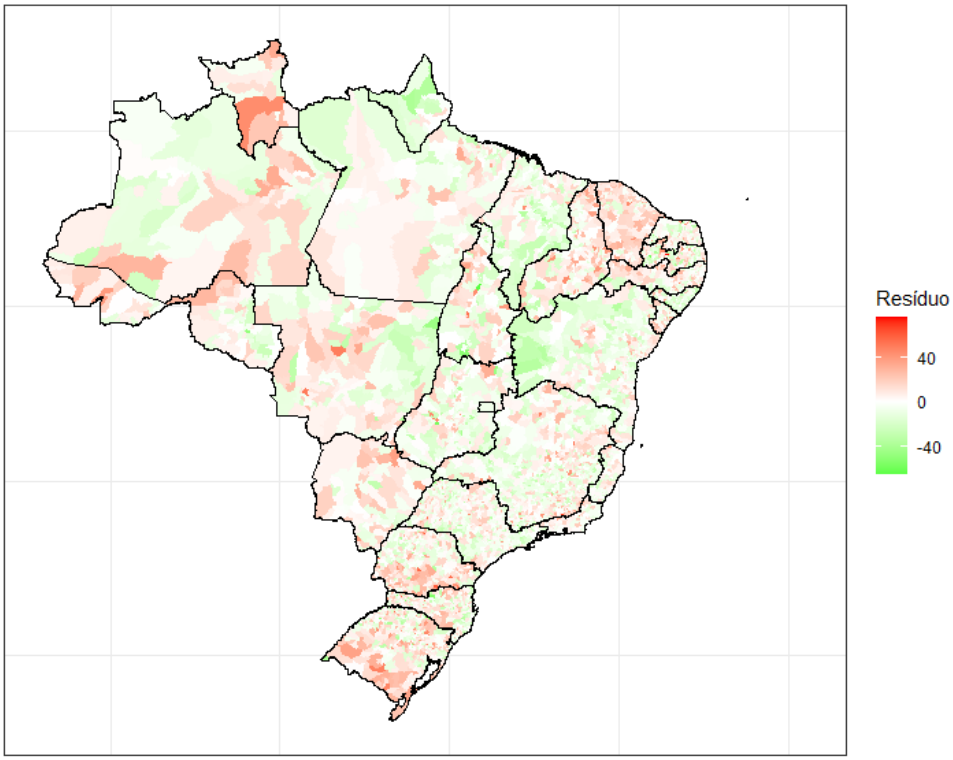


Figura 12 - Resíduos do modelo de eXtreme Gradient Boosting (XGB) para o agregado de todos os cânceres consolidados no nível estadual.

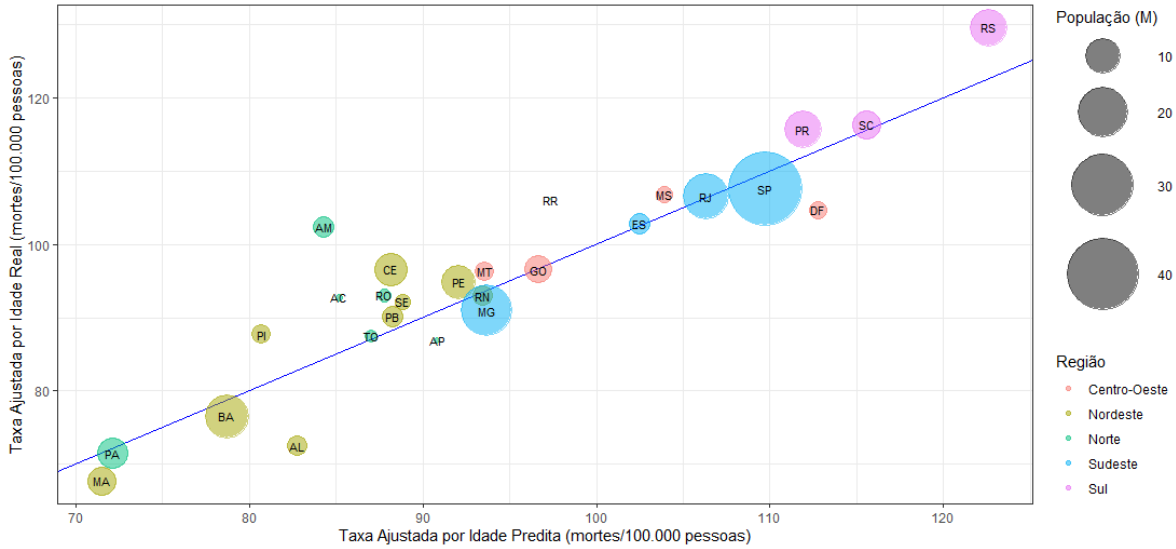
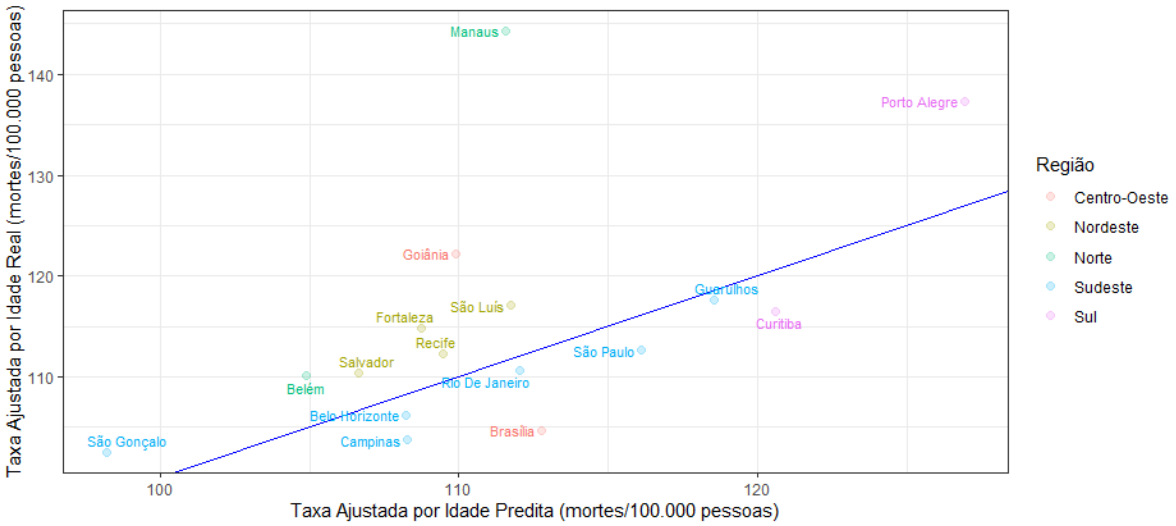


Tabela 3 - Resíduos do modelo de eXtreme Gradient Boosting (XGB) para o agregado de todos os cânceres agregados no nível estadual.

Região	UF	População	Mortes Previstas	Mortes Reais	Taxa Prevista (Mortes / 100.000)	Taxa Real (Mortes / 100.000)	Taxa Residual (Mortes / 100.000)
Centro-Oeste	MS	2.549.296	2.649	2.722	103,9	106,8	2,9
Centro-Oeste	MT	3.138.822	2.936	3.022	93,5	96,3	2,7
Centro-Oeste	GO	6.343.136	6.131	6.125	96,7	96,6	-0,1
Centro-Oeste	DF	2.727.098	3.075	2.855	112,8	104,7	-8,1
Centro-Oeste		**14.758.352**	**14.791**	**14.723**	**100,2**	**99,8**	**-0,5**
Nordeste	CE	8.711.659	7.680	8.414	88,2	96,6	8,4
Nordeste	PI	3.172.210	2.560	2.782	80,7	87,7	7,0
Nordeste	SE	2.171.137	1.929	2.002	88,9	92,2	3,3
Nordeste	PE	9.136.697	8.413	8.666	92,1	94,8	2,8
Nordeste	PB	3.883.822	3.427	3.499	88,2	90,1	1,9
Nordeste	RN	3.338.489	3.120	3.106	93,4	93,0	-0,4
Nordeste	BA	14.957.177	11.774	11.451	78,7	76,6	-2,2
Nordeste	MA	6.734.353	4.815	4.558	71,5	67,7	-3,8
Nordeste	AL	3.279.289	2.714	2.378	82,8	72,5	-10,3
Nordeste		**55.384.833**	**46.432**	**46.856**	**83,8**	**84,6**	**0,8**
Norte	AM	3.740.976	3.152	3.828	84,3	102,3	18,1
Norte	RR	479.073	467	508	97,4	106,1	8,7
Norte	AC	762.631	650	708	85,2	92,8	7,6
Norte	RO	1.707.272	1.499	1.589	87,8	93,1	5,3
Norte	TO	1.458.965	1.270	1.276	87,0	87,5	0,4
Norte	PA	7.847.213	5.659	5.610	72,1	71,5	-0,6
Norte	AP	718.906	653	625	90,8	86,9	-3,9
Norte		**16.715.036**	**13.349**	**14.144**	**79,9**	**84,6**	**4,8**
Sudeste	ES	3.792.874	3.888	3.900	102,5	102,8	0,3
Sudeste	RJ	16.273.984	17.299	17.340	106,3	106,6	0,3
Sudeste	SP	43.281.358	47.494	46.569	109,7	107,6	-2,1
Sudeste	MG	20.446.840	19.150	18.594	93,7	90,9	-2,7
Sudeste		**83.795.056**	**87.830**	**86.403**	**104,8**	**103,1**	**-1,7**
Sul	RS	11.115.607	13.626	14.408	122,6	129,6	7,0
Sul	PR	10.910.374	12.206	12.620	111,9	115,7	3,8
Sul	SC	6.519.554	7.536	7.588	115,6	116,4	0,8
Sul		**28.545.535**	**33.368**	**34.616**	**116,9**	**121,3**	**4,4**
Brasil		**199.198.812**	**195.770**	**196.742**	**98,3**	**98,8**	**0,5**

Figura 13 - Resíduos do modelo de eXtreme Gradient Boosting (XGB) para o agregado de todos os cânceres para os Municípios do Brasil com mais de 1 milhão de habitantes.



4.5. RESULTADOS DO TESTE DE MORAN

Através da matriz de vizinhança por contiguidade tipo Queen, adicionada de aproximação por *K-Nearest Neighbors* para os municípios de Ilha Bela e Fernando de Noronha, foram obtidos uma média de 5,9 municípios vizinhos por cidade do Brasil.

A Figura 14 mostra um gráfico de Moran com o resíduo para o modelo obtido pelo método de XGB para todos os tumores agregados. Pode-se observar uma tendência positiva para autocorrelação espacial.

Na Figura 15 foram selecionados os municípios com resíduo para o modelo obtido pelo método de XGB para todos os tumores agregados superior à média mais um desvio padrão e que possuíam um p-valor de associação inferior a 0,05 e cercados por vizinhos com resíduo superior à média. Foram observados diversos municípios com essas características principalmente concentrados nos estados do Rio Grande do Sul (Figura 16) e no Paraná (Figura 17).

Os resultados obtidos pelo método de Moran não permitem a avaliação do aglomerado espacial, apenas a avaliação de cada cidade em relação a seus vizinhos. Buscou-se, portanto, na sequência um método que permitisse uma estatística única para todo o aglomerado espacial.

Figura 14 - Gráfico de Moran, mostrando a correlação espacial positiva entre a mortalidade adicional por câncer além da prevista pelo modelo de eXtreme Gradient Boosting para todos os tumores agregados e o adicional de mortalidade dos municípios vizinhos.

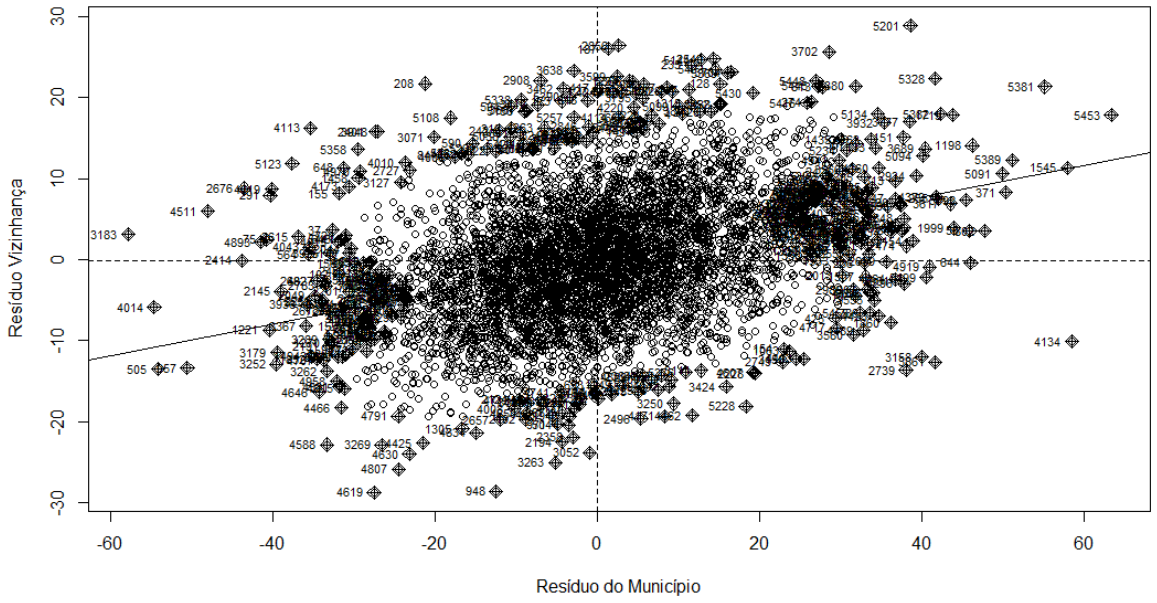


Figura 15 - Mapa do Brasil com os municípios identificados pelo método de Moran (LISA) com alta taxa de mortalidade adicional por câncer em relação ao modelo de eXtreme Gradient Boosting para todos os tumores agregados e que estejam cercados por municípios com altas taxas adicionais (superior à média mais um desvio padrão).

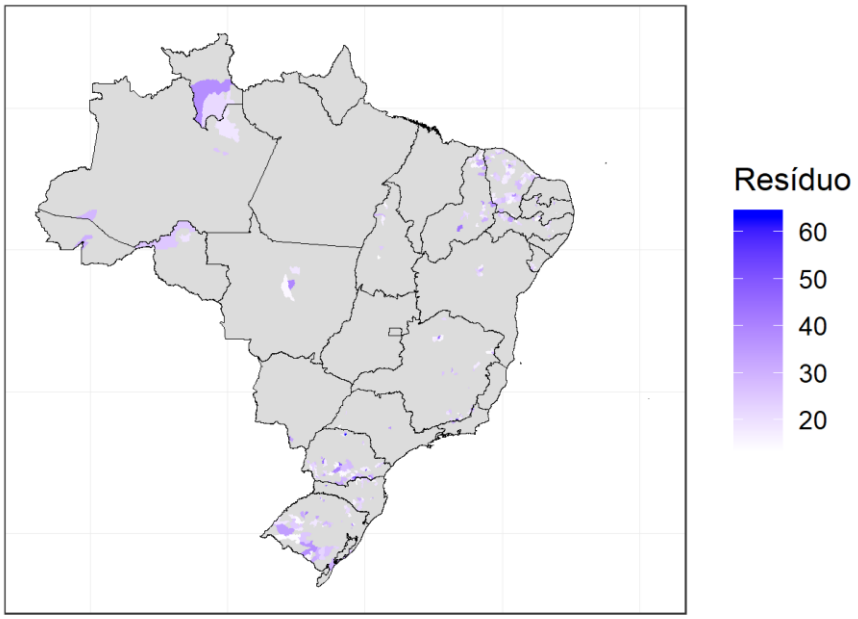


Figura 16 - Aproximação do gráfico anterior para as cidades do estado do Rio Grande do Sul.

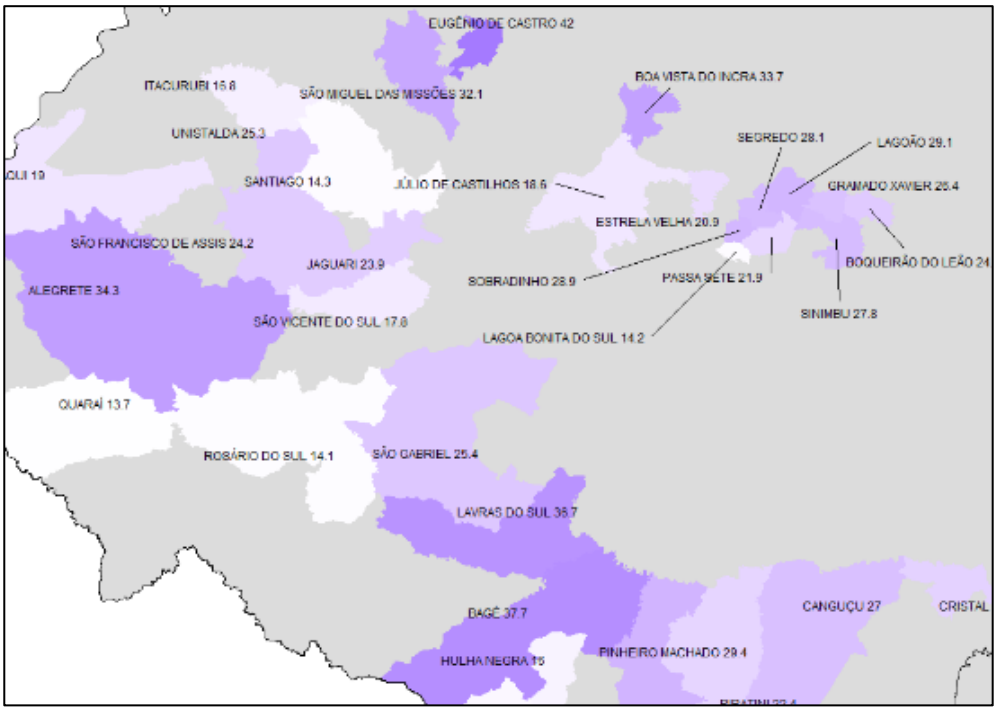
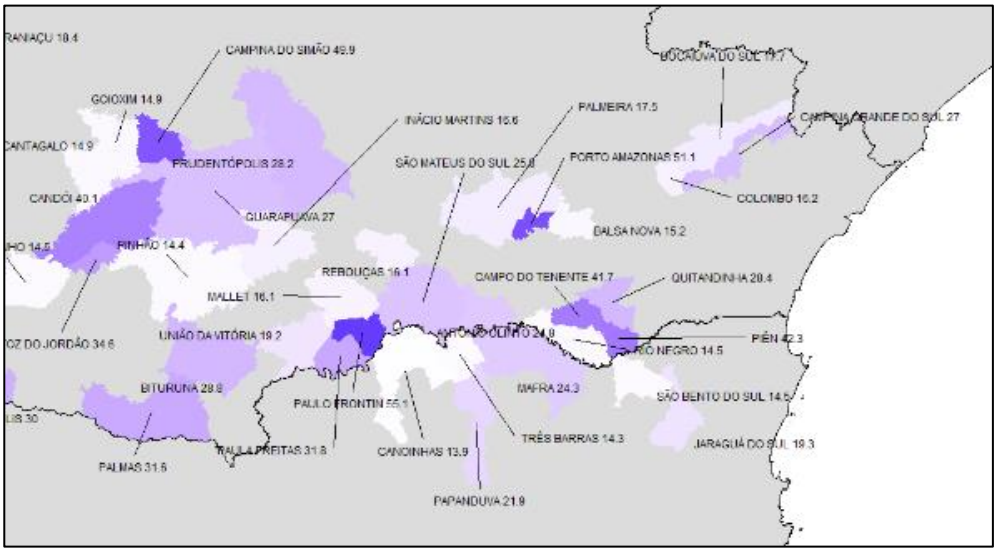


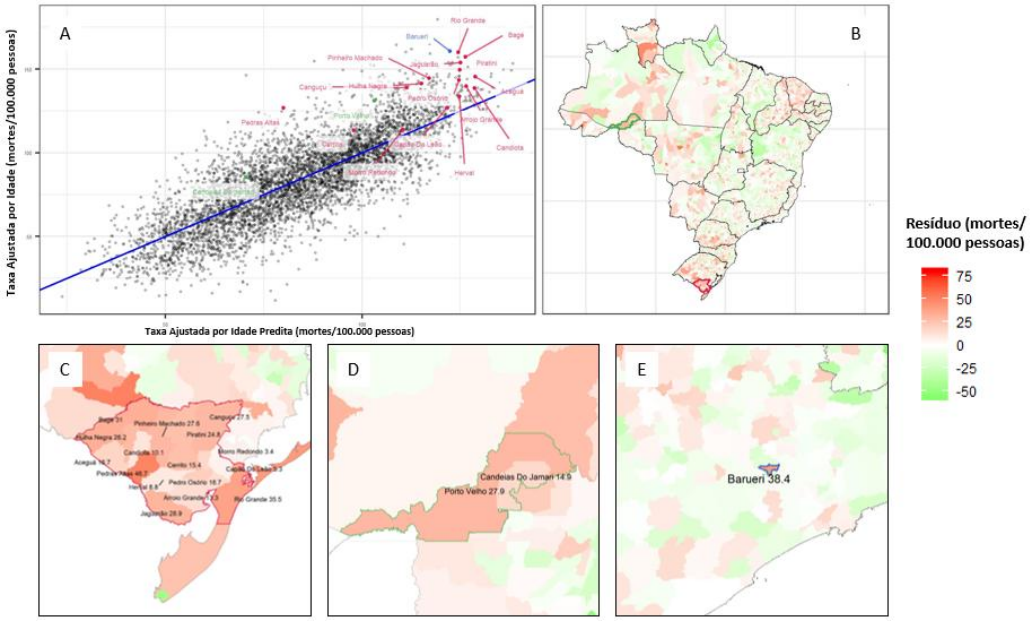
Figura 17 - Aproximação para os estados do Paraná e norte de Santa Catarina.



4.6. VARREDURA ESPACIAL DE KULLDORFF

O método de varredura espacial de Kulldorff possibilitou a identificação de três aglomerados espaciais considerando o modelo de XGB para todos os tumores agregados.

Figura 18 - Resultados do modelo XGB: (A) Gráfico de correlação com R^2 0,66 para Mortalidade Ajustada por Câncer por 100.000 habitantes em Municípios Brasileiros (colorido por clusters Kulldorff) (B) Resíduos plotados no mapa brasileiro com clusters identificados por cores. Amplie Barueri (C), Cluster Bagé-Rio Grande (D) e Cluster Porto Velho e Arredores (E).



O cluster primário, com o menor valor de p (p = 0,001), foi a região entre Rio Grande e Bagé no estado do Rio Grande do Sul (RS), com excesso de 28,6 óbitos por 100.000 habitantes (p = 0,001). Os clusters secundários situaram-se na região de Porto Velho no estado de Rondônia (RO), com excesso de 27,3 óbitos por 100.000 habitantes (p = 0,001), e na cidade de Barueri no estado de São Paulo (SP) com uma taxa de mortalidade excessiva de 38,4 mortes por 100.000 residentes (detalhes na Tabela 4).

Tabela 4 - Taxas de mortalidade por câncer por cluster identificado pela varredura espacial de Kulldorff e Municípios pelo modelo de XGB.

Aglomerado/Município	Taxa Ajustada (mortes por 100.000)	Taxa Predita (mortes por 100.000)	Resíduo (mortes por 100.000)	População (milhares)	Casos adicionais
Aglomerado 1	**151,3**	**122,7**	**28,6**	**538,2**	**154,1**
Rio Grande	160,0	124,5	35,5	205,2	72,8
Bagé	157,3	126,3	31,0	121,0	37,5
Canguçu	138,9	111,4	27,5	55,3	15,2
Jaguarão	154,0	125,1	28,9	28,6	8,3

Capão Do Leão	127,0	121,7	5,3	25,2	1,3
Piratini	149,6	124,9	24,8	20,6	5,1
Arroio Grande	139,8	126,5	13,3	19,0	2,5
Pinheiro Machado	144,6	117,0	27,6	13,1	3,6
Candiota	138,6	128,6	10,1	9,2	0,9
Pedro Osório	143,4	124,7	18,7	8,0	1,5
Herval	133,6	124,8	8,8	7,0	0,6
Cerrito	113,3	97,9	15,4	6,5	1,0
Morro Redondo	113,8	110,4	3,4	6,5	0,2
Hulha Negra	141,3	115,1	26,2	6,3	1,7
Aceguá	145,5	128,8	16,7	4,6	0,8
Pedras Altas	126,8	80,1	46,7	2,2	1,0
Aglomerado 2	**129,0**	**101,7**	**27,3**	**498,0**	**136,0**
Porto Velho	131,0	103,2	27,9	475,7	132,7
Candeias Do Jamari	85,4	70,5	14,9	22,4	3,3
Aglomerado 3	**160,7**	**122,4**	**38,4**	**253,9**	**97,4**
Barueri	160,7	122,4	38,4	253,9	97,4
Brasil - Outros	**98,5**	**98,2**	**0,3**	**197.908,7**	**584,6**

Comparando diferentes algoritmos para os tipos específicos de câncer, os algoritmos RF e XGB tiveram a melhor performance geral de predição. Foram encontrados aglomerados espaciais com o método de varredura de Kulldorff para os tumores de esôfago, estômago, colorretal, pulmão e tumores de localização não especificada. Nenhum cluster significativo (p<0,05) foi encontrado para os tumores de fígado, pâncreas, mama, próstata e cérebro em nenhum dos modelos especificados (detalhes na Tabela 5).

O maior número de clusters foi identificado pelo algoritmo de LASSO.

Tabela 5 - Performance do modelo e número de aglomerados identificados para cada tipo de câncer e algoritmo. XGB - eXtreme Gradient Boosting, RF - Florestas Aleatórias, pSVM - Máquinas de Vetores de Suporte Polinomiais, LASSO - Regressão Linear com LASSO.

Tipo de Câncer	Algoritmo	R2 no Conjunto Geral	Número de Clusters identificados	Número de Clusters identificados (p<0.05)
Todos os Cânceres	XGB	0,66	3	3
Todos os Cânceres	RF	0,65	4	4
Todos os Cânceres	pSVM	0,61	4	4
Todos os Cânceres	LASSO	0,59	12	12

C15 - Esôfago	XGB	0,32	5	5
C15 - Esôfago	RF	0,32	5	5
C15 - Esôfago	pSVM	0,30	8	8
C15 - Esôfago	LASSO	0,24	5	5
C16 – Estômago	XGB	0,13	1	1
C16 – Estômago	RF	0,15	1	1
C16 – Estômago	pSVM	0,08	5	5
C16 – Estômago	LASSO	0,08	5	5
C18-C21 Colorretal	XGB	0,42	1	0
C18-C21 Colorretal	RF	0,43	1	0
C18-C21 Colorretal	pSVM	0,41	1	1
C18-C21 Colorretal	LASSO	0,41	1	1
C22 – Fígado	XGB	0,07	1	0
C22 – Fígado	RF	0,08	1	0
C22 – Fígado	pSVM	0,07	1	0
C22 – Fígado	LASSO	0,07	1	0
C25 – Pâncreas	XGB	0,25	1	0
C25 – Pâncreas	RF	0,27	1	0
C25 – Pâncreas	pSVM	0,25	1	0
C25 – Pâncreas	LASSO	0,25	1	0
C34 – Pulmão	XGB	0,51	1	0
C34 – Pulmão	RF	0,53	1	0
C34 – Pulmão	pSVM	0,45	5	5
C34 – Pulmão	LASSO	0,45	7	7
C50 – Mama	XGB	0,24	1	0
C50 – Mama	RF	0,26	1	0
C50 – Mama	pSVM	0,24	1	0
C50 – Mama	LASSO	0,24	1	0
C61 – Próstata	XGB	0,11	1	0
C61 – Próstata	RF	0,12	1	0
C61 – Próstata	pSVM	0,08	1	0
C61 – Próstata	LASSO	0,07	1	0
C71 - Cérebro	XGB	0,15	1	0
C71 - Cérebro	RF	0,16	1	0
C71 - Cérebro	pSVM	0,15	1	0
C71 - Cérebro	LASSO	0,15	1	0
C80 – Local. N. Especificada	XGB	0,11	7	7
C80 – Local. N. Especificada	RF	0,12	8	8
C80 – Local. N. Especificada	pSVM	0,09	10	10
C80 – Local. N. Especificada	LASSO	0,09	7	7

A Figura 19 e a Tabela 6 fornecem detalhes sobre os aglomerados e suas sobreposições identificados por cada modelo e tipos de câncer. Para o grupo de todos os cânceres, um aglomerado composto por 16 cidades (Figura 19 H1) foi identificado por quatro algoritmos na região entre as cidades de Bagé e Rio Grande com uma taxa de mortalidade 27% superior à média prevista pelos modelos. Nessa mesma região um aglomerado para câncer colorretal foi identificado por 2 modelos (Figura 19 I5) com um excesso de 70% em relação ao predito.

Vários aglomerados sobrepostos do grupo de todos os cânceres, câncer de pulmão e câncer de estômago foram identificados na Região de Porto Velho e arredores (Figura 19 D). Dois desses aglomerados, relacionados a todos os cânceres, se sobrepuseram, sendo que o primeiro (Figura 19 D2) apresentou um excesso de mortalidade 27% superior ao predito pelo modelo XGB e o segundo (Figura 19 D1) um excesso de 40% em relação à média predita pelos modelos LASSO e pSVM.

Na região de Macapá, foram identificados dois aglomerados de câncer de estômago sobrepostos na cidade de Macapá: o primeiro, identificado pelos modelos XGB, pSVM e LASSO (Figura 19 B2), mostrou uma taxa de mortalidade 82% superior à média predita pelos modelos. O segundo aglomerado, que inclui a cidade de Santana, foi identificado pelo modelo RF (Figura 19 B1) com uma taxa de mortalidade 58% superior à predita pelo modelo.

Diversos aglomerados foram identificados no estado do Ceará para diferentes tipos de câncer, pelos modelos LASSO e pSVM (Figura 19 C). Para o conjunto de todos os cânceres, foram identificados 4 aglomerados distintos com uma taxa de mortalidade em média 27% superior ao predito (Figura 19: C1, C3, C6 e C7) e todos estes aglomerados foram identificados pelo modelo LASSO. Para câncer de pulmão, foi identificado um aglomerado por dois modelos distintos, LASSO e pSVM, com uma taxa de mortalidade 75% superior à predita. Adicionalmente foram também identificados dois aglomerados distintos para câncer de estômago pelos modelos LASSO e pSVM, com um excesso de 85% na região mostrada na Figura 19 C4 e um excesso de 71% na região mostrada na Figura 19 C5.

No estado do Rio Grande do Sul foram também observados diversos aglomerados de câncer de pulmão. Na região de sua capital Porto Alegre (Figura 19: H5, H6 e H7), três aglomerados distintos foram observados ao redor da cidade com taxas adicionais de mortalidade com média de 43% acima do esperado.

Em relação a câncer de esôfago, foram identificados dois aglomerados sobrepostos na região entre Paraná e Santa Catarina (Figura 19: E4 e E5) por quatro modelos. Três aglomerados sobrepostos foram identificados na região oeste do estado do Rio Grande do Sul (Figura 19: I1, I2 e I3), e situação semelhante foi identificada na região do entorno da cidade de Teófilo Otoni (Figura 19: F2 e F3). Quatro aglomerados foram identificados no estado do Espírito Santo (Figura 19: F4, F5, F6 e F7). Câncer de esôfago foi o tipo de câncer avaliado neste estudo com o maior número de aglomerados identificados, com um total de 15.

Cânceres de localização não especificada tiveram um total de 11 aglomerados identificados, em seis deles por todos os algoritmos testados, que mostraram uma distribuição dispersa ao longo do território brasileiro com 4 deles compostos por apenas um município. O que mais se destacou foi o município de Itaperuçu-PR (Figura 19 E6), uma cidade de 25 mil habitantes que apresentou uma taxa de mortalidade 13 vezes superior à esperada pela média de todos os modelos.

A taxa média de excesso de mortalidade não explicável por características sociodemográficas foi mais alta para cânceres não especificados (média ponderada: 126% ± desvio padrão ponderado: 143% com peso nos casos esperados), seguida por câncer de estômago (73% ± 10%), colorretal câncer (um aglomerado com 70%), câncer de esôfago (67% ± 24%), câncer de pulmão (51% ± 11%) e todos os cânceres (28% ± 4%).

Detalhes sobre cada um dos aglomerados podem ser observados na Tabela 6.

Figura 19 - Localização dos aglomerados geográficos para vários tipos de câncer. ***(A)*** *Brasil.* ***(B)*** *Região de Macapá com dois aglomerados sobrepostos de Câncer de Estômago.* ***(C)*** *estado do Ceará com diferentes aglomerados para Todos os Cânceres, Câncer de Pulmão e Câncer de Estômago.* ***(D)*** *Região de Porto Velho com quatro aglomerados sobrepostos para Todos os Cânceres, Câncer de Pulmão e Câncer de Estômago.* ***(E)*** *Região do estado do Paraná com diferentes combinações de aglomerados para Todos os Cânceres, Câncer de Esôfago e Câncer de Localização Não Especificada.* ***(F)*** *Sudeste do Brasil com 7 aglomerados para câncer de esôfago.* ***(G)*** *estado do Rio Grande do Sul com combinação variada de aglomerados para diferentes tipos, especificados para Todos os Cânceres e de Pulmão* ***(H)*** *e Câncer de Esôfago, Colorretal e de Localização Não Especificada* ***(I)****.*

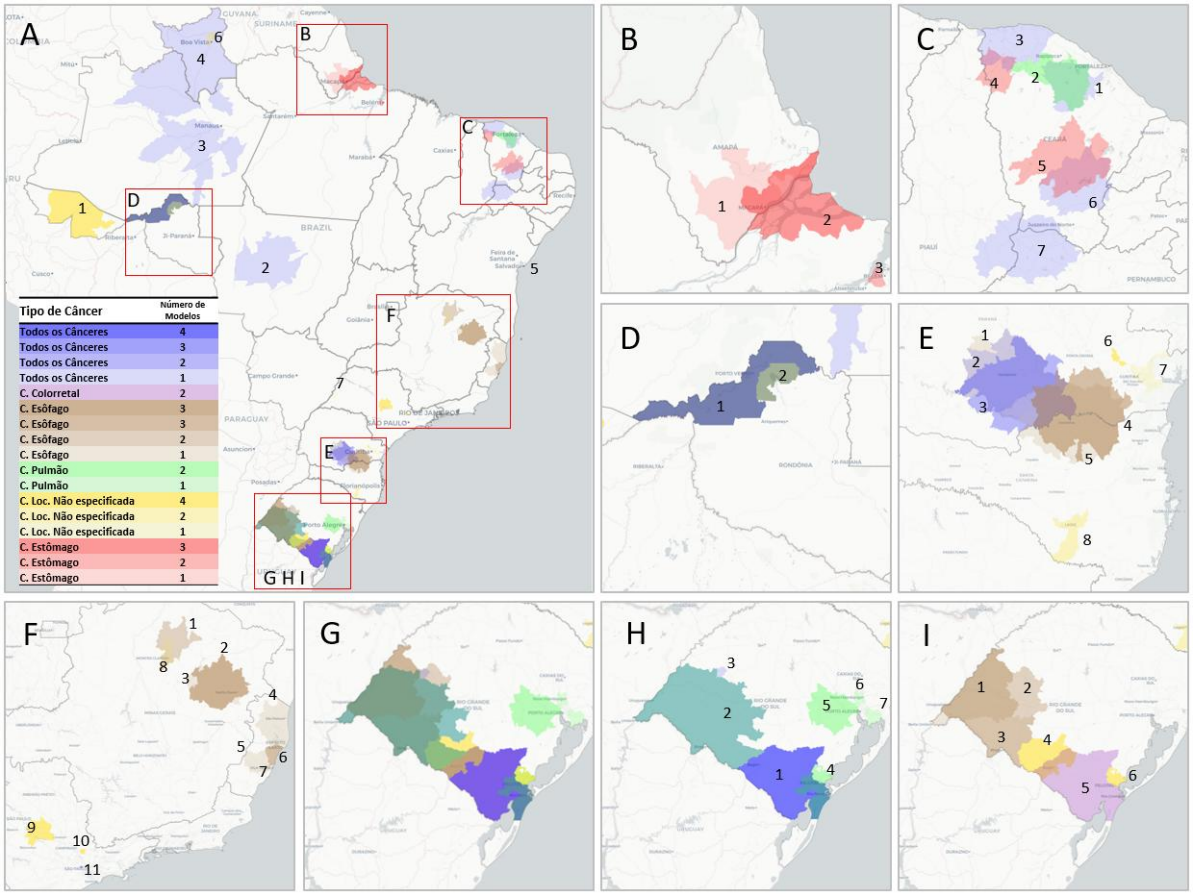


Tabela 6 - Detalhes dos aglomerados geográficos da Figura 19.

Tipo de Câncer	# de modelos	Loc. Fig. 15	Modelos	Número de Cidades	População	Média Estimada	Número de Casos	P-Valor min - max	Casos em excesso
Todos os Cânceres	4	H1	XGB, RF, pSVM, LASSO	16	538.220	643	815	0.001 - 0.002	27%
Todos os Cânceres	3	F10	XGB, RF, LASSO	1	253.877	309	408	0.001 - 0.005	32%
Todos os Cânceres	2	D1	pSVM, LASSO	1	475.691	447	623	0.001 - 0.001	40%
Todos os Cânceres	2	E3	pSVM, LASSO	20	587.156	641	820	0.001 - 0.001	28%
Todos os Cânceres	1	C1	LASSO	18	500.042	334	445	0.001 - 0.001	33%
Todos os Cânceres	1	A4	LASSO	18	556.312	416	550	0.001 - 0.001	32%
Todos os Cânceres	1	H2	LASSO	16	574.415	635	806	0.001 - 0.001	27%
Todos os Cânceres	1	C3	LASSO	23	592.999	392	496	0.002 - 0.002	27%
Todos os Cânceres	1	C7	LASSO	32	539.167	356	442	0.046 - 0.046	24%
Todos os Cânceres	1	C6	LASSO	26	586.597	428	537	0.002 - 0.002	25%
Todos os Cânceres	1	A2	LASSO	22	474.579	425	544	0.001 - 0.001	28%
Todos os Cânceres	1	A5	LASSO	2	308.748	296	375	0.047 - 0.047	27%
Todos os Cânceres	1	D2	XGB	2	498.045	507	642	0.001 - 0.001	27%
Todos os Cânceres	1	A3	RF	12	446.095	236	315	0.04 - 0.04	33%
Todos os Cânceres	1	E2	RF	13	401.607	455	564	0.047 - 0.047	24%
Todos os Cânceres	1	H(2,3)	pSVM	17	576.921	641	809	0.003 - 0.003	26%
C. Colorretal	2	I5	pSVM, LASSO	16	538.220	47	79	0.001 - 0.001	70%
C. Esôfago	3	F2	XGB, RF, LASSO	28	571.144	41	61	0.002 - 0.007	50%
C. Esôfago	3	E4	XGB, RF, LASSO	28	565.803	45	80	0.001 - 0.001	76%
C. Esôfago	3	F6	XGB, RF, LASSO	5	586.686	26	47	0.001 - 0.003	78%
C. Esôfago	2	F1	RF, pSVM	10	573.875	31	47	0.033 - 0.035	52%
C. Esôfago	2	I(1,2)	XGB, RF	16	584.172	44	65	0.001 - 0.004	48%
C. Esôfago	1	E5	pSVM	27	583.542	44	80	0.001 - 0.001	82%
C. Esôfago	1	E1	pSVM	15	431.730	32	48	0.033 - 0.033	50%
C. Esôfago	1	F	XGB	2	443.871	21	37	0.013 - 0.013	76%
C. Esôfago	1	F3	pSVM	31	591.580	47	64	0.017 - 0.017	36%
C. Esôfago	1	F4	LASSO	17	587.888	27	45	0.031 - 0.031	67%
C. Esôfago	1	I(1,3)	LASSO	10	476.257	28	55	0.001 - 0.001	96%
C. Esôfago	1	F(5,7)	pSVM	9	551.007	29	46	0.015 - 0.015	59%
C. Esôfago	1	F7	pSVM	2	522.051	14	31	0.002 - 0.002	121%
C. Esôfago	1	I1	pSVM	6	323.479	21	39	0.004 - 0.004	86%
C. Esôfago	1	F7	pSVM	1	353.043	8	21	0.013 - 0.013	163%
C. Pulmão	2	H2	pSVM, LASSO	16	574.415	97	152	0.001 - 0.001	57%
C. Pulmão	2	H4	pSVM, LASSO	6	591.722	107	161	0.001 - 0.001	50%
C. Pulmão	2	H5	pSVM, LASSO	20	506.420	100	146	0.002 - 0.007	46%
C. Pulmão	2	D2	pSVM, LASSO	2	498.045	59	99	0.002 - 0.002	68%
C. Pulmão	2	C2	pSVM, LASSO	16	555.057	40	70	0.015 - 0.019	75%
C. Pulmão	1	H6	LASSO	9	524.379	106	147	0.044 - 0.044	39%
C. Pulmão	1	H7	LASSO	2	452.728	86	124	0.047 - 0.047	44%
C. Loc. n Especif.	4	F9	XGB, RF, pSVM, LASSO	16	422.324	24	37	0.013 - 0.037	56%
C. Loc. n Especif.	4	A1	XGB, RF, pSVM, LASSO	9	508.427	19	37	0.001 - 0.001	100%
C. Loc. n Especif.	4	I6	XGB, RF, pSVM, LASSO	1	340.257	16	35	0.001 - 0.001	122%
C. Loc. n Especif.	4	I4	XGB, RF, pSVM, LASSO	3	168.828	7	26	0.001 - 0.001	285%
C. Loc. n Especif.	4	E6	XGB, RF, pSVM, LASSO	1	25.389	1	14	0.001 - 0.001	1300%
C. Loc. n Especif.	4	F10	XGB, RF, pSVM, LASSO	1	108.145	4	15	0.001 - 0.003	275%
C. Loc. n Especif.	2	A6	pSVM, LASSO	1	303.002	7	19	0.001 - 0.029	171%
C. Loc. n Especif.	2	E8	RF, pSVM	1	159.076	7	17	0.023 - 0.04	162%
C. Loc. n Especif.	1	F8	RF	1	381.407	15	27	0.019 - 0.019	80%
C. Loc. n Especif.	1	A7	pSVM	1	39.190	1	8	0.01 - 0.01	700%
C. Loc. n Especif.	1	E7	pSVM	6	524.942	20	35	0.013 - 0.013	75%
C. Estômago	3	B2	XGB, pSVM, LASSO	5	597.394	51	92	0.001 - 0.022	82%
C. Estômago	2	C4	pSVM, LASSO	13	364.233	27	50	0.021 - 0.028	85%
C. Estômago	2	D2	pSVM, LASSO	2	498.045	33	58	0.017 - 0.023	76%
C. Estômago	2	C5	pSVM, LASSO	17	585.956	38	65	0.011 - 0.014	71%
C. Estômago	2	B3	pSVM, LASSO	2	594.948	46	80	0.001 - 0.002	74%
C. Estômago	1	B1	RF	5	577.729	57	90	0.006 - 0.006	58%

5. DISCUSSÃO

Os algoritmos de *machine learning* conseguiram predizer com relativo sucesso a taxa de mortalidade ajustada por idade por câncer nos municípios do Brasil utilizando apenas dados socioeconômicos como variáveis preditoras. O melhor modelo (*eXtreme Gradient Boosting*) obteve, para o agregado de todos os cânceres, um R^2 de 0,65 no conjunto de teste, o que significa que 65% de toda a variabilidade da taxa de mortalidade nos municípios do Brasil pode ser explicada apenas por variáveis socioeconômicas.

Esse nível de precisão é ainda mais interessante por não levar em consideração nenhum dos principais fatores de risco para o câncer estabelecidos pela literatura. O principal deles, tabagismo, não foi considerado no modelo, principalmente pelo fato de que não existem dados de tabagismo no nível municipal no Brasil. Entretanto, dada a associação complexa de variáveis possibilitada por modelos de *machine learning* (Karim et al., 2018), é possível que alguns desses fatores de risco sejam indiretamente associados à predição por correlacionarem-se a variáveis observadas ou a mediadores de outros fatores de risco, uma vez que é conhecida a associação de fatores socioeconômicos com alguns dos os principais fatores de risco de câncer (Akinyemiju et al., 2017).

Essa boa capacidade preditiva possibilita novas análises associativas mais precisas em estudos epidemiológicos ecológicos. Conforme previsto no Teorema de Frisch-Waugh-Lovell (Frisch and Waugh, 1933; Lovell, 2006), uma regressão com variáveis independentes Z, utilizando como variável dependente os resíduos de um modelo com variáveis independentes X (neste caso sendo X as variáveis socioeconômicas, e o resíduo a porção de mortalidade não explicada por X) resultaria nos mesmos coeficientes de uma regressão múltipla considerando todas as variáveis X + Z. Dessa forma, o uso de uma regressão utilizando como variável dependente os resíduos de um modelo de *machine learning*, que por sua vez são capazes de obter melhores resultados em comparação a modelos de regressão tradicionais (aqui demonstrado na Figura 7), resultariam em melhores estimativas de coeficientes do que aplicando esses mesmos métodos tradicionais desde o início para todas as variáveis em regressões múltiplas. Em outras palavras, tomando o resíduo aqui obtido e aplicando-o como variável dependente em outro modelo utilizando Z como variável dependente (ex. concentração de um

contaminante ambiental, prevalência de alguma condição genética, etc) resultaria em um modelo muito melhor do que um modelo multivariado utilizando Z mais variáveis socioeconômicas.

Essa melhoria na acurácia possibilita também uma melhor estimativa de associação geográfica. A epidemiologia espacial permite identificar variações na incidência de doenças em relação a fatores demográficos, ambientais, comportamentais, socioeconômicos e genéticos, dado que todos esses fatores possuem uma forte associação espacial (Elliott and Wartenberg, 2004). Dessa forma, a análise espacial de resíduos de um modelo de *machine learning* utilizando variáveis socioeconômicas possibilita isolar nestas associações apenas os fatores demográficos, ambientais, comportamentais e genéticos em suas associações espaciais.

A análise espacial utilizando o método de Moran Local com base no modelo de XGB nos dados de taxa de mortalidade pelo agregado de todos os cânceres, mostrou associação espacial dispersa em vários municípios espalhados pelo país, entretanto com uma forte concentração nos estados do Rio Grande do Sul e Paraná. Essas mesmas regiões foram também identificadas pelo método de varredura de Kulldorff. A diferença entre essas metodologias é que o método de varredura permite uma estimativa de risco adicional para todo o aglomerado, enquanto o método de Moran permite uma estimativa individual de cada cidade. Dessa forma, optou-se pelo método de Kulldorff para a sequência das análises.

As análises de varredura utilizando o método de Kulldorff mostraram-se sensíveis aos diferentes algoritmos utilizados. Como observado na Tabela 5, para o agregado de todos os cânceres para o método de XGB, foram identificados 3 aglomerados enquanto para o método de LASSO foram identificados 12 aglomerados. Esses aglomerados identificados por diferentes modelos às vezes continham exatamente as mesas cidades (Ex. Figura 19 H1), entretanto em outros casos esses aglomerados estavam sobrepostos em apenas algumas cidades (Ex. Figura 19 E2 e E3, Figura 19 D1 e D2).

A região entre as cidades de Bagé e Rio Grande (Ex. Figura 19 H1) foi identificada por todos os modelos para o agregado de todos os cânceres com um excesso de 27% em relação ao previsto. Este mesmo aglomerado, com

exatamente as mesas cidades, foi também identificado para câncer colorretal com um excesso de 70% em relação ao previsto, o único aglomerado identificado para esse tumor em todo o Brasil.

Toda a região do extremo sul do Brasil, fronteira com o Uruguai, foi foco de uma miríade de 11 aglomerados de 5 tipos diferentes de câncer avaliados. Esses, associados aos outros 3 aglomerados da região de Porto Alegre, fazem do estado do Rio Grande do Sul o estado com o maior número de aglomerados do país. Também na análise específica para câncer de pulmão, onde dos 7 aglomerados identificados no Brasil, 5 estão no Rio Grande do Sul. Essas duas observações são consistentes com dados do Instituto Nacional do Câncer (INCA) em que apontam o estado do Rio Grande do Sul como o estado do Brasil com a maior taxa de incidência de todos os cânceres e específica para câncer de pulmão. Entretanto, áreas de alta incidência não necessariamente identificam taxas anômalas, pois, por não considerarem variáveis socioeconômicas, podem estar dentro do esperado nessas áreas. Este estudo, portanto, permite mostrar que o estado do Rio Grande do Sul não só tem as maiores taxas de mortalidade por câncer do Brasil como também as maiores preditas. Entretanto, o estado com a maior taxa anômala é o estado do Amazonas que, tendo uma taxa ajustada de mortalidade por câncer de 102 mortes por 100.000 habitantes, próxima da média nacional de 98 mortes por 100.000 habitantes, está 21% acima da taxa esperada dadas as suas características sociodemográficas (Figura 12 e Tabela 3).

Também no Rio Grande do Sul, foram identificados 3 aglomerados parcialmente sobrepostos na região Oeste do Estado com altas taxas não explicadas de câncer de estômago (Figura 19 I1, I2 e I3). Esses aglomerados diferem no modelo utilizado e no excesso de mortalidade, variando entre 96% e 42% (Tabela 6Tabela 6 - Detalhes dos aglomerados geográficos da Figura 19.). Sewram e colaboradores identificaram o consumo de mate quente associado a um risco três vezes maior (OR, 2.95; IC95% 1,30–6,74) de desenvolvimento de câncer de esôfago de células escamosas (Sewram et al., 2003). O consumo de mate quente é frequente no Rio Grande do Sul e pode estar associado a esses aglomerados. Entretanto esta mesma associação não pode ser feita para outros aglomerados desse tumor encontrados no Brasil como os identificados no Paraná com um excesso de mortalidade de 82 e 77% (Figura 19 E4 e E5), Norte de Minas

com adicionais de mortalidade entre 50 e 61% (Figura 19 F1, F2 e F3) e Espírito Santo com 4 aglomerados com adicionais entre 58 e 78% (Figura 19 F4, F5, F6 e F7).

Na Região Norte, observaram-se diversos agregados sobrepostos na cidade de Porto Velho e região. Os aglomerados destacados incluíam câncer de pulmão (com um adicional de mortalidade médio de 67.8%), câncer de estômago (com um adicional de mortalidade médio de 75.8%) e todos os cânceres com 2 aglomerados parcialmente sobrepostos (com um adicional de mortalidade 26.6% e 39.5%)(Figura 19 D1 e D2). Esse aglomerado de todos os cânceres é contíguo ao aglomerado que liga a cidade de Porto Velho à cidade de Manaus (Figura 19 A3). A cidade de Manaus não aparece em um aglomerado pois a sua população supera o limite populacional de 600.000 habitantes para a formação de aglomerados, entretanto, como pode ser observado na Figura 13, Manaus tem a maior taxa de mortalidade por câncer do Brasil (considerando-se os municípios com mais de 1 milhão de habitantes). Portanto, toda a região entre Porto Velho, Manaus e o estado de Roraima merecem especial atenção em relação às taxas de mortalidade por câncer, muito acima do esperado para cidades em condições socioeconômicas semelhantes.

A cidade de Macapá e região apresentaram 2 aglomerados parcialmente sobrepostos que incluíam todos os modelos para câncer gástrico com um adicional de mortalidade que chegou a 82% (Figura 19 B1 e B2). Esse achado é, também, semelhante aos dados de incidência do INCA que o identificou como o estado com a maior incidência de câncer gástrico do país (Ministério da Saúde do Brasil, 2019). Entretanto, como ressaltamos anteriormente, a estimativa do INCA não faz estimação de valor esperado segundo as características locais.

No estado do Ceará foram observados 7 diferentes aglomerados para os cânceres de pulmão, estômago e o consolidado de todos os cânceres (Figura 19 C). O aglomerado entre as cidades de Viçosa do Ceará e Guaraciaba do Norte, identificado pelos modelos de LASSO e pSVM, mostrou um adicional de 85,2% de mortalidade por câncer de estômago além do predito (Figura 19 C4). O segundo caso com maiores excessos foi o aglomerado identificado entre as cidades de Sobral e Canindé, com excesso de mortalidade Câncer de Pulmão de 75% além do predito pelos modelos de LASSO e pSVM.

Os tumores de localização não-especificada foram os que tiveram os excessos mais claros em relação ao predito com valores chegando a 1300% na cidade de Itaperuçu, situada na região próxima a Curitiba-PR. Oliveira e colaboradores demonstraram que 67% dos casos de tumores de localização mal definidas de Goiânia poderiam ter suas localizações corrigidas quando comparadas àquelas registradas no registro de câncer de base populacional (Oliveira et al., 2014). Dessa forma, pode-se inferir que nesses aglomerados de grande concentração de tumores de localização não-especificadas pode haver um grave problema de classificação da localização primária de tumores, podendo servir como um guia para identificação de locais para investimento em ferramentas de apoio à identificação de tumores.

É importante notar que nenhum aglomerado significativo de mortalidade por câncer não explicada por fatores socioeconômicos foi encontrado para os cânceres de mama, próstata, pâncreas, fígado e cérebro. Alguns destes tumores possuem baixa letalidade e, por isso, estão mais sujeitos a fontes de variabilidade não observadas pela variação da mortalidade analisada neste estudo, o que diminui a capacidade preditiva dos modelos de *machine learning* avaliados. Outra possibilidade é que esses tumores sejam pouco afetados por outras variáveis que não sociodemográficas e, portanto, não estejam associados geograficamente.

A análise inferencial possibilitada pelo método SHAP identificou a proporção de brancos como a principal variável para a predição da taxa de mortalidade por câncer nos municípios do Brasil. Outras variáveis importantes foram o percentual de casas com computador, geladeira, automóvel e percentual de analfabetos. Nota-se, portanto, uma participação significativa de elementos de renda relacionados a uma alta predição da mortalidade por câncer. A relação de fatores socioeconômicos com a incidência e mortalidade por câncer são conhecidos e são baseados na capacidade estendida de locais mais desenvolvidos em tratar fatores de risco concorrentes como diabetes e doenças cardiovasculares, além de um importante efeito relacionado à dieta e ao estilo de vida (Barrett et al., 1998; Gersten and Wilmoth, 2002; Omran, 1971). Uma associação semelhante foi observada por Borges e colaboradores nas capitais do Brasil para câncer de boca (Borges et al., 2009) na qual foi observada uma associação forte e positiva da incidência deste tumor com o IDH municipal, longevidade e renda.

Discussão

Houve também uma contribuição negativa importante da variável percentual de óbitos não-classificados, um marcador da qualidade do dado de classificação de mortes em cada município (Figura 10). Dessa forma, os municípios com pior classificação de mortes tiveram também menos mortes classificadas por câncer; nesse caso, o modelo ajusta sua predição baseada nessa variável diminuindo o impacto da qualidade dos dados no resultado da previsão.

Entre as principais limitações deste estudo é importante mencionar que não é capaz de gerar qualquer explicação etiológica para a existência dos aglomerados de alta mortalidade não explicada por fatores socioeconômicos, apenas identifica-os dentro de regiões específicas. Os resultados do estudo, porém, têm um potencial gerador de hipóteses e podem servir como guias para a condução de estudos epidemiológicos de campo com foco em fatores de risco ambientais, genéticos, comportamentais e socioeconômicos, além de permitir identificar áreas de foco para políticas públicas de redução de risco de câncer. Em segundo lugar, o estudo analisou a mortalidade ao invés de incidência uma vez que dados de incidência confiáveis somente podem ser obtidos no nível de granularidade necessário para o estudo em um Registro de Câncer de Base Populacional de cobertura nacional, o que não existe no Brasil. Ainda que a mortalidade possa ser um bom proxy para a incidência, ela pode incluir novas fontes de variabilidade e erro, especialmente no caso de cânceres mais tratáveis (como câncer de mama e câncer de próstata).

6. CONCLUSÃO

O uso de técnicas de aprendizado supervisionado com *machine learning* se mostrou-se eficiente para a predição da taxa de mortalidade em nível ecológico e uma estratégia versátil, especialmente para o caso da análise de resíduos. Diversos aglomerados de excesso de mortalidade de câncer foram identificados para os tumores de pulmão, estômago, colorretal, tumores de localização não especificada e para todos os cânceres agrupados. O método de interpretabilidade do modelo mostrou a relação entre mortalidade por câncer e variáveis socioeconômicas e suas características. Esses achados podem gerar novas

hipóteses para pesquisa epidemiológica, bem como servir de guia para ações futuras específicas de prevenção, rastreamento e tratamento.

7. REFERÊNCIAS

Agencia Nacional de Saúde Suplementar - ANS. **ANS TABNET** 2019. http://www.ans.gov.br/anstabnet/ (acessado em 02/17/2019).

Akinyemiju T, Ogunsina K, Okwali M, Sakhuja S, Braithwaite D. **Lifecourse socioeconomic status and cancer-related risk factors: Analysis of the WHO study on global ageing and adult health (SAGE)**. Int J Cancer 2017;140:777–87. https://doi.org/10.1002/ijc.30499.

Albuquerque MV de, Viana AL d'Ávila, Lima LD de, Ferreira MP, Fusaro ER, Iozzi FL. **Desigualdades regionais na saúde: mudanças observadas no Brasil de 2000 a 2016**. Cien Saude Colet 2017;22:1055–64. https://doi.org/10.1590/1413-81232017224.26862016.

Anand S, Sen A. **Human Development Index: Methodology and Measurement.** Oxford: 1994.

Angwin J, Larson J, Mattu S, Kirchner L. **Machine Bias**. ProPublica 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (acessado em 03/15/2020).

ANS AN de SSu. **Beneficiários de planos privados de saúde, por cobertura assistencial (Brasil – 2009-2019)** 2019.

Anselin L. **Local Indicators of Spatial Association-LISA**. Geogr Anal 2010;27:93–115. https://doi.org/10.1111/j.1538-4632.1995.tb00338.x.

Azzoni CR. **Economic growth and regional income inequality in Brazil**. Ann Reg Sci 2001;35:133–52. https://doi.org/10.1007/s001680000038.

Baine WB, Yu W, Summe JP. **The Epidemiology of Hospitalization of Elderly Americans for Septicemia or Bacteremia in 1991–1998: Application of Medicare Claims Data**. Ann Epidemiol 2001;11:118–26. https://doi.org/10.1016/S1047-2797(00)00184-8.

Barbosa IR, Costa Í do CC, de Souza DLB, Pérez MB. **Desigualdades Socioespaciais na distribuição da mortalidade por câncer no Brasil**. Geogr Médica e Da Saúde 2016;12:122–32.

Barnston AG. **Correspondence among the Correlation, RMSE, and Heidke**

Forecast Verification Measures; Refinement of the Heidke Score. Weather Forecast 1992;7:699–709. https://doi.org/10.1175/1520-0434(1992)007<0699:CATCRA>2.0.CO;2.

Barrett R, Kuzawa CW, McDade T, Armelagos GJ. **Emerging and re-emerging infectious diseases: the third epidemiologic transition**. Emerg Re-Emerging Infect Dis Third Epidemiol Transit 1998;27:247–71. https://doi.org/10.1146/annurev.anthro.27.1.247.

Black RJ, Bray F, Ferlay J, Parkin DM. **Cancer incidence and mortality in the European union: Cancer registry data and estimates of national incidence for 1990**. Eur J Cancer Part A 1997;33:1075–107. https://doi.org/10.1016/S0959-8049(96)00492-3.

Borges DM de L, Sena MF de, Ferreira MAF, Roncalli AG. **Mortalidade por câncer de boca e condição sócio-econômica no Brasil**. Cad Saude Publica 2009;25:321–7. https://doi.org/10.1590/S0102-311X2009000200010.

Brasil M. **Municípios do Brasil - Grupos étnico-raciais predominantes** 2019. https://commons.wikimedia.org/wiki/File:Municípios_do_Brasil_-_Grupos_étnico-raciais_predominantes.png (acessado em 03/15/2020).

Breiman L. **Statistical Modeling: The Two Cultures**. Stat Sci 2001;16:199–231.

Chiavegatto Filho ADP, Kawachi I, Wang YP, Viana MC, Andrade LHSG. **Does income inequality get under the skin? A multilevel analysis of depression, anxiety and mental disorders in São Paulo, Brazil**. J Epidemiol Community Health 2013;67:966–72. https://doi.org/10.1136/jech-2013-202626.

Chiavegatto Filho ADP, dos Santos HG, do Nascimento CF, Massa K, Kawachi I. **Overachieving Municipalities in Public Health**. Epidemiology 2018;29:836–40. https://doi.org/10.1097/EDE.0000000000000919.

Cirkovic BRA, Cvetkovic AM, Ninkovic SM, Filipovic ND. **Prediction models for estimation of survival rate and relapse for breast cancer patients**. 2015 IEEE 15th Int. Conf. Bioinforma. Bioeng., IEEE; 2015, p. 1–6. https://doi.org/10.1109/BIBE.2015.7367658.

Cogliano VJ, Baan R, Straif K, Grosse Y, Lauby-Secretan B, Ghissassi F El, et al.

Preventable exposures associated with human cancers. J Natl Cancer Inst 2010;103:1827–39. https://doi.org/10.1093/jnci/djr483.

Correia LO dos S, Padilha BM, Vasconcelos SML. **Métodos para avaliar a completitude dos dados dos sistemas de informação em saúde do Brasil: uma revisão sistemática**. Cien Saude Colet 2014;19:4467–78. https://doi.org/10.1590/1413-812320141911.02822013.

Cosmatos I, Matcho A, Weinstein R, Montgomery MO, Stang P. **Analysis of patient claims data to determine the prevalence of hidradenitis suppurativa in the United States**. J Am Acad Dermatol 2013;68:412–9. https://doi.org/10.1016/J.JAAD.2012.07.027.

Departamento de Análise de Situação e Saúde. **VIGILÂNCIA DE FATORES DE RISCO E PROTEÇÃO PARA DOENÇAS CRÔNICAS POR INQUÉRITO TELEFÔNICO - VIGITEL**. 1st ed. Brasilia: MINISTÉRIO DA SAÚDE; 2007.

Dominguez RL, Cherry CB, Estevez-Ordonez D, Mera R, Escamilla V, Pawlita M, et al. **Geospatial analyses identify regional hot spots of diffuse gastric cancer in rural Central America**. BMC Cancer 2019;19:1–8. https://doi.org/10.1186/s12885-019-5726-x.

Elfiky AA, Pany MJ, Parikh RB, Obermeyer Z. **Development and Application of a Machine Learning Approach to Assess Short-term Mortality Risk Among Patients With Cancer Starting Chemotherapy**. JAMA Netw Open 2018;1:e180926. https://doi.org/10.1001/jamanetworkopen.2018.0926.

Elliott P, Wartenberg D. **Spatial epidemiology: Current approaches and future challenges**. Environ Health Perspect 2004;112:998–1006. https://doi.org/10.1289/ehp.6735.

Faggiano F, Partanen T, Kogevinas M, Boffetta P. **Socioeconomic differences in cancer incidence and mortality.** IARC Sci Publ 1997:65–176.

Fawcett T. **An introduction to ROC analysis**. Pattern Recognit Lett 2006;27:861–74. https://doi.org/10.1016/J.PATREC.2005.10.010.

Ferlay J, Lam F, Colombet M, Mery L, Pineros M, Znaor A. **Global cancer observatory: cancer today**. Int Agency Res Cancer 2020:Available from

https://gco.iarc.fr/today.

Frisch R, Waugh F V. **Partial Time Regressions as Compared with Individual Trends**. Econometrica 1933;1:387. https://doi.org/10.2307/1907330.

FUCCIO L, ZAGARI RM, MINARDI ME, BAZZOLI F. **Systematic review: Helicobacter pylori eradication for the prevention of gastric cancer**. Aliment Pharmacol Ther 2006;25:133–41. https://doi.org/10.1111/j.1365-2036.2006.03183.x.

Fundação Nacional de Saúde (FUNASA). **Manual de Procedimentos do Sistema de Informações sobre Mortalidade Manual de Procedimentos do Sistema de Informações sobre Mortalidade**. 2001.

Gapminder Tools. 2020. https://www.gapminder.org/tools/#$state$time$value=2015;&marker$axis_x$which=hdi_human_development_index&domainMin:null&domainMax:null&zoomedMin:null&zoomedMax:null&scaleType=linear&spaceRef:null;&axis_y$which=colonandrectum_cancer_new_cases_per_100000_wo (acessado em 03/22/2020).

Gersten O, Wilmoth JR. **The Cancer Transition in Japan since 1951**. Source Demogr Res 2002;7:271–306. https://doi.org/10.4054/DemRes.2002.7.5.

Health Knowledge. **Health information: important regional and international differences in populations | Health Knowledge** n.d. https://www.healthknowledge.org.uk/e-learning/health-information/population-health-specialists/regional-international-differences-populations (acessado em 04/01/2020).

Hosseinpoor AR, Parker LA, Tursan d'Espaignet E, Chatterji S. **Social determinants of smoking in low- and middle-income countries: Results from the world health survey**. PLoS One 2011;6. https://doi.org/10.1371/journal.pone.0020331.

Hurley SF, Matthews JP. **Cost-effectiveness of the Australian National Tobacco Campaign**. Tob Control 2008;17:379–84. https://doi.org/10.1136/tc.2008.025213.

IARC. **Latest global cancer data: Cancer burden rises to 18.1 million new cases and 9.6 million cancer deaths in 2018**. 2018.

IBGE. **Projeção da população do Brasil e das Unidades da Federação**. Inst Bras Geogr e Estatística 2019.

IBGE. **IBGE | mapas | político administrativo** 2018. https://mapas.ibge.gov.br/politico-administrativo (acessado em 02/10/2019).

IBGE. **Base de Informações por Setor Censitário** 2012. ftp://ftp.ibge.gov.br/Censos/Censo_Demografico_2010/Resultados_do_Universo/Agregados_por_Setores_Censitarios/ (acessado em 09/17/2017).

IBGE. **Censo Demográfico 2010** 2010. https://www.ibge.gov.br/home/estatistica/populacao/censo2010/default.shtm (acessado em 09/17/2017).

Instituto Brasileiro de Geografia e Estatística (IBGE). **Áreas Territoriais**. Brasília: 2020.

International Monetary Fund (IMF). **World Economic Outlook Database**. 2020.

Jerez JM, Molina I, García-Laencina PJ, Alba E, Ribelles N, Martín M, et al. **Missing data imputation using statistical and machine learning methods in a real breast cancer problem**. Artif Intell Med 2010;50:105–15. https://doi.org/10.1016/J.ARTMED.2010.05.002.

Karagiannis-Voules DA, Scholte RGC, Guimarães LH, Utzinger J, Vounatsou P. **Bayesian Geostatistical Modeling of Leishmaniasis Incidence in Brazil**. PLoS Negl Trop Dis 2013;7. https://doi.org/10.1371/journal.pntd.0002213.

Karim ME, Pang M, Platt RW. **Can We Train Machine Learning Methods to Outperform the High-dimensional Propensity Score Algorithm?**. Epidemiology 2018;29:191–8. https://doi.org/10.1097/EDE.0000000000000787.

Kim S, Park T, Kon M. **Cancer survival classification using integrated data sets and intermediate information**. Artif Intell Med 2014;62:23–31. https://doi.org/10.1016/j.artmed.2014.06.003.

Kourou K, Exarchos TP, Exarchos KP, Karamouzis M V., Fotiadis DI. **Machine learning applications in cancer prognosis and prediction**. Comput Struct Biotechnol J 2015;13:8–17. https://doi.org/10.1016/J.CSBJ.2014.11.005.

Kuhn M, Weston S, Williams A, Keefer C, Engelhardt A, Cooper T, et al. **caret: Classification and Regression Training** 2018.

Kulldorff M. **A spatial scan statistic**. Commun Stat - Theory Methods 1997;26:1481–96. https://doi.org/10.1080/03610929708831995.

Kulldorff M, Feuer EJ, Miller BA, Freedma LS. **Breast Cancer Clusters in the Northeast United States: A Geographic Analysis**. Am J Epidemiol 1997;146:161–70. https://doi.org/10.1093/oxfordjournals.aje.a009247.

Kulldorff M, Nagarwalla N. **Spatial disease clusters: Detection and inference**. Stat Med 1995;14:799–810. https://doi.org/10.1002/sim.4780140809.

Lansdorp-Vogelaar I, Sharp L. **Cost-effectiveness of screening and treating Helicobacter pylori for gastric cancer prevention**. Best Pract Res Clin Gastroenterol 2013;27:933–47. https://doi.org/10.1016/j.bpg.2013.09.005.

Lima CR de A, Leal CD, Dias EP, Gonzalez FL, dos Santos HL, da Silva MEM, et al. **Departamento de Informática do SUS – DATASUS A Experiência de Disseminação de Informações em Saúde**. A experiência Bras. em Sist. informação em saúde, MINISTÉRIO DA SAÚDE; 2009, p. 109–28.

Lin H, Ning B, Li J, Ho SC, Huss A, Vermeulen R, et al. **Lung cancer mortality among women in Xuan Wei, China: A comparison of spatial clustering detection methods**. Asia-Pacific J Public Heal 2015;27:NP392–401. https://doi.org/10.1177/1010539512444778.

Loda M, Mucci LA, Mittelstadt ML, Van Hemelrijck M, Cotter MB. **Pathology and epidemiology of cancer**. Springer International Publishing; 2016. https://doi.org/10.1007/978-3-319-35153-7.

Lovell MC. **A Simple Proof of the FWL (Frisch-Waugh-Lovell) Theorem**. SSRN Electron J 2006. https://doi.org/10.2139/ssrn.887345.

Lundberg SM, Erion G, Chen H, Degrave A, Prutkin JM, Nair B, et al. **Explainable AI for Trees: From Local Explanations to Global Understanding**. n.d.

Lundberg SM, Erion GG, Lee S-I. **Consistent Individualized Feature Attribution for Tree Ensembles** 2018.

Lundberg SM, Lee S-I. **A unified approach to interpreting model predictions**. 31 st Conf. Neural Inf. Process. Syst., vol. 2017- Decem, Neural information processing systems foundation; 2017, p. 4766–75.

Lundqvist A, Andersson E, Ahlberg I, Nilbert M, Gerdtham U. **Socioeconomic inequalities in breast cancer incidence and mortality in Europe - A systematic review and meta-analysis**. Eur J Public Health 2016;26:804–13. https://doi.org/10.1093/eurpub/ckw070.

Ma Y, Ding Z, Qian Y, Shi X, Castranova V, Harner EJ, et al. **Predicting Cancer Drug Response by Proteomic Profiling**. Clin Cancer Res 2006;12:4583–9. https://doi.org/10.1158/1078-0432.CCR-06-0290.

Machado JP, Martins M, Leite I da C, Machado JP, Martins M, Leite I da C. **Qualidade das bases de dados hospitalares no Brasil: alguns elementos**. Rev Bras Epidemiol 2016;19:567–81. https://doi.org/10.1590/1980-5497201600030008.

Milyo J, Mellor JM. **On the Importance of Age-Adjustment Methods in Ecological Studies of Social Determinants of Mortality**. Health Serv Res 2003;38:1781–90. https://doi.org/10.1111/j.1475-6773.2003.00202.x.

Ministério da Saúde do Brasil. **Estimativa INCA 2020: Incidência de Câncer no Brasil**. 2019.

Mucaki EJ, Zhao JZL, Lizotte DJ, Rogan PK. **Predicting responses to platin chemotherapy agents with biochemically-inspired machine learning**. Signal Transduct Target Ther 2019;4:1. https://doi.org/10.1038/s41392-018-0034-5.

Mukherjee S. **O Imperador de Todos os Males**. 1st ed. São Paulo: Companhia das Letras; 2012.

National Cancer Institute (NCI). **World (WHO 2000-2025) Standard - Standard Populations - SEER Datasets** 2013. https://seer.cancer.gov/stdpopulations/world.who.html (acessado em 02/10/2019).

Oliveira PPV de, Silva GA e, Curado MP, Malta DC, Moura L de. **Confiabilidade da causa basica de obito por cancer entre Sistema de Informacoes sobre Mortalidade do Brasil e Registro de Cancer de Base Populacional de Goiania, Goias, Brasil**. Cad Saude Publica 2014;30:296–304. https://doi.org/10.1590/0102-

311X00024813.

Omran AR. **The epidemiologic transition: A theory of the epidemiology of population change**. Milbank Mem Fund Q 1971;49:509–38. https://doi.org/10.1111/j.1468-0009.2005.00398.x.

Park K, Ali A, Kim D, An Y, Kim M, Shin H. **Robust predictive model for evaluating breast cancer survivability**. Eng Appl Artif Intell 2013;26:2194–205. https://doi.org/10.1016/J.ENGAPPAI.2013.06.013.

Presidência da República. **Lei de Acesso à Informação**. LEI Nº 12527, 18 NOVEMBRO 2011 2011. http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.htm (acessado em 03/15/2020).

Queiroz BL, Freire FHM de A, Gonzaga MR, Lima EEC de, Queiroz BL, Freire FHM de A, et al. **Estimativas do grau de cobertura e da mortalidade adulta (45q15) para as unidades da federação no Brasil entre 1980 e 2010**. Rev Bras Epidemiol 2017;20:21–33. https://doi.org/10.1590/1980-5497201700050003.

R Core Team. **R: A Language and Environment for Statistical Computing** 2018.

Rigotti JIR. **A (re)distribuição espacial da população brasileira e possíveis impactos sobre a metropolização O USO DOS QUESITOS CENSITÁRIOS PARA O ESTUDO DAS MIGRAÇÕES View project The effect of varying population estimates on the calculation of enrolment rates and ou**. 32º Encontro Anu. da Anpocs, Caxambu: 2008, p. 1.

Rios-Neto ELG, Guimarães RR de M. **The demography of education in Brazil: inequality of educational opportunities based on Grade Progression Probability (1986-2008)**. Vienna Yearb Popul Res 2010;8:283–306. https://doi.org/10.2307/23025518.

Romero Canal M, da Silva Ferreira ER, Estofolete CF, Martiniano Dias A, Tukasan C, Bertoque AC, et al. **Spatiotemporal-based clusters as a method for dengue surveillance**. Rev Panam Salud Pública 2017;41:1–6. https://doi.org/10.26633/rpsp.2017.162.

Samuel AL. **Some Studies in Machine Learning Using the Game of Checkers**. IBM J Res Dev 1959;3:210–29. https://doi.org/10.1147/rd.33.0210.

Saran R, Robinson B, Abbott KC, Agodoa LYC, Albertus P, Ayanian J, et al. **US Renal Data System 2016 Annual Data Report: Epidemiology of Kidney Disease in the United States.** Am J Kidney Dis 2017;69:A7–8. https://doi.org/10.1053/j.ajkd.2016.12.004.

Sewram V, De Stefani E, Brennan P, Boffetta P. **Maté consumption and the risk of squamous cell esophageal cancer in Uruguay**. Cancer Epidemiol Biomarkers Prev 2003;12:508–13.

Shapley LS. **Notes on the n-Person Game -- II: The Value of an n-Person Game**. Santa Monica: 1951.

Sherman RL, Henry KA, Tannenbaum SL, Feaster DJ, Kobetz E, Lee DJ. **Applying spatial analysis tools in public health: An example using satscan to detect geographic targets for colorectal cancer screening interventions**. Prev Chronic Dis 2014;11. https://doi.org/10.5888/pcd11.130264.

Singal AG, Mukherjee A, Joseph Elmunzer B, Higgins PDR, Lok AS, Zhu J, et al. **Machine learning algorithms outperform conventional regression models in predicting development of hepatocellular carcinoma**. Am J Gastroenterol 2013;108:1723–30. https://doi.org/10.1038/ajg.2013.332.

Sondik E. **Measurement of progress against cancer. Extramural Committee to Assess Measures of Progress Against Cancer.** J Natl Cancer Inst 1990;82:825–35.

Taylor RA, Pare JR, Venkatesh AK, Mowafi H, Melnick ER, Fleischman W, et al. **Prediction of In-hospital Mortality in Emergency Department Patients with Sepsis: A Local Big Data-Driven, Machine Learning Approach**. Acad Emerg Med 2016;23:269–78. https://doi.org/10.1111/acem.12876.

Thorsen-Meyer H-C, Nielsen AB, Nielsen AP, Kaas-Hansen BS, Toft P, Schierbeck J, et al. **Dynamic and explainable machine learning prediction of mortality in patients in the intensive care unit: a retrospective study of high-frequency data in electronic patient records**. Lancet Digit Heal 2020;2:e179–91. https://doi.org/10.1016/S2589-7500(20)30018-2.

United Nations Development Program (UNDP). **Human Development Reports**. 2019.

Vieira NN, Cançado ALF. **IDENTIFICAÇÃO DE CONGLOMERADOS ESPACIAIS DE ACIDENTES AÉREOS NO BRASIL**. Rev Conex SIPAER 2013;4:64–76.

Wang Y, Tetko I V., Hall MA, Frank E, Facius A, Mayer KFX, et al. **Gene selection from microarray data for cancer classification—a machine learning approach**. Comput Biol Chem 2005;29:37–46. https://doi.org/10.1016/J.COMPBIOLCHEM.2004.11.001.

Wild CP. **Complementing the genome with an "exposome": The outstanding challenge of environmental exposure measurement in molecular epidemiology**. Cancer Epidemiol Biomarkers Prev 2005;14:1847–50. https://doi.org/10.1158/1055-9965.EPI-05-0456.

World Bank. **Gini index (World Bank estimate) - Brazil** 2018a. https://data.worldbank.org/indicator/SI.POV.GINI?locations=BR (acessado em 10/25/2020).

World Bank. **Poverty Data - Brazil** 2018b. https://data.worldbank.org/indicator/SI.POV.UMIC?locations=BR (acessado em 10/25/2020).

World Health Organization. **Mortality Database**. Cancer Mortal Database 2019. https://www-dep.iarc.fr/WHOdb/WHOdb.htm (acessado em 09/04/2020).

Zhang S. **Parimputation : From Imputation and Null-Imputation to Partially Imputation**. IEEE Intell Informatics Bull 2008;9:32–8.

8. CURRÍCULO LATTES

Bruno Casaes Teixeira

Endereço para acessar este CV: http://lattes.cnpq.br/6105897959048452
ID Lattes: **6105897959048452**
Última atualização do currículo em 08/02/2021

Possui graduação em Farmácia e Bioquímica pela Universidade de São Paulo (2008). É mestrando no Programa de Saúde Pública da faculdade de Saúde Pública da USP com foco em estudos epidemiológicos em oncologia utilizando Machine Learning e Estatística Espacial. **(Texto informado pelo autor)**

Identificação

Nome	Bruno Casaes Teixeira
Nome em citações bibliográficas	TEIXEIRA, B. C.
Lattes iD	http://lattes.cnpq.br/6105897959048452

Endereço

Formação acadêmica/titulação

2002 - 2008	Graduação em Farmácia e Bioquímica. Universidade de São Paulo, USP, Brasil. Orientador: -.

Formação Complementar

2020 - 2020	Survival Analysis. (Carga horária: 40h). Johns Hopkins University, JHU, Estados Unidos.
2020 - 2020	Critical reading of epidemiologic literature. (Carga horária: 8h). Johns Hopkins University, JHU, Estados Unidos.
2020 - 2020	Methods and applications of Cohort Studies. (Carga horária: 40h). Johns Hopkins University, JHU, Estados Unidos.
2018 - 2018	Statistics. (Carga horária: 40h). Duke University, DUKE, Estados Unidos.
2017 - 2017	Machine Learning. (Carga horária: 2017h). Stanford University, STANFORD, Estados Unidos.
2016 - 2016	Change Management. (Carga horária: 40h). Harvard University, HARVARD, Estados Unidos.
2014 - 2014	Data Analysis. (Carga horária: 40h). Johns Hopkins University, JHU, Estados Unidos.

Atuação Profissional

Amgen Biotecnologia Ltda, AMG, Brasil.

Vínculo institucional

2015 - 2019	Vínculo: Celetista, Enquadramento Funcional: Real World Evidence Manager

Alexandre Dias Porto Chiavegatto Filho

Bolsista de Produtividade em Pesquisa do CNPq - Nível 2

Endereço para acessar este CV: http://lattes.cnpq.br/5517850224634709
ID Lattes: **5517850224634709**
Última atualização do currículo em 01/02/2021

Possui graduação em Economia pela FEA/USP, doutorado direto em Saúde Pública pela FSP/USP e pós-doutorado na Universidade de Harvard. É Professor Livre Docente do Departamento de Epidemiologia da FSP/USP e orientador dos programas de pós-graduação de Saúde Pública e de Epidemiologia da USP. Atuou como professor convidado (2016 e 2020) e pesquisador visitante (2017 e 2019) na Universidade de Harvard. Em 2020, recebeu o Prêmio Abril e Dasa de Inovação Médica (Categoria: Prevenção). Nos últimos anos, tem sido o Pesquisador Principal de projetos de inteligência artificial em saúde financiados pela FAPESP, CNPq, Microsoft e Fundação Lemann. Em 2015-2016 foi responsável pelo curso online Big Data em Saúde no Brasil, da parceria USP-Coursera, que teve mais de 8.500 alunos matriculados e representantes de todos os Estados brasileiros. É o diretor do Laboratório de Big Data e Análise Preditiva em Saúde (Labdaps) da FSP/USP. Atualmente é o coordenador da rede IACOV-BR (Inteligência Artificial para Covid-19 no Brasil), que tem como objetivo desenvolver algoritmos de machine learning para o diagnóstico e prognóstico de covid-19 nas cinco regiões brasileiras. Tem experiência em pesquisas na área de saúde pública, com ênfase em estatísticas de saúde e machine learning. **(Texto informado pelo autor)**

Identificação

Nome Alexandre Dias Porto Chiavegatto Filho

Nome em citações bibliográficas Chiavegatto Filho, A.D.P.;CHIAVEGATTO FILHO, A. D. P.;Chiavegatto Filho, Alexandre Dias Porto;Filho, Alexandre Dias Porto Chiavegatto;Chiavegatto Filho, Alexandre DP;Chiavegatto, Alexandre Dias Porto;Chiavegatto Filho, Alexandre;Filho, Alexandre Chiavegatto;CHIAVEGATTO FILHO, ALEXANDRE D P;CHIAVEGATTO FILHO, A D P;CHIAVEGATTO FILHO, ALEXANDRE D. P.;DIAS PORTO CHIAVEGATTO FILHO, ALEXANDRE;Chiavegatto Filho, A. D. P.

Lattes iD http://lattes.cnpq.br/5517850224634709

Endereço

Endereço Profissional Faculdade de Saúde Pública.
Av. Dr. Arnaldo, 715
Cerqueira César
01246904 - São Paulo, SP - Brasil
Telefone: (11) 30617914
URL da Homepage: http://www.fsp.usp.br/alexandre

Formação acadêmica/titulação

2007 - 2010 Doutorado em Saúde Pública (Conceito CAPES 6).
Universidade de São Paulo, USP, Brasil.
com **período sanduíche** em Harvard School of Public Health (Orientador: Ichiro Kawachi).
Título: Efeito da desigualdade de renda na mortalidade do Município de São Paulo., Ano de obtenção: 2010.
Orientador: Sabina Léa Davidson Gotlieb.
Bolsista do(a): Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, CAPES, Brasil.
Palavras-chave: desigualdade de renda; epidemiologia; mortalidade; São Paulo.
Grande área: Ciências da Saúde
Grande Área: Ciências da Saúde / Área: Saúde Coletiva / Subárea: Saúde Pública.

2008 - 2014 Graduação em Economia.
Universidade de São Paulo, USP, Brasil.

9. ANEXO: ARTIGO SUBMETIDO

SPATIAL CLUSTERS OF CANCER MORTALITY IN BRAZIL: A MACHINE LEARNING MODELLING APPROACH

Teixeira, BC [1]; Toporcov, TN[1]; Chiaravalloti-Neto F[1]; Chiavegatto Filho, ADP[1]

[1]Department of Epidemiology, School of Public Health, University of São Paulo, Brazil

ABSTRACT

Objectives: This study aimed to test if machine learning algorithms are able to predict the incidence of cancer mortality, and then to use the results to identify cancer clusters not explained by local characteristics. **Methodology:** The outcome of interest was age-standardized mortality data, extracted from the Mortality Information System. Predictive features included publicly-available sociodemographic and health coverage variables. Machine learning algorithms were selected and trained with 70% of the data to predict the age-adjusted cancer mortality rate at the municipal level, while performance was tested with the remaining data (30%). Spatial clusters of municipalities with higher-than-expected mortality rates were identified using Kulldorff statistics. Specific analyzes were also performed for the ten most frequent cancers types. **Results:** The algorithm with highest R^2 was the gradient boosting trees (R^2=0.66). For total cancer, all algorithms found a spatial cluster between Bagé and Rio Grande (27% excess mortality); while three algorithms identified clusters in Porto Velho City (27%-40% excess). For esophageal cancer, clusters of all algorithms overlapped in the west of Rio Grande do Sul (48%-96% excess); while other significant clusters were found in southern Paraná, northern Minas Gerais and Espírito Santo. For stomach cancer, the two most significant clusters in Macapá (82% excess) and Porto Velho (85% excess). Variables with most impact on mortality predictions were percentage of white population proportion and proportion of houses with computers. **Conclusion:** We found a few consistent and well-defined Brazilian geographic regions with significantly higher-than-expected cancer mortality, even after taking into consideration local characteristics.

INTRODUCTION

Cancer incidence varies greatly across geographies and types. At a global scale it evident when analyzing age-adjusted incidence rates, that in 2018 was 419 per 100,000 residents in Oceania, 350 in North America, 217 in Latin America and the Caribbean and 130 in Africa[1]. In Brazil, cancer is the second main cause of death, corresponding to 227,920 deaths in 2018. In relative terms, cancer has a age adjusted mortality rate of 111 per 100,000 for males and 95 per 100,000 for females[2].

These rates also vary considerably within regions of Brazil, being higher in the southern and southeastern regions, which are the most developed of the country[3]. Brazil, with its extensive territory and massive socioeconomic disparity, is also populated by diverse ethnical groups, but has a relatively homogeneous nation-wide healthcare system, which can be a portentous environment for eco-epidemiologic modeling.

Machine learning methods have been applied to various health-related areas, mostly for individualized prediction algorithms such as mortality risk during chemotherapy[4,5], in-hospital mortality[6,7] and prognostic prediction[8–10]. Currently, there are only limited applications in ecological epidemiology modelling[11].

Spatial epidemiology techniques can help to identify the variations of disease incidence in relation to demographic, environmental, behavioral, socioeconomic and genetic risk factors[12]. Spatial clustering analysis, specifically the scan statistic method proposed by Kulldorff[13], have been widely used to identify areas of high rates of health-related outcomes. In oncology, the Kulldorff scan statistic method has been applied to lung cancer in China[14], colorectal cancer in Florida[15], and breast cancer in the United States[16]. The use of scan statistics with machine learning models remains underexplored and can help to identify clusters with a higher incidence than expected given local characteristics.

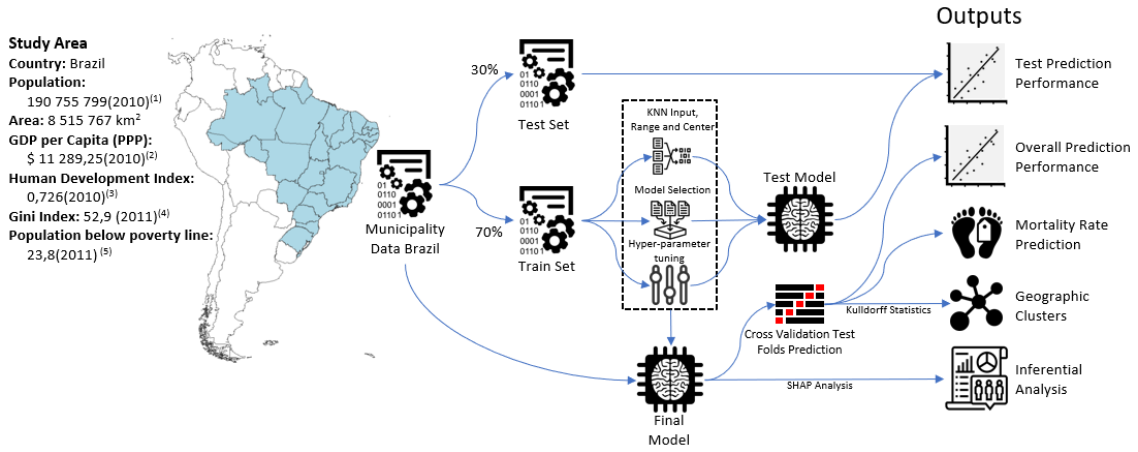
The objective of this study was to first test if machine learning algorithms are able to predict the incidence of cancer mortality in the municipalities of Brazil, and then to use these results to identify cancer clusters that are not explained by local socioeconomic characteristics.

METHODS

DATA COLLECTION

The schematic representation of the study is shown in Figure 1. We first extracted crude mortality data from each of the 5,565 municipalities of Brazil using the Mortality Information System (SIM) of the Ministry of Health, which has been shown to have high coverage, capturing over 95% of deaths in the Brazilian territory[17]. Cancer mortality was selected using the ICD-10 codification for malignant tumors (Chapter 2 of the ICD-10) and aggregated to the Municipality level from 2007 to 2016. Age-adjusted mortality rates for each municipality were calculated using the World Health Organization Standard population from 2000 to 2025[18]. Missing data for covariates were treated using K-Nearest Neighbors imputation method and adjusted to the range between 0-1 using the caret package[19] from R software[20].

Figure 1 – Study Area and Schematic diagram of methods



(1) 2010 Census[21], (2) Brazilian Territory 2020[22], (3) Gross Domestic Product (GDP) by Purchasing Power Parity (PPP) and 2017 International Dollars[23], (4) Human Development Index (HDI)[24], (5) Gini Index (World Bank Estimate)[25], (6) Poverty headcount ratio at $5.50 a day (2011 PPP) (% of population)[26]

A total of 40 sociodemographic variables focusing on income, assets, demography and urbanization were collected from the last Census (from 2010) with municipalities as the aggregated level[21]. The percentage of private healthcare coverage was obtained from the Ministry of Health[27]. Details on all variables are listed in Appendix 1. Geographical coordinates of municipalities for the spatial analyzes were obtained from the Brazilian Institute of Geography and Statistics (IBGE)[21].

MACHINE LEARNING MODELS

Municipalities were randomly split into train and test sets (80% and 20% of the total sample, respectively). The train set was used to tune the hyperparameters with 3 repeated 10-fold cross-validation. We first tested the predictive performance of 9 popular machine learning algorithms: linear regression, LASSO regression, Ridge regression, random forests (RF), extreme gradient boosting (XGB), linear support vector machines (LSVM), polynomial support vector machines (PSVM), conditional inference model tree and decision trees. Model performance in the train set was measured with R-squared with 95% confidence intervals (95%CI) and the four best algorithms according to this metric were selected (XGB, RF, pSVM and LASSO).

For each algorithm we performed hyperparameter selection with 10-fold cross-validation algorithm trained with a random search algorithm using 3 repetitions with standard variations provided by the caret package[19]. Model performance was measured solely in the test set.

After selecting the best-performing combination of hyperparameters of the algorithms, each was trained on the whole set with 10-fold cross-validation and the results of the test folds were the predictive values for the next steps of the analysis, in order to guarantee that every municipality has a test set result. Variable importance analysis was performed with SHAP (Shapley Additive Explanation)[28] in the best performing model.

GEOGRAPHICAL ANALYSIS

The prediction residuals of the machine learning algorithms was used to identify geographical clusters of higher-than-expected cancer mortality rates with Kulldorff scan-statistics[13] using total number of cases, predicted values obtained from the machine learning algorithms, 2012 projected population[29] and municipalities centroids obtained from IBGE[22]. The only parameter in the Kulldorff scan statistic is the maximum cluster size, to be determined by spatial territory or by the population at risk. Kulldorff[30] reported that a scan window up to 50% of the population at risk is the ideal rule of thumb to avoid negative cluster detection. Due to the low value and high variance of population density in Brazil, a cluster size of 0.3% of the total Brazilian population was considered to capture a meaningful territorial extension for the objective of this study.

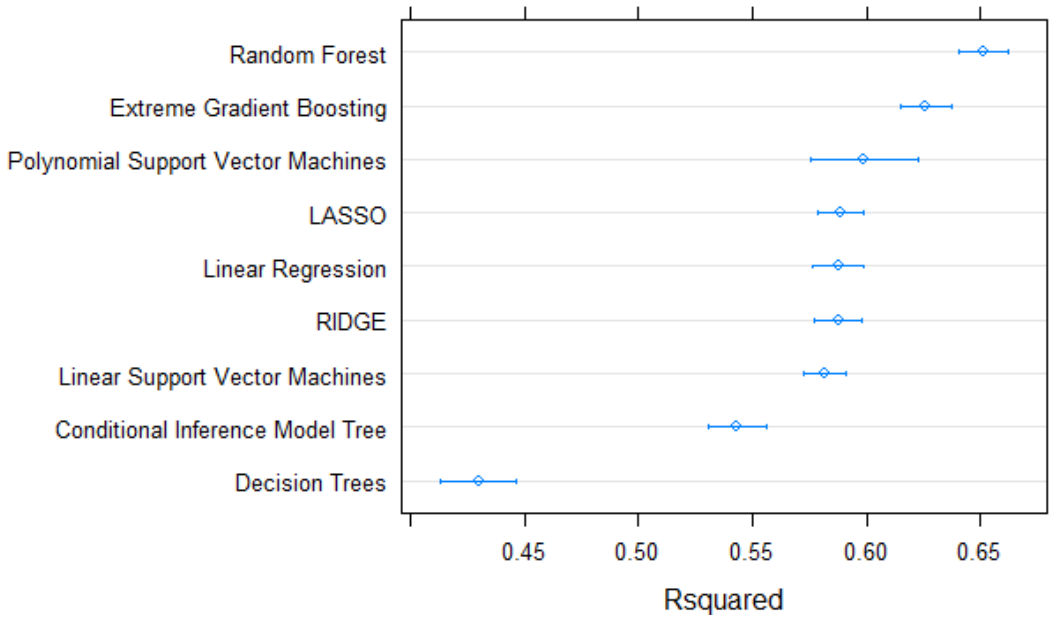
SUB ANALYSIS FOR SPECIFIC TYPES OF CANCER

The analysis was first performed for all cancers combined (Chapter 2 of ICD-10) and then specific analyses were performed for the ten types of cancer with the highest number of deaths (Annex 2).

RESULTS

In the algorithm selection phase, best performing algorithms were: RF (R^2 0.651, 95%CI 0.640-0.662), XGB (R^2 0.626, 95%CI 0.615-0.637), pSVM (R^2 0.599, 95%CI 0.576-0.622) and LASSO (R^2 0.588, 95%CI 0.578-0.598). The results for every algorithm in the selection phase are presented in Figure 2.

Figure 2 - Resample plot of model performance in the train set with 95% Confidence Intervals.



After hyperparameter tuning, the XGB model presented the best performance for predicting total cancer mortality rates (R^2 0.65 in the test set and 0.66 in the whole database). A correlation plot for the XGB model is shown in Figure 4A.

The SHAP analysis of the XGB model shows that the most important predictive variable was the percentage of White residents (Figure 3). This variable had a positive relation with the prediction in the entire range of the distribution. The second most important variable was computer ownership. This variable had a non-linear relation with a growing contribution until 30%, stabilizing above this value. Per capita births was the third most

important variable, with a positive relationship until up to 25 births per thousand inhabitants, also stabilizing above this rate.

A total of three geographic clusters of residual mortality rates (high predictive error in the overall set) were identified by the Kulldorff Statistics (Figure 4). The primary cluster, with the lowest p-value (p = 0.001), was the region between Rio Grande and Bagé in the State of Rio Grande do Sul (RS), with an excess of 28.6 deaths per 100,000 residents (p = 0.001). The secondary clusters were the region of Porto Velho in the State of Rondônia (RO), with an excess of 27.3 deaths per 100,000 people (p = 0.001), and the city of Barueri in the State of São Paulo (SP) with an excess mortality rate of 38.4 deaths per 100,000 residents (details in Table 1).

Figure 3 - SHAP analysis of the top 12 contributing variables contribution to the model.

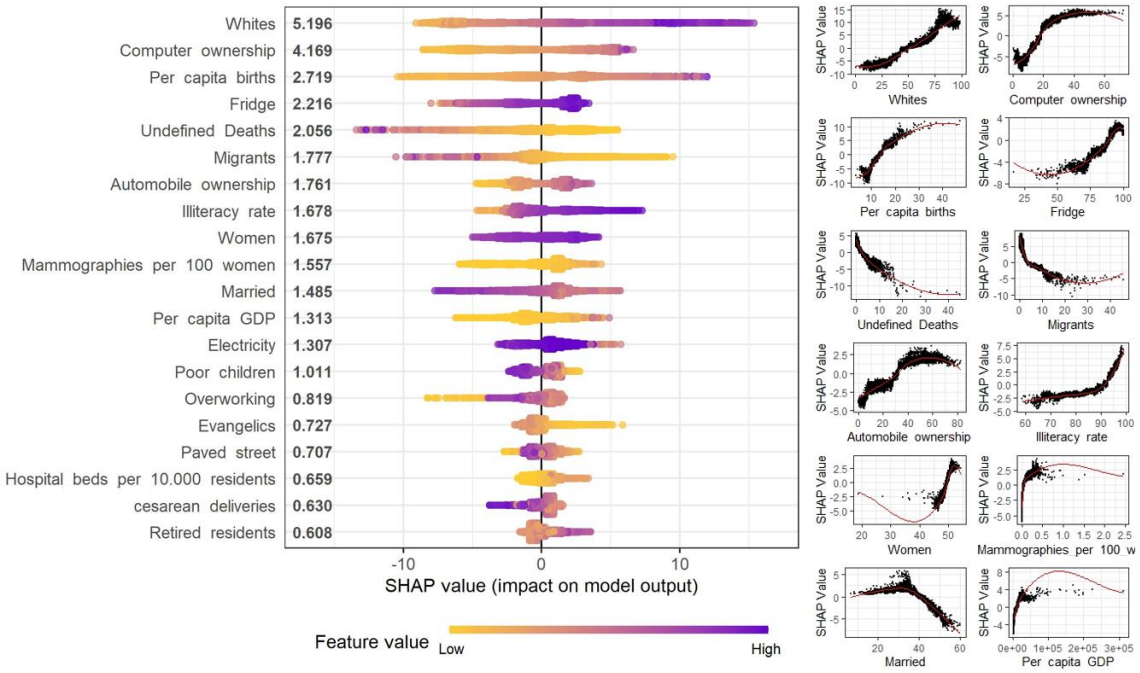


Figure 4 – XGB model results: (A) Correlation plot with R^2 0.66 for Cancer Adjusted Mortality per 100.000 people in Brazilian Municipalities (colored by Kulldorff clusters) (B) Residuals plotted in Brazilian map with clusters identified by color. Zoom in Barueri (C), Bagé-Rio Grande Cluster (D) and Porto Velho and Surroundings Cluster (E).

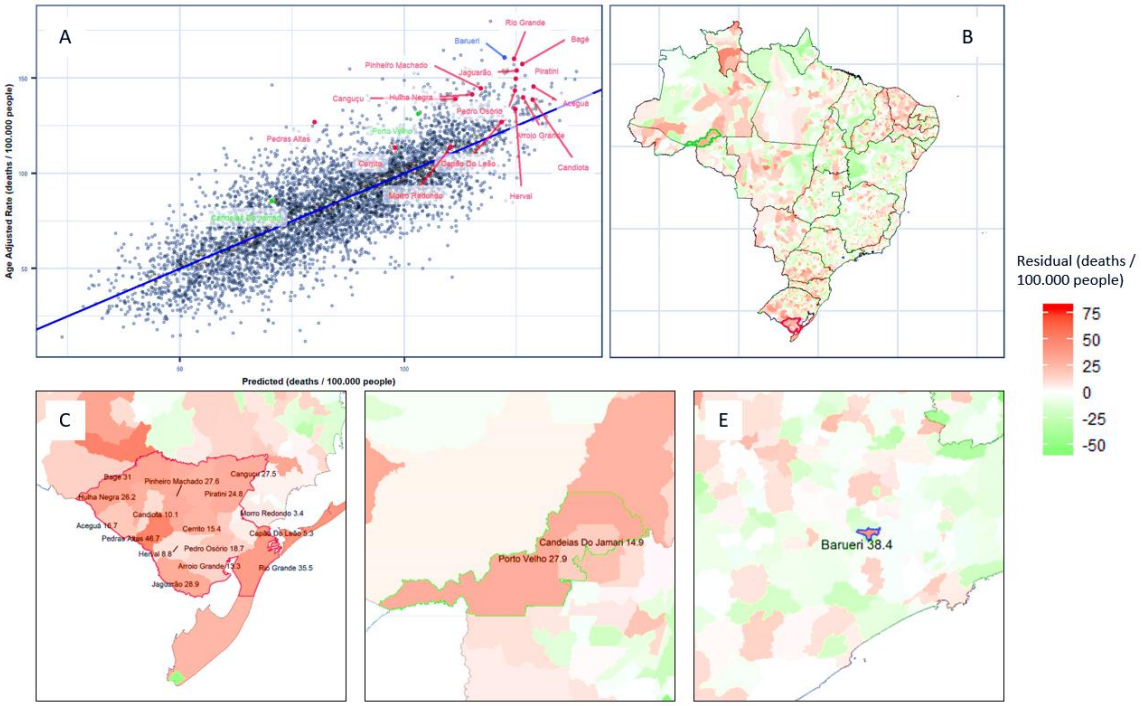


Table 1 - Cancer Mortality Rates by Kulldorff Statistic cluster and Municipality by the XGB Algorithm (Brazil 2008-2016)

Cluster/Municipality	Adjusted Rate (deaths per 100.000)	Predicted Rate (deaths per 100.000)	Residual (deaths per 100.000)	Population (thousands)	Additional Cases
Cluster 1	**151.3**	**122.7**	**28.6**	**538.2**	**15 4.1**
Rio Grande	160.0	124.5	35.5	205.2	72.8
Bagé	157.3	126.3	31.0	121.0	37.5
Canguçu	138.9	111.4	27.5	55.3	15.2
Jaguarão	154.0	125.1	28.9	28.6	8.3
Capão Do Leão	127.0	121.7	5.3	25.2	1.3
Piratini	149.6	124.9	24.8	20.6	5.1
Arroio Grande	139.8	126.5	13.3	19.0	2.5
Pinheiro Machado	144.6	117.0	27.6	13.1	3.6
Candiota	138.6	128.6	10.1	9.2	0.9
Pedro Osório	143.4	124.7	18.7	8.0	1.5
Herval	133.6	124.8	8.8	7.0	0.6
Cerrito	113.3	97.9	15.4	6.5	1.0
Morro Redondo	113.8	110.4	3.4	6.5	0.2
Hulha Negra	141.3	115.1	26.2	6.3	1.7
Aceguá	145.5	128.8	16.7	4.6	0.8
Pedras Altas	126.8	80.1	46.7	2.2	1.0
Cluster 2	**129.0**	**101.7**	**27.3**	**498.0**	**13 6.0**

Porto Velho	131.0	103.2	27.9	475.7	13 2.7
Candeias Do Jamari	85.4	70.5	14.9	22. 4	3.3
Cluster 3	**160.7**	**122.4**	**38.4**	**253 .9**	**97. 4**
Barueri	160.7	122.4	38.4	253 .9	97. 4
Brazil - Others	**98.5**	**98.2**	**0.3**	**197 ,908.7**	**58 4.6**

When comparing different prediction algorithms for each specific cancer type, the RF and XGB algorithms had the highest overall predictive performance. Using the Kulldorff Statistic, the highest number of clusters was identified by the LASSO model. No significant clusters were identified for liver, pancreatic, breast, prostate and brain cancers in any of the models (details on Table 2).

Table 2 - Prediction Accuracy and Clusters Identified by the Kulldorff Statistic for different Machine Learning Algorithms (Brazil 2008-2016)

Cancer Type	**Algorithm**	**R2 in Overall Set**	**Number of Clusters Identified**	**Number of Clusters Identified (p<0.05)**
Total Cancer	XGB	0.66	3	3
Total Cancer	Random Forest	0.65	4	4
Total Cancer	pSVM	0.61	4	4
Total Cancer	LASSO	0.59	12	12
C15 - Esophagus	XGB	0.32	5	5
C15 – Esophagus	Random Forest	0.32	5	5
C15 – Esophagus	pSVM	0.30	8	8
C15 – Esophagus	LASSO	0.24	5	5
C16 – Stomach	XGB	0.13	1	1
C16 – Stomach	Random Forest	0.15	1	1

C16 – Stomach	pSVM	0.08	5	5
C16 – Stomach	LASSO	0.08	5	5
C22 – Liver	XGB	0.07	1	0
C22 – Liver	Random Forest	0.08	1	0
C22 – Liver	pSVM	0.07	1	0
C22 – Liver	LASSO	0.07	1	0
C25 – Pancreatic	XGB	0.25	1	0
C25 – Pancreatic	Random Forest	0.27	1	0
C25 – Pancreatic	pSVM	0.25	1	0
C25 – Pancreatic	LASSO	0.25	1	0
C34 – Lung	XGB	0.51	1	0
C34 – Lung	Random Forest	0.53	1	0
C34 – Lung	pSVM	0.45	5	5
C34 – Lung	LASSO	0.45	7	7
C50 – Breast	XGB	0.24	1	0
C50 – Breast	Random Forest	0.26	1	0
C50 – Breast	pSVM	0.24	1	0
C50 – Breast	LASSO	0.24	1	0
C61 – Prostate	XGB	0.11	1	0
C61 – Prostate	Random Forest	0.12	1	0
C61 – Prostate	pSVM	0.08	1	0
C61 – Prostate	LASSO	0.07	1	0
C71 - Brain	XGB	0.15	1	0
C71 – Brain	Random Forest	0.16	1	0
C71 – Brain	pSVM	0.15	1	0
C71 - Brain	LASSO	0.15	1	0
C80 – Unsp. Location	XGB	0.11	7	7
C80 – Unsp. Location	Random Forest	0.12	8	8
C80 – Unsp. Location	pSVM	0.09	10	10
C80 – Unsp. Location	LASSO	0.09	7	7

Figure 5 and Table 3 provide details on the overlapping clusters identified by each model and cancer types. For total cancer, a cluster composed of 16 cities (Figure 5 H1) was identified by four algorithms in the region between Bagé and Rio Grande, and in this same region a cluster for colorectal cancer was identified by 2 models (Figure 5 I5). Several overlapping clusters of total cancer, lung cancer and stomach cancer were identified in the Region of Porto Velho and surrounding areas (Figure 5D). In the region of Macapá, two clusters for stomach cancer were identified: the first, identified by three models (Figure

5B2), intersects the city of Macapá and the second, around Santana, was identified by one model (Figure 5B1). Multiple clusters were identified in the State of Ceará for different cancer types, by the LASSO and pSVM models (Figure 5C).

*Figure 5 – Geographic clusters location for various cancer types and regions. **(A)** Brazil. **(B)** Macapá region with two overlapping clusters of Stomach Cancer. **(C)** Ceará State with different clusters for Total Cancer, Lung Cancer and Stomach Cancer. **(D)** Porto Velho Region with four overlapping clusters for Total Cancer, Lung Cancer and Stomach Cancer. **(E)** Region in the State of Paraná with different combinations of cluster for Total Cancer, Esophagus Cancer and Non-Specified Location Cancer. **(F)** Southeast Brazil with 7 clusters for Esophagus Cancer. **(G)** Rio Grande do Sul State with a varied combination of clusters for different types, specified for Total and Lung Cancer **(H)** and Esophagus, Colorectal and Non-Specified Location Cancers **(I)**.*

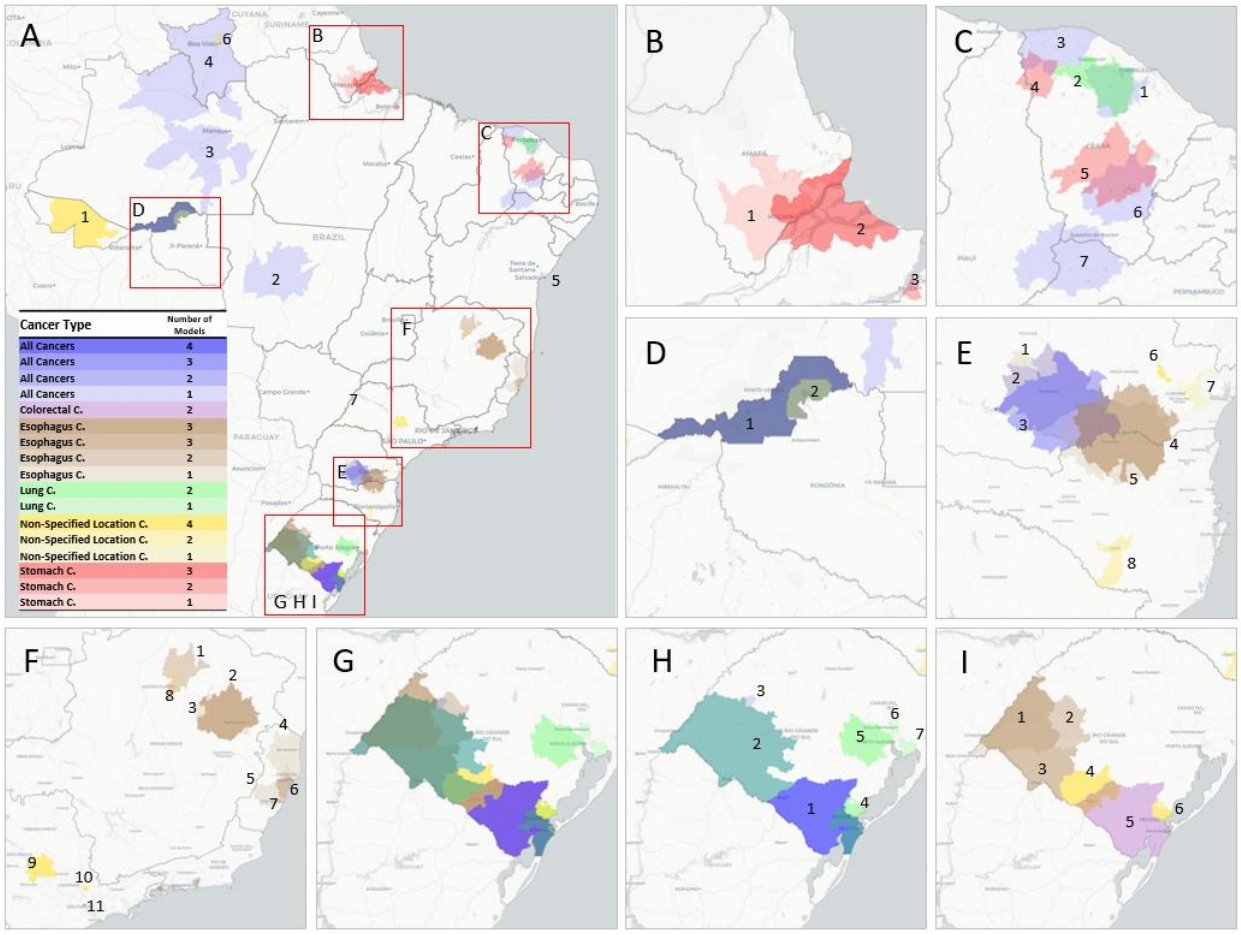


Table 3 - Figure 5 Clusters Details

Cancer Type	umber of Models	ig.4 Location	Models	umber of Cities	opulation	P ean Estimated	umber of Cases	luster Value Range	C p ean Excess Cases
Total Cancer	1		XGB, RF, SVM, LASSP	6	38.220	5 43	15	0. 001 - 0.002	7%
Total Cancer	11		XGB, RF, LASSO		53.877	2 09	08	0. 001 - 0.005	2%
Total Cancer	1		SVM, LASSO		75.691	4 47	23	0. 001 - 0.001	0%
Total Cancer	3		SVM, LASSO	0	87.156	5 41	20	0. 001 - 0.001	8%
Total Cancer			LASSO			5		0.	

	1			8	00.042	34	45	001 - 0.001	3%
Total Cancer	4		LASSO	8	56.312	5 16	50	0. 001 - 0.001	2%
Total Cancer	2		LASSO	6	74.415	5 35	06	0. 001 - 0.001	7%
Total Cancer	3		LASSO	3	92.999	5 92	96	0. 002 - 0.002	7%
Total Cancer	7		LASSO	2	39.167	5 56	42	0. 046 - 0.046	4%
Total Cancer	6		LASSO	6	86.597	5 28	37	0. 002 - 0.002	5%
Total Cancer	2		LASSO	2	74.579	4 25	44	0. 001 - 0.001	8%
Total Cancer	5		LASSO		08.748	3 96	75	0. 047 - 0.047	7%
Total Cancer	2		XGB		98.045	4 07	42	0. 001 - 0.001	7%
Total Cancer	3		RF	2	46.095	4 36	15	0. 04 - 0.04	3%
Total Cancer	2		RF	3	01.607	4 55	64	0. 047 - 0.047	4%
Total Cancer	(2,3)		SVM	7	76.921	5 41	09	0. 003 - 0.003	6%
Colorectal C.	5	LASSO	SVM,	6	38.220	5 7	9	0. 001 - 0.001	0%
Esophagus C.	2	LASSO	XGB, RF,	8	71.144	5 1	1	0. 002 - 0.007	0%
Esophagus C.	4	LASSO	XGB, RF,	8	65.803	5 5	0	0. 001 - 0.001	6%
Esophagus C.	6	LASSO	XGB, RF,		86.686	5 6	7	0. 001 - 0.003	8%
Esophagus C.	1		RF, SVM	0	73.875	5 1	7	0. 033 - 0.035	2%
Esophagus C.	(1,2)		XGB, RF	6	84.172	5 4	5	0. 001 - 0.004	8%
Esophagus C.	5		SVM	7	83.542	5 4	0	0. 001 - 0.001	2%
Esophagus C.	1		SVM	5	31.730	4 2	8	0. 033 - 0.033	0%
Esophagus C.			XGB		43.871	4 1	7	0. 013 - 0.013	6%
Esophagus C.	3		SVM	1	91.580	5 7	4	0. 017 - 0.017	6%
Esophagus C.	4		LASSO	7	87.888	5 7	5	0. 031 - 0.031	7%
Esophagus C.	(1,3)		LASSO	0	76.257	4 8	5	0. 001 - 0.001	6%
Esophagus C.	(5,7)		SVM		51.007	5 9	6	0. 015 - 0.015	9%
Esophagus C.	7		SVM		22.051	5 4	1	0. 002 - 0.002	21%
Esophagus C.	1		SVM		23.479	3 1	9	0. 004 - 0.004	6%
Esophagus C.	7		SVM		53.043	3	1	0. 013 - 0.013	63%
Lung C.	2	LASSO	SVM,	6	74.415	5 7	52	0. 001 - 0.001	7%
Lung C.	4	LASSO	SVM,		91.722	5 07	61	0. 001 - 0.001	0%
Lung C.	5	LASSO	SVM,	0	06.420	5 00	46	0. 002 - 0.007	6%
Lung C.	2	LASSO	SVM,		98.045	4 9	9	0. 002 - 0.002	8%
Lung C.	2	LASSO	SVM,	6	55.057	5 0	0	0. 015 - 0.019	5%
Lung C.	6		LASSO		24.379	5 06	47	0. 044 - 0.044	9%
Lung C.	7		LASSO		52.728	4 6	24	0. 047 - 0.047	4%
Non-Specified Location C.	9	SVM, LASSO	XGB, RF,	6	22.324	4 4	7	0. 013 - 0.037	6%
Non-Specified Location C.	1	SVM, LASSO	XGB, RF,		08.427	5 9	7	0. 001 - 0.001	00%
Non-Specified Location C.	6	SVM, LASSO	XGB, RF,		40.257	3 6	5	0. 001 - 0.001	22%

Non-Specified Location C.	4	XGB, RF, SVM, LASSO		68.828	1	7	6	0.001 - 0.001	85%
Non-Specified Location C.	6	XGB, RF, SVM, LASSO		5.389	2	1	4	0.001 - 0.001	300%
Non-Specified Location C.	10	XGB, RF, SVM, LASSO		08.145	1	4	5	0.001 - 0.003	75%
Non-Specified Location C.	6	SVM, LASSO		03.002	3	1	9	0.001 - 0.029	71%
Non-Specified Location C.	8	RF, SVM		59.076	1	7	7	0.023 - 0.04	62%
Non-Specified Location C.	8	RF		81.407	3	5	7	0.019 - 0.019	0%
Non-Specified Location C.	7	SVM		9.190	3	1		0.01 - 0.01	00%
Non-Specified Location C.	7	SVM		24.942	5	0	5	0.013 - 0.013	5%
Stomach C.	2	XGB, SVM, LASSO		97.394	5	1	2	0.001 - 0.022	2%
Stomach C.	4	SVM, LASSO	3	64.233	3	7	0	0.021 - 0.028	5%
Stomach C.	2	SVM, LASSO		98.045	4	3	8	0.017 - 0.023	6%
Stomach C.	5	SVM, LASSO	7	85.956	5	8	5	0.011 - 0.014	1%
Stomach C.	3	SVM, LASSO		94.948	5	6	0	0.001 - 0.002	4%
Stomach C.	1	RF		77.729	5	7	0	0.006 - 0.006	8%

Regarding esophagus cancer, two overlapping clusters were identified in the region between Parana and Santa Catarina (Figure 5: E4 and E5) by four models. Three overlapping clusters were identified in the western region of the State of Rio Grande do Sul (Figure 5: I1, I2 and I3), and a similar situation was identified in the region around the city of Teófilo Otoni (Figure 5: F2 and F3). Four clusters were identified in the State of Espírito Santo (Figure 5: F4, F5, F6 and F7).

The mean excess mortality rate not explainable by sociodemographic characteristics was highest for non-specified cancer (weighted mean: 126% ± weighted standard deviation: 143% with weights on expected cases), followed by stomach cancer (73% ± 10%), colorectal cancer (one cluster 70%), esophagus cancer (67% ± 24%), lung cancer (51% ± 11%) and total cancer (28% ± 4%).

DISCUSSION

Machine learning algorithms were able to predict cancer mortality with high overall performance using only socioeconomic and health care coverage factors. Cancer mortality is the consequence of incidence and lethality, and both have a strong association with socioeconomic characteristics[31,32]. Therefore, machine learning algorithms may provide a better modelling approach when compared to traditional statistical tools. We used the predicted values to identify statistically significant clusters of excess cancer mortality (i.e. higher than expected rates) throughout Brazil. There were consistent and significant cluster

overlaps for the different algorithms, especially in the Southern and Northern regions of the country.

The area between Bagé and Rio Grande (Figure 5 H1), the southernmost region of Brazil, was identified by all models, for both total cancer and colorectal cancer. Lung cancer clusters were also particularly common in this State, with three different clusters around the capital Porto Alegre, one cluster in Pelotas/Rio Grande and another in the western area of the State. For stomach cancer, the region of Macapá in Amapá State, in the North of Brazil, showed an 82% excess in mortality. These cancer types are etiologically related to tobacco smoking[33]. Interestingly, the three regions of Macapá, Porto Alegre and Porto Velho, are the top three state capitals with highest rates of smoking habits in males[34]. A cluster analysis from stomach cancer cases in Central America has identified a possible association with germline as well as a hotspot for *H. pylory* infection and might indicate a focus of future epidemiologic research in this region[35].

Although the clusters of lung cancer in Rio Grande do Sul and stomach cancer in Amapá are geographically consistent with cancer incidence analyzes of National Institute of Cancer (INCA)[3], high incidence rates do not necessarily coincide with areas with an anomalous number of cases, as the large variance of sociodemographic characteristics throughout Brazil could mean that even high cancer rates are within the expected value, given these local characteristics. Our study, by first predicting the cancer mortality rate of Brazilian municipalities using sociodemographic characteristics (and showing that they have a high predictive ability), was then able to identify spatial clusters with higher-than-expected cancer mortality rates considering socioeconomic and health care coverage factors.

It is important to note that no significant clusters were found for breast, prostate, liver, pancreatic and brain cancers. One possible reason is that some of these specific cancers have low incidence and are therefore amenable to random local variations that decrease the predictive ability of the machine learning models. Another possibility is that these cancers are less significantly affected by other factors beyond sociodemographic characteristics. Eleven clusters of non-specified location cancer were also found, six of them by all four models tested. These clusters may indicate regions with a lack of specialization of the mortality registration services in identifying specific tumors.

Even though etiological reasoning regarding increased cancer mortality rates is beyond the scope of this study, we encourage further studies to collect new local data to

confirm our findings. However, most of the cancer types with clusters identified in this study are associated with avoidable risk factors[33]. For example, for lung cancer and esophagus cancers, the main risk factor is tobacco smoking[33] and anti-tobacco campaigns have shown important results in reducing tobacco consumption, thus reducing local rates of lung cancer[36]. It is important, though, to notice that part of tobacco consumption influence at the local areas may have been attenuated due to its relation with socioeconomic factors[37]. Additionally, the cultural uniqueness of the southern region of Brazil of hot mate tea consumption and intensive barbecue grill (coal burning), are topics for further investigation due to its observed association with gastric[38] and lung[33] cancer, respectively. Further, for gastric cancer the most important risk factor is the presence of *Helicobacter Pylori*[33], and its screening was shown to reduce gastric cancer incidence by 35%[39] and in a cost-effective way[40]. The use of spatial clustering methods associated with machine learning modeling may improve geography-targeted screening, campaigns and epidemiological field studies.

Variable importance analysis of the machine learning algorithms found that computer ownership, automobile ownership, electricity coverage, percentage of houses with fridge and literacy rate increased the probability of a high prediction of cancer mortality rates. Socioeconomic factors have been associated with cancer incidence and mortality, especially given that high-income individuals are able to treat competing risk factors such as cardiovascular diseases and diabetes, as well as due to the presence of differences in dietary and life-style behaviors[41–43].

There are important limitations to this study. First, it was not able to provide reasoning for the excessive mortality rates found, although it can provide important guidance for future epidemiological field research regarding environmental, genetic, behavioral and socioeconomic risk factors[12]. Second, this study analyzed mortality instead of incidence since there is no national coverage of a Population Based Cancer registry in Brazil and mortality data is of higher quality and available at the municipal level. Although mortality can be a good proxy for incidence[44], it may include other sources of variability especially in cancer types with a high survival rates.

CONCLUSION

The combination of clustering statistics with machine learning modeling is a promising and versatile tool that can be used with many different types of outcomes. Several clusters of excessive mortality were identified in various regions of Brazil for lung, stomach,

esophagus, colorectal, and total cancer. These findings provide insights for further epidemiological field work as well as guidance for focused healthcare actions.

REFERENCES

1. Ferlay, J. *et al.* Global cancer observatory: cancer today. *Int. Agency Res. Cancer* Available from https://gco.iarc.fr/today (2020).
2. World Health Organization. Mortality Database. *Cancer Mortality Database* (2019). Available at: https://www-dep.iarc.fr/WHOdb/WHOdb.htm. (Accessed: 4th September 2020)
3. Ministério da Saúde do Brasil. *Estimativa INCA 2020: Incidência de Câncer no Brasil*. (2019).
4. Elfiky, A., Pany, M., Parikh, R. & Obermeyer, Z. A machine learning approach to predicting short-term mortality risk in patients starting chemotherapy. *bioRxiv* 204081 (2017). doi:10.1101/204081
5. Mucaki, E. J., Zhao, J. Z. L., Lizotte, D. J. & Rogan, P. K. Predicting responses to platin chemotherapy agents with biochemically-inspired machine learning. *Signal Transduct. Target. Ther.* **4**, 1 (2019).
6. Thorsen-Meyer, H.-C. *et al.* Dynamic and explainable machine learning prediction of mortality in patients in the intensive care unit: a retrospective study of high-frequency data in electronic patient records. *Lancet Digit. Heal.* **2**, e179–e191 (2020).
7. Taylor, R. A. *et al.* Prediction of In-hospital Mortality in Emergency Department Patients with Sepsis: A Local Big Data-Driven, Machine Learning Approach. *Acad. Emerg. Med.* (2016). doi:10.1111/acem.12876
8. Singal, A. G. *et al.* Machine learning algorithms outperform conventional regression models in predicting development of hepatocellular carcinoma. *Am. J. Gastroenterol.* **108**, 1723–1730 (2013).
9. Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V. & Fotiadis, D. I. Machine learning applications in cancer prognosis and prediction. *Comput. Struct. Biotechnol. J.* **13**, 8–17 (2015).
10. Park, K. *et al.* Robust predictive model for evaluating breast cancer survivability. *Eng. Appl. Artif. Intell.* **26**, 2194–2205 (2013).
11. Chiavegatto Filho, A. D. P., dos Santos, H. G., do Nascimento, C. F., Massa, K. & Kawachi, I. Overachieving Municipalities in Public Health. *Epidemiology* **29**, 836–840 (2018).
12. Elliott, P. & Wartenberg, D. Spatial epidemiology: Current approaches and future challenges. *Environmental Health Perspectives* **112**, 998–1006 (2004).
13. Kulldorff, M. A spatial scan statistic. *Commun. Stat. - Theory Methods* **26**, 1481–1496 (1997).
14. Lin, H. *et al.* Lung cancer mortality among women in Xuan Wei, China: A

comparison of spatial clustering detection methods. *Asia-Pacific J. Public Heal.* **27**, NP392–NP401 (2015).

15. Sherman, R. L. *et al.* Applying spatial analysis tools in public health: An example using satscan to detect geographic targets for colorectal cancer screening interventions. *Prev. Chronic Dis.* **11**, (2014).

16. Kulldorff, M., Feuer, E. J., Miller, B. A. & Freedma, L. S. Breast Cancer Clusters in the Northeast United States: A Geographic Analysis. *Am. J. Epidemiol.* **146**, 161–170 (1997).

17. Queiroz, B. L. *et al.* Estimativas do grau de cobertura e da mortalidade adulta (45q15) para as unidades da federação no Brasil entre 1980 e 2010. *Rev. Bras. Epidemiol.* **20**, 21–33 (2017).

18. National Cancer Institute (NCI). World (WHO 2000-2025) Standard - Standard Populations - SEER Datasets. (2013). Available at: https://seer.cancer.gov/stdpopulations/world.who.html. (Accessed: 10th February 2019)

19. Kuhn, M. *et al.* caret: Classification and Regression Training. (2018).

20. R Core Team. R: A Language and Environment for Statistical Computing. (2018).

21. IBGE. Censo Demográfico 2010. (2010). Available at: https://www.ibge.gov.br/home/estatistica/populacao/censo2010/default.shtm. (Accessed: 17th September 2017)

22. Instituto Brasileiro de Geografia e Estatística (IBGE). *Áreas Territoriais*. (2020).

23. International Monetary Fund (IMF). *World Economic Outlook Database*. (2020).

24. United Nations Development Program (UNDP). *Human Development Reports*. (2019).

25. World Bank. *Gini index (World Bank estimate) - Brazil*.

26. World Bank. *Poverty Data - Brazil*.

27. ANS, A. N. de S. Su. Beneficiários de planos privados de saúde, por cobertura assistencial (Brasil – 2009-2019). (2019).

28. Lundberg, S. M. *et al. Explainable AI for Trees: From Local Explanations to Global Understanding*.

29. IBGE. Projeção da população do Brasil e das Unidades da Federação. *Instituto Brasileiro de Geografia e Estatística* (2019).

30. Kulldorff, M. & Nagarwalla, N. Spatial disease clusters: Detection and inference. *Stat. Med.* **14**, 799–810 (1995).

31. Lundqvist, A., Andersson, E., Ahlberg, I., Nilbert, M. & Gerdtham, U. Socioeconomic inequalities in breast cancer incidence and mortality in Europe - A systematic review and meta-analysis. *Eur. J. Public Health* **26**, 804–813 (2016).

32. Faggiano, F., Partanen, T., Kogevinas, M. & Boffetta, P. Socioeconomic differences in cancer incidence and mortality. *IARC Sci. Publ.* 65–176 (1997).

33. Cogliano, V. J. *et al.* Preventable exposures associated with human cancers. *Journal of the National Cancer Institute* **103**, 1827–1839 (2010).

34. Departamento de Análise de Situação e Saúde. *VIGILÂNCIA DE FATORES DE RISCO E PROTEÇÃO PARA DOENÇAS CRÔNICAS POR INQUÉRITO TELEFÔNICO - VIGITEL*. (MINISTÉRIO DA SAÚDE, 2007).

35. Dominguez, R. L. *et al.* Geospatial analyses identify regional hot spots of diffuse gastric cancer in rural Central America. *BMC Cancer* **19**, 1–8 (2019).

36. Hurley, S. F. & Matthews, J. P. Cost-effectiveness of the Australian National Tobacco Campaign. *Tob. Control* **17**, 379–384 (2008).

37. Hosseinpoor, A. R., Parker, L. A., Tursan d'Espaignet, E. & Chatterji, S. Social determinants of smoking in low- and middle-income countries: Results from the world health survey. *PLoS One* **6**, (2011).

38. Sewram, V., De Stefani, E., Brennan, P. & Boffetta, P. Maté consumption and the risk of squamous cell esophageal cancer in Uruguay. *Cancer Epidemiol. Biomarkers Prev.* **12**, 508–513 (2003).

39. FUCCIO, L., ZAGARI, R. M., MINARDI, M. E. & BAZZOLI, F. Systematic review: Helicobacter pylori eradication for the prevention of gastric cancer. *Aliment. Pharmacol. Ther.* **25**, 133–141 (2006).

40. Lansdorp-Vogelaar, I. & Sharp, L. Cost-effectiveness of screening and treating Helicobacter pylori for gastric cancer prevention. *Best Practice and Research: Clinical Gastroenterology* **27**, 933–947 (2013).

41. Gersten, O. & Wilmoth, J. R. The Cancer Transition in Japan since 1951. *Source Demogr. Res.* **7**, 271–306 (2002).

42. Omran, A. R. The epidemiologic transition: A theory of the epidemiology of population change. *Milbank Mem. Fund Q.* **49**, 509–538 (1971).

43. Barrett, R., Kuzawa, C. W., McDade, T. & Armelagos, G. J. *Emerging and re-emerging infectious diseases: the third epidemiologic transition*. *Emerging and re-emerging infectious diseases: the third epidemiologic transition* (Annual Reviews Inc, 1998). doi:10.1146/annurev.anthro.27.1.247

44. Black, R. J., Bray, F., Ferlay, J. & Parkin, D. M. Cancer incidence and mortality in the European union: Cancer registry data and estimates of national incidence for 1990. *Eur. J. Cancer Part A* **33**, 1075–1107 (1997).

SUPPLEMENTARY MATERIALS

Annex 1 - Sociodemographic varialbles summary statistics

	Overall (N=5565)

Variable	Mean (SD)	Range
Life expectance	73.089 (2.681)	65.300 - 78.640
Residents	34277.772 (203112.622)	805.000 - 11253503.000
Median age	29.025 (4.367)	13.590 - 46.740
Elderly	12.090 (3.286)	2.550 - 29.190
Dependency ratio	60.255 (9.010)	15.940 - 122.980
Women	49.504 (1.569)	18.910 - 54.240
Per capita births	13.936 (3.609)	3.400 - 47.460
Married	36.535 (8.683)	6.550 - 64.040
Evangelics	17.100 (9.465)	0.420 - 85.840
Disability rate	24.565 (4.746)	2.540 - 44.330
Municipal density	108.202 (572.445)	0.130 - 13024.600
Urban area	63.835 (22.036)	4.180 - 100.000
Fridge	88.620 (11.627)	16.660 - 100.000
Computer ownership	21.330 (13.982)	0.440 - 72.700
Automobile ownership	31.869 (19.948)	0.000 - 90.720
Household density	3.391 (0.431)	2.560 - 6.920
Favela (slums) residents	0.132 (0.402)	0.000 - 9.880
Electricity	97.076 (5.801)	29.520 - 100.000
Green spaces	0.433 (0.245)	0.000 - 0.980
Paved street	0.465 (0.236)	0.000 - 0.990
Whites	46.948 (24.049)	0.860 - 99.160
Illiteracy rate	85.259 (8.936)	58.400 - 99.100
College education	5.507 (3.275)	0.280 - 33.840
Highschool completion	16.183 (6.051)	1.860 - 47.470
Migrants	5.501 (4.476)	0.000 - 45.750
Foreigners	0.076 (0.546)	0.000 - 37.720
Median income	566.155 (265.771)	128.770 - 2210.720
Unemployment	3.752 (1.993)	0.000 - 16.990
Child labor	13.026 (8.301)	0.000 - 72.090
Retired residents	16.712 (4.669)	2.320 - 40.180
Overworking	28.615 (10.956)	0.900 - 73.090
Poor children	59.951 (22.496)	2.440 - 95.530
Per capita GDP	12587.939 (14676.853)	2257.990 - 312257.340
Gini coefficient	0.503 (0.066)	0.280 - 0.810
Bolsa Familia Coverage	75.927 (18.751)	0.000 - 100.000
Communiting	10.816 (9.815)	0.000 - 69.330
Priv Insurance	9.120 (11.848)	9.120 (11.848)
Mammographies per 100 women	0.104 (0.112)	0.000 - 2.460
Oral Health Strategy coverage	65.462 (37.253)	65.462 (37.253)
Primary health coverage for poor residents	79.617 (27.193)	0.000 - 100.000
Vaccination coverage	79.098 (10.723)	0.000 - 182.370
Low birth weight	7.677 (3.473)	0.000 - 40.000
Hospital beds per 10.000 residents	1.546 (1.843)	0.000 - 24.540
Family Health Strategy teams per 10.000	0.296 (0.141)	0.000 - 1.230

residents		
Ultrasound machines. per 10.000 live births	4.573 (8.390)	0.000 - 125.000
cesarean deliveries	50.896 (18.411)	3.330 - 100.000
Xray machines per 10.000 residents	0.111 (0.144)	0.000 - 1.460
Life support equipment per 10.000 residents	0.303 (0.369)	0.000 - 3.250
Undefined Deaths	3.134 (3.284)	0.000 - 45.238

Annex 2 - Top-10 Cancer Types by mortality (2007 - 2016) and ICD-10 codes.

Cancer Type	**Number of Deaths accounted**
C34 - Lung Cancer	212.379
C18-C21 - Colorectal Cancer	132.802
C50 - Breast Cancer	125.008
C16 - Stomach Cancer	123.188
C61 - Prostate Cancer	120.999
C22 - Liver Cancer	76.704
C25 - Pancreatic Cancer	74.123
C15 - Esophagus Cancer	70.386
C71 - Brain Cancer	64.741
C80 – Non-Specified Location Cancer	56.340