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

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