Submitted to the European Population Conference – 2020

DETERMINANTS OF ADULT MORTALITY IN BRAZILIAN MICRORGIONS IN 2010: AN ANALYSIS BASED ON MACHINE LEARNING MODELS

Natália Martins Arruda[1], Tiago Carvalho[1] , Luciana Correia Alves[2]

Abstract

The aim of the present study was to investigate the relationships among socioeconomic, structural, contextual and health factors and the probability of adult mortality in the Brazilian small areas in 2010. The analyses were based on data from the 2010 Demographic Census and Mortality Information System of DATASUS. For this purpose, the correction of underreporting of deaths was done by the TOPALS method with Bayesian estimation for mortality in small areas. In addition, the machine learning method was used to establish the determinants of probability of adult mortality. Machine learning methods have great potential for this type of analysis, since they allow a better understanding of the interactions between the different factors. The results showed that mortality rates due to external causes, unemployment rate, proportion of blacks, vaccination coverage and proportion of whites were the ones that obtained the greatest predictive power in the risk of adult mortality probability using the algorithms Random Forest, Extreme Boosted Trees, Support Vector Machine and Naive Bayes. The algorithms obtained good performance and were effective in analyzing the variables, although some correlated, with the outcome of adult mortality probability. Identifying the determinants of adult mortality and the main disparities between social groups and in small areas is extremely important in helping to build public policies that respond adequately to the specific needs of each region and social group, thus contributing to reduce the socioeconomic inequalities and mortality.

Keywords - Mortality; Adult; Machine Learning; Socioeconomic Factors; Inequalities

[1] Federal Institute of São Paulo, Campinas-SP, Brazil.

[2] Department of Demography, Population Studies Center Elza Berquó (NEPO), University of Campinas (Unicamp), Campinas-SP, Brazil

Introduction

The identification of household and regional characteristics influencing the risk of adult mortality can help in the development of policies and actions aimed at addressing this health and socioeconomic problem. Increases in life expectancy will be less achieved by greater reductions only in infant mortality. In all countries, future progress in life expectancy will depend on adult mortality to some degree.

Brazil is a country marked by social and economic inequalities. It is among the countries with the highest degree of social inequality in the world. (LIMA COSTA; MATOS; CAMARANO, 2006). In general, studies dealing with adult mortality focused mostly on data quality and coverage rather than on their inequality determinants (MOURA et al., 2016; WALQUE; WILMER, 2013; VASCONCELOS; FRANCA, 2012). Sttudies on determinants of adult mortality are little explored in Brazil (QUEIROZ et al, 2017).

Levels in small areas, defined as microregions, are considered best to avoid bias created by heterogeneity in mortality levels and socioeconomic characteristics of municipalities, and to detect geographic mortality patterns that are sometimes not evident using larger areas (RICHARDSON et al., 2004).

In this Brazilian context, the principal aim of this study was investigate the relationships among socioeconomic, structural, contextual, health factors and the probability of adult mortality (45q15) in the Brazilian microregions in 2010.

Data and Methods

The data came from the 2010 Demographic Census conducted by the Brazilian Institute of Geography and Statistics (IBGE); *Sistema de Informação sobre Mortalidade* (SIM - Mortality Information System), *Cadastro Nacional de Estabelecimentos de Saúde* (CNES - National Register of Health Facilities) and the *Sistema de Informação de Atenção Básica* (SIAB - Information System of the Primary Care) organized by the Ministry of Health through the Department of Informatics of the Brazilian Unified Health System (DATASUS) also for the year 2010. To calculate the probability of adult mortality (45q15), we used the following deaths from the SIM using the TOPALS approach with Bayesian estimation of correction proposed by Schmertmann and Gonzaga (2018) for age specific mortality in small areas with defective vital records incorporating uncertainty about levels of mortality.

The combination of the variables from the different sources was done through the microregion code. By microregions is meant the grouping of bordering municipalities based on economic and social similarities. Brazil is divided between 558 microregions and consists of the analysis sample of this study.

Independent Variables: We sought to combine socioeconomic, demographic, and health access variables at the microregion level to understand the relationships of these variables to the probability of adult mortality through machine learning algorithms. Predictor variables may follow a symmetrical or asymmetric distribution (for continuous variables). Variables within a dataset may or may not have an underlying relationship with a response. (KUHN; JOHNSON, 2013; JAMES et al., 2017).

Brazil is a country of great regional heterogeneity and inequalities in which different levels coexist with regard to health, demographic and socioeconomic indicators. For example, most of the Brazilian microregions have water and sewage coverage with over 90% of households, and some microregions have low coverage (Table 1). Other example is in the socioeconomic indicators - poverty rate, child labor rate, per capita household average income, per capita GDP and unemployment rate - also have significant regional differences, with Brazil's average poverty rate at 7% reaching 30% in the Traipu microregion in the state of Alagoas, Northeast Region of Brazil while the lowest rates are located in the microregions of the South and Southeast.

Table 1 - Summary of the independent variables statistics.

Variables	Name	Min	Median	Mean	Max
gdp per capita	Gross Domestic Products per capita	2867	12105	14105	77872
income	Household Income	162	542	542	1665
higherdegree	Proportion of population with higher education (college)	0.02	0.07	0.07	0.24
cov_garbage	Proportion of households waste collection	0.18	0.79	0.76	1.00
urbanization	Degree of Urbanization	22.34	72.72	72.01	100
hospitalBeds	Hospital beds per 1000 people	0.39	2.12	2.29	9.34
cov_water	Proportion of households with piped water	0.45	0.99	0.96	1.00
cov_sewerage	Proportion of households with sewage collection system	0.31	0.91	0.87	1.00
white	Proportion of white people	0.08	0.40	0.45	0.92
aging_index	Aging Index	8.47	41.82	42.42	87.76
circsystem	Mortality Rate by Circulatory System Diseases	8.05	171.45	169.78	342.55
child.labor	Child Labor Rate	3.31	10.79	11.51	31.27
sex_ratio	Sex Ratio	88	99	100	114
cov_vaccination	Vaccination Coverage	56	77	77	100
cov_FHS	Coverage of Family Health Strategy	0.03	0.77	0.74	1.69
illiteracy	Illiteracy Rate	2.30	11.25	14.43	42.40
poverty_rate	Poverty Rate	0.00	0.04	0.07	0.30
externalcauses	External Causes Mortality Rate	0.00	68.99	70.88	154.87
unemployment	Unemploymente Rate	1.10	6.74	6.93	20.47
blacks	Proportion of black/brown	0.07	0.58	0.53	0.87
young_prop	Proportion of young people (15 to 29 years)	0.21	0.27	0.27	0.32

Source: IBGE(2010), Datasus (2010)

Outcome Variable – Adult Mortality Probability: Initially a continuous variable, the probability was transformed into a two-class variable using the mean probability of adult mortality value. The microregions that had adult death probability below 0.1406 (midpoint) were classified as class P0 and the remainder as class P50. This approach was chosen to improve the performance of the machine learning algorithms that have high robustness to high data dimensionality. The study by Lustgarten et al. (2008) shows that transforming the continuous variable into two classes can help to significantly improve the classification performance of the algorithms

Methods

The Machine Learning method is useful as a replacement or complement to parametric regression. Unlike traditional regression based approaches, machine learning does not impose a parametric model linking a dependent variable with independent variables. The key idea is to let the algorithm find the path to the result and connections between the predictor variables. Moreover, collinearity and assumption violations are not important concerns depending on the chosen algorithm (BILLARI; FÜRNKRANZ; PRSKAWETZ, 2006). We choose in this study

the following models: **Support Vector Machine (SVM)**: It creates a border that best separates the data through support vectors; **Naive Bayes (NB):** based on the predictors that we have observed, what is the probability that the outcome is from Class P0 or P50; **Random Forest (RF):** Construction of decision trees, evaluate the importance of each predictor so that it is possible to identify the relevant variables in the construction of each tree; **Extreme Gradient Boosting (XGBoost):** Decision trees that differ from RF since they start with weak learning trees and build more reliable trees based on the residuals from the predictions of each previous trees.

Pre-processing: as the selected variables have very different scales, such as vaccine coverage, which is a proportion (%) and per capita GDP ranging from 2000 to 20,000, it was decided to perform some tranformation in the variables. (KUHN; JOHNSON, 2013). Two transformations were conducted: centralization and transformation on the same scale. To center a predictor variable, the average predictor value is subtracted from all values. As a result of centralization, the predictor has a zero mean. Similarly, for the data to have the same scale, each value of the predictor variable is divided by its standard deviation. These manipulations are generally used to improve the numerical stability of some Machine Learning algorithms and benefit from predictors being on a common scale. (KUHN; JOHNSON, 2013).

Metrics: to measure the performance of the models we use the confusion matrix generated for each of the models and it was possible to calculate the main measures that were used to compare the algorithms. The following metrics were calculated: Accuracy: Determines the number of predictions made correctly by the model over all predictions made; Sensitivity: ratio of true positives to true positives plus false negatives; Specificity: ratio of true negatives to true negatives plus false positives; Precision: tells us what proportion of Class 1 observations that were classified as such were actually Class 1.

Results and some findings

As we can see on Table 2, Accuracy was similar between decision tree models (XGB and RF) and SVM and lower in the model generated by the Naive Bayes algorithm. Regarding Sensitivity, the most outstanding model, again, was SVM, followed by XGB and RF, the worst performance being presented, also by the NB model, that is, it calculates the proportion of observations that were really Class 1 that were diagnosed by the algorithm as being Class 1, that is, the proportion of true positives.

Table 2 - Model performance comparison

	XGBoost	SVM	RF	NB
Accuracy	75.82%	76.11%	75.81%	67.74%
Sensibility	75.97%	76.82%	73.60%	61.29%
Specificity	75.67%	75.39%	78.02%	74.19%
Precision	75.74%	75.74%	77.00%	70.37%

Source: IBGE(2010), Datasus (2010)

In Figure 1, we show the feature importance, a measure that can point out the features with major relevance in our models. Using the two models with the greatest performance,

XGBoost and SVM, the results showed that the variables presenting higher degree of importance (highly correlated with label class) were Mortality Rate by External Causes, Unemployment Rate, Proportion of Black/Brown, Proportion of White and Sewerage Coverage.

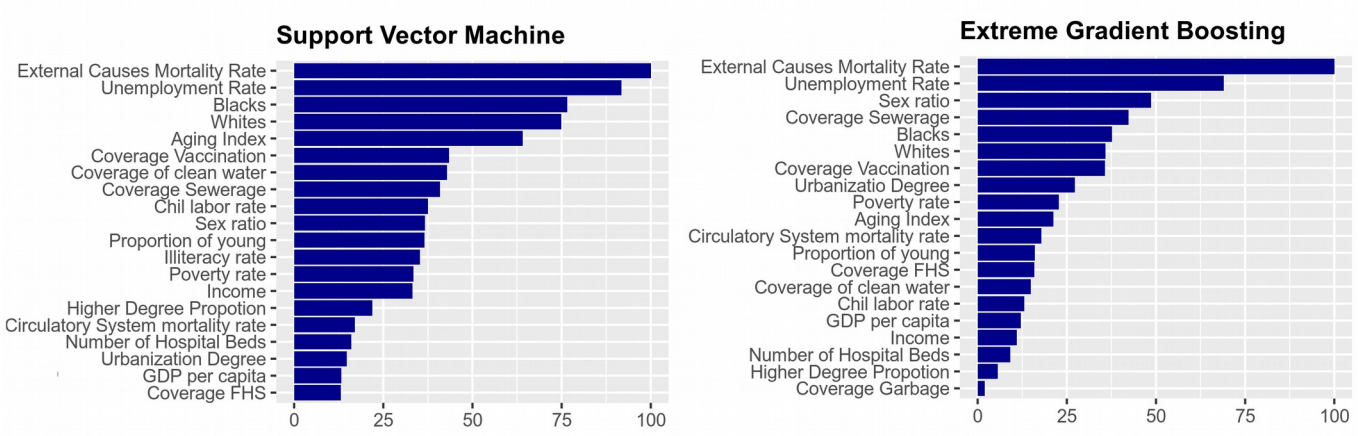
In relation to mortality Rate by External Causes, some literature point out that homicide is the main cause of deaths in the age group of 15 to 29 years and 60% of deaths due to external causes. The age range from 30 to 44 years and from 45 to 59 years begins to have the highest proportion of mortality due to traffic accidents (FERREIRA & ARAÚJO, 2006)

Others stusdies point out that the unemployment rate reveals an association between disadvantage in the labor market, health and mortality (IVERSEN et al, MOSER, 1987; QUEIROZ et al, 2017). Unemployment among the young and transition to adult life, from 18 to 24 years, is related to higher risks of general mortality, homicides and all causes of death (DAVILA et al, 2010)

The variables of the proportion of blacks and proportion of whites show that “Blacks and whites occupy unequal places in society and bring with them unequal experiences at birth, living and dying" (FIORIO et al, 2011, p.529).

The relationship of the place where one lives and the health condition of the population is directly related to the social and material deprivation of the less favored population leading to worse health conditions and higher mortality and this is associated with the sewerage coverage (SANTOS & NORONHA, 2001).

Figure 1 - Comparison of feature importance between SVM and XGBoost



Source: IBGE(2010), Datasus (2010)

References

LIMA-COSTA, M. F.; MATOS, D. L.; CAMARANO, A. A. Evolução das desigualdades sociais em saúde entre idosos e adultos brasileiros: um estudo baseado na Pesquisa Nacional por Amostra de Domicílios (PNAD 1998, 2003). Ciência & Saúde Coletiva, Rio de Janeiro, RJ, v. 11, n. 4, p. 941-950, 2006.

MOURA, E. C. et al. Mortality in Brazil according to gender perspective, years 2000 and 2010. Revista Brasileira de Epidemiologia, São Paulo, SP, v. 19, n. 2, p. 326-338, jun. 2016

WALQUE, D.; FILMER, D. Trends and socioeconomic gradients in adult mortality around the developing world. Population and Development Review, New York, NY, v. 39, n. 1, p. 1-29, 2013.

VASCONCELLOS, A. M. N.; FRANÇA, E. Measuring adult mortality in Brazil: improving quality of cause of death data. In: CHAIRE QUETELET, ADULT MORTALITY AND MORBIDITY, 2012, Louvain-la-Neuve, Bélgica. 2012.

QUEIROZ, B. L. et al. Adult mortality differentials and regional development at the local level in Brazil, 1980-2010. In: ANNUAL MEETING OF THE POPULATION ASSOCIATION OF AMERICA, 2017, Chicago. Anais... PAA: [S. l.], 2017

RICHARDSON, S. et al. Interpreting posterior relative risk estimates in disease-mapping studies. Environmental Health Perspectives, US, v. 112, n. 9, p. 1016-1025, 2004

SCHMERTMANN, C. P.; GONZAGA, M. R. Bayesian estimation of age-specific mortality and life expectancy for small areas with defective vital records. Demography, New York, NY, v. 55, n. 4, p. 1363-1388, 2018.

KUHN, M.; JOHNSON, K. Applied predictive modeling. New York, NY: Springer, 2013.

JAMES, G. et al. An introduction to statistical learning with applications in R. 8. ed. New York, NY: Springer, 2017.

LUSTGARTEN, J. L. et al. Improving classification performance with discretization on biomedical datasets. In: ANNUAL SYMPOSIUM, 11., 2008, Washington, DC. Proceedin s. Washington, DC: American Medical Informatics Association, 2008

BILLARI, F. C.; FÜRNKRANZ, J.; PRSKAWETZ, A. Timing, sequencing, and quantum of life course events: a machine learning approach. European Journal Of Population, Amsterdam, v. 22, n. 1, p. 37-65, 2006.

FERREIRA, H.; ARAÚJO, H. E. Transições negadas: homicídios entre os jovens brasileiros. In: CAMARANO, A. (org.). Transição para a vida adulta ou vida adulta em transição? Rio de Janeiro, RJ: IPEA, 2006. p. 291-318.

IVERSEN, L. et al. Unemployment and mortality in Denmark, 1970-80. British Medical Journal, London, v. 295, n. 6603, p. 879-884, 1987.

MOSER, K. A. et al. Unemployment and mortality: comparison of the 1971 and 1981 longitudinal study census samples. British Medical Journal, London, v. 294, n. 6564, p. 86-90, jan. 1987

DAVILA, E. P. et al. Young adults, mortality, and employment. Journal of Occupational and Environmental Medicine, Baltimore, v. 52, n. 5, p. 501-510, 2010.

FIORIO, N. M. et al. Mortalidade por raça/cor: evidências de desigualdades sociais em Vitória (ES), Brasil. Revista Brasileira de Epidemiologia, São Paulo, SP, v. 14, n. 3, p. 522-530, 2011.

SANTOS, S. M.; NORONHA, C. P. Padrões espaciais de mortalidade e diferenciais sócioeconômicos na cidade do Rio de Janeiro. Cadernos de Saúde Pública, Rio de Janeiro, RJ, v.17, n. 5, p. 1099-1110, 2001.