UNIVERSIDADE ESTADUAL DE CAMPINAS
INSTITUTO DE FILOSOFIA E CIÊNCIAS HUMANAS

NATÁLIA MARTINS ARRUDA

DETERMINANTES DA MORTALIDADE ADULTA NAS MICRORREGIÕES BRASILEIRAS EM 2010: UMA ANÁLISE BASEADA EM MODELOS DE APRENDIZADO DE MÁQUINA

CAMPINAS
2019

NATÁLIA MARTINS ARRUDA

DETERMINANTES DA MORTALIDADE ADULTA NAS MICRORREGIÕES BRASILEIRAS EM 2010: UMA ANÁLISE BASEADA EM MODELOS DE APRENDIZADO DE MÁQUINA

Dissertação apresentada ao Programa de Pós-Graduação em Demografia da Universidade Estadual de Campinas como requisito para a obtenção do título de Mestra em Demografia.

Supervisor/Orientadora: Profª Drª Luciana Correia Alves.

ESTE EXEMPLAR CORRESPONDE À VERSÃO FINAL DA DISSERTAÇÃO DEFENDIDA PELA ALUNA NATÁLIA MARTINS ARRUDA, E ORIENTADA PELA PROFA. DRA. LUCIANA CORREIA ALVES.

CAMPINAS
2019

FICHA CATALOGRÁFICA

Agência(s) de fomento e nº(s) de processo(s): CAPES, 1696637
ORCID: https://orcid.org/0000-0002-8503-1652

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca do Instituto de Filosofia e Ciências Humanas
Paulo Roberto de Oliveira - CRB 8/6272

Ar69d

Arruda, Natália Martins, 1992-
Determinantes da mortalidade adulta nas microrregiões brasileiras em 2010 : uma análise baseada em modelos de aprendizado de máquina / Natália Martins Arruda. – Campinas, SP : [s.n.], 2019.

Orientador: Luciana Correia Alves.
Dissertação (mestrado) – Universidade Estadual de Campinas, Instituto de Filosofia e Ciências Humanas.

1. Mortalidade. 2. Adulto. 3. Aprendizado de máquina. 4. Fatores socioeconômicos. 5. Desigualdades em saúde. 6. Desigualdades sociais. I. Alves, Luciana Correia, 1975-. II. Universidade Estadual de Campinas. Instituto de Filosofia e Ciências Humanas. III. Título.

Informações para Biblioteca Digital

Título em outro idioma: Determinants of adult mortality in the brazilian microregions in 2010 : an analysis based on machine learning models
Palavras-chave em inglês:
Mortality
Adult
Machine learning
Socioeconomic factors
Health inequalities
Social inequalities
Área de concentração: Demografia
Titulação: Mestra em Demografia
Banca examinadora:
Luciana Correia Alves [Orientador]
Tiago José de Carvalho
Marcos Roberto Gonzaga
Data de defesa: 18-03-2019
Programa de Pós-Graduação: Demografia

UNIVERSIDADE ESTADUAL DE CAMPINAS
INSTITUTO DE FILOSOFIA E CIÊNCIAS HUMANAS

A comissão julgadora dos trabalhos de Defesa de Dissertação de Mestrado, composta pelos professores Doutores a seguir descritos, em sessão pública realizada em 18 de março de 2019, considerou a candidata NATÁLIA MARTINS ARRUDA aprovada.

Profª Drª Luciana Correia Alves (Orientadora) – Universidade Estadual de Campinas.

Prof Dr. Tiago José de Carvalho – Instituto Federal de Educação, Ciência e Tecnologia de São Paulo.

Prof. Dr. Marcos Roberto Gonzaga – Universidade Federal do Rio Grande do Norte.

A Ata de Defesa com as respectivas assinaturas dos membros encontra-se no SIGA/Sistema de Fluxo de Dissertações/Teses e na Secretaria do Programa de Pós-Graduação em Demografia do Instituto de Filosofia e Ciências Humanas

Dedico aos meus pais e a minha irmã que sempre me apoiaram e me ensinaram a lutar pelo que eu acredito e aos meus avós (In Memorian) que se foram, em 2018, em meio ao processo deste trabalho.

AGRADECIMENTOS

Mais um ciclo se encerra e todas as experiências e conhecimentos adquiridos pedem alguns agradecimentos.

Em primeiro lugar, gostaria de agradecer minha orientadora Luciana, por ter me guiado, ajudado, mostrado caminhos que nem eu sabia que era capaz de ir e acreditado em mim. Muito obrigada por todas as oportunidades e portas que se abriram graças à você. Serei eternamente grata. Obrigada pela disponibilidade e ajuda em todos os momentos durante o processo de construção da dissertação e por ter contribuído de forma vital para minha formação como demógrafa.

Em segundo, gostaria de agradecer todos os professores do Programa de Pós-Graduação em Demografia pelos ensinamentos e pela dedicação na formação dos demógrafos. À todos os funcionários do Núcleo de Estudo de População “Elza Berquó” e do Instituto de Filosofia e Ciências Humanas pela disponibilidade e ajuda sempre que foi necessário.

Agradeço ao professor Tiago Carvalho pelas sugestões, críticas e conselhos que ajudaram a construir o que apresento nessas páginas. Além disso, gostaria de agradecer o professor Marcos Gonzaga pela ajuda, disponibilidade e por ter aceitado o convite para fazer parte da banca de defesa.

Aos meus amigos e familiares, muito obrigada por todo o apoio, atenção, cumplicidade e carinho que recebi durante todo este processo de construção do mestrado, sem vocês o caminho teria sido tortuoso.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

RESUMO

A identificação das características domiciliares e regionais com influência no risco de mortalidade adulta pode ajudar no desenvolvimento de políticas e ações direcionadas a tratar esse problema de saúde e socioeconômico. Aumentos na expectativa de vida serão menos alcançados por maiores reduções somente na mortalidade infantil. Em todos os países, progressos futuros na expectativa de vida irão depender da mortalidade adulta em algum grau. O objetivo principal do presente estudo foi investigar as relações entre os fatores socioeconômicos, estruturais, contextuais e de saúde e a probabilidade de morte adulta nas microrregiões brasileiras no ano de 2010. As análises baseiam-se nos dados retirados do Censo Demográfico de 2010 e do Sistema de Informações de Mortalidade do DATASUS. Para tal, procedeu-se a correção dos subregistros dos óbitos nas microrregiões brasileiras por meio do método TOPALS com estimação bayesiana para mortalidade em pequenas áreas. Além disso, foi utilizado o método de aprendizado de máquina para estabelecer os determinantes da probabilidade de morte adulta. Métodos de aprendizado de máquina possuem um grande potencial para este tipo de análise, uma vez que permitem um melhor entendimento das interações entre os diferentes fatores. Os resultados encontrados mostraram que as variáveis taxas de mortalidade por causas externas, taxa de desemprego, proporção de negros, cobertura de vacinação e proporção de brancos foram as que obtiveram maior poder preditivo nas probabilidades de morte adulta usando os algoritmos *Random Forest, Extreme Boosted Trees, Support Vector Machine e Naive Bayes.* Os algoritmos obtiveram bom desempenho e mostraram ser eficazes ao analisar as variáveis, ainda que algumas correlacionadas, com o desfecho de probabilidade de morte adulta. Identificar os determinantes da mortalidade adulta e as principais disparidades entre grupos sociais e em pequenas áreas é de extrema importância no auxílio de construção de políticas públicas que visem responder adequadamente as necessidades específicas de cada região e grupo social, contribuindo para a redução das desigualdades socioeconômicas e mortalidade.

Palavras-chave: Mortalidade; Adulto; Aprendizado de máquina; Fatores socioeconômicos; Desigualdades.

ABSTRACT

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. The main objective of the present study was to investigate the relationships among socioeconomic, structural, contextual and health factors and the adult probability of death in the Brazilian *microregions* 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 in the Brazilian *microregions* 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 the adult probability of death. 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 odds of adult death 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 death 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.

Key-words: Mortality; Adult; Machine learning; Socioeconomic factors; Inequalities.

LISTA DE GRÁFICOS

LISTA DE TABELAS

LISTA DE FIGURAS

SUMÁRIO

INTRODUÇÃO

Mudanças no nível de renda ou educação, no grau de acesso aos serviços públicos, ou qualquer transformação na organização socioeconômica e política, possui influências significativas no sentido de melhorar ou piorar a qualidade de vida de um grupo em relação ao outro, modificando o padrão dos diferenciais, por estratos sociais, da mortalidade (WOOD; CARVALHO, 1994).

A esperança de vida ao nascer (uma medida da duração da vida de uma população) aumentou em várias décadas em muitas nações desenvolvidas no século XX, com notáveis aumentos que também ocorreram na maioria dos países em desenvolvimento nas últimas décadas (VAUPEL, 2010).

A queda das taxas de mortalidade ocorreu primeiramente nas primeiras idades. O principal motivo para os ganhos expressivos de anos adicionais na esperança de vida neste período foi a queda da mortalidade nos primeiros anos de vida. No Brasil, a mortalidade infantil reduziu consideravelmente entre 1980 e 2000. Em 1980, esta taxa correspondia a 69,12 óbitos para cada mil nascidos vivos, e em 2009 passou para 22,47 para cada mil nascidos vivos (IBGE, 2010a). Da mesma forma, os ganhos na esperança de vida também foram substanciais. Em 1980, a esperança de vida alcançou, para ambos os sexos, 62,57 anos e, em 2009, 73,17 anos (IBGE, 2010a).

O declínio da mortalidade infantil ocorreu, principalmente, devido à uma redução das doenças infecciosas, em virtude de melhorias estruturais como, por exemplo, oferta de água encanada e sistema de esgoto (PRATA, 1992).

Na outra ponta dos grupos etários, após 1980, no Brasil, houve uma diminuição nas taxas de mortalidade entre os indivíduos com mais de 60 anos, especialmente para os idosos com mais de 80 anos, fator que contribuiu com uma maior longevidade e para ganhos adicionais na esperança de vida (CAMPOS; RODRIGUES, 2004).

Paralelamente à transição da mortalidade, deu-se a transição epidemiológica que tem como foco principal entender as causas por trás das mudanças da predominância de doenças transmissíveis para não transmissíveis. Neste sentido, o primeiro período da transição epidemiológica foi causado por uma combinação de avanços na saúde pública e inovações tecnológicas e comportamentais (OMRAN, 1971). Os exemplos incluem reduções nas doenças transmissíveis, disponibilidade de água potável, encanamento interno, antibióticos e vacinas (OMRAN, 1971).

Ademais, observou-se além do processo de declínio da mortalidade, a sua compressão, ou seja, a distribuição etária das mortes move-se em direção as idades mais avançadas concentrando-se nessas idades (FRIES, 1980; GONZAGA; QUEIROZ; LIMA, 2018). Conforme ocorre a diminuição da probabilidade de morte, maior número de pessoas sobrevive às idades adultas e avançadas, e a curva de sobrevivência se desloca para a direita (FRIES, 1980; GONZAGA; QUEIROZ; LIMA, 2018).

Diante deste panorama, tornou-se necessário entender a mortalidade a partir de sua natureza e de seus fatores determinantes, visto que importantes mudanças ocorrem continuamente na mortalidade nos diferentes grupos etários (AGOSTINHO; QUEIROZ, 2008).

As desigualdades socioeconômicas têm um papel importante para o entendimento do declínio acelerado da mortalidade nos países em desenvolvimento. Santos e Noronha (2001), por exemplo, mostraram que grupos mais favorecidos economicamente apresentaram baixa mortalidade por todas as causas de morte. O declínio da mortalidade é maior entre a população mais escolarizada, com melhores condições de moradia e sanitária (KUNITZ, 1987). Cutler; Deaton e Lleras-Muney (2006) afirmam que os indivíduos com baixa renda, baixa educação ou baixo *status* social frequentemente morrem mais jovens do que aqueles que são mais escolarizados e com maior renda.

O Brasil, nessa perspectiva, é um país marcado por desigualdades sociais e econômicas. Está entre os países com grau mais elevado de desigualdade social no mundo (LIMA-COSTA; MATOS; CAMARANO, 2006). Apesar das duas últimas décadas terem sido de mudanças sociais significativas com redução da desigualdade e da pobreza (ANDRADE et al, 2013). Os estados com maior desigualdade de renda, de acordo com o coeficiente de Gini (indicador que mede o grau de concentração de renda) estão localizadas na área urbana das regiões Norte e Nordeste do país (NORONHA; ANDRADE, 2007). No Brasil, entre 1990 e 2009, o coeficiente de Gini diminuiu de 0,61 para 0,54. Essa redução ocorreu em todas as regiões, sendo mais acentuadas na região Sul e Sudeste com 15% e 11,5% de queda, respectivamente. A região com menor queda foi a Centro-Oeste com 8,3%. Em relação à taxa de pobreza, em 1990, era igual a 41,92 e em 2009, 11,60. Apesar dessa queda, diferenças regionais significativas ainda estão presentes. Em 2009, as taxas de pobreza nas regiões Sul, Sudeste e Centro-Oeste estavam em torno de 12%, no Norte e Nordeste esse valor era de 32% e 40%, respectivamente (ANDRADE et al, 2003).

As desigualdades socioeconômicas encontradas quando se fala sobre mortalidade infantil foram amplamente investigadas no Brasil e no mundo (MOSLEY; CHEN, 1984; VICTORA et al., 2000; CASTRO; SIMÕES, 2009). Em contrapartida, os estudos que tratam de mortalidade adulta focaram em sua maioria na qualidade dos dados e cobertura e não em seus fatores de desigualdades (MOURA et al., 2016; WALQUE; WILMER, 2013; VASCONCELOS; FRANÇA, 2012). Em geral, estudos sobre determinantes de mortalidade adulta são pouco explorados no Brasil (QUEIROZ et al, 2017).

Queiroz et al. (2017) mostraram que a mortalidade adulta no nível das mesorregiões brasileiras convergiu entre as regiões do país no período entre 1980 e 2010. Ainda assim, há diferenças significativas entre as regiões com foco de alta mortalidade em algumas áreas. Além disso, no período estudado, redução dos níveis de mortalidade em áreas com baixo nível socioeconômico poderiam indicar melhorias nas condições de vida.

Espera-se uma diminuição geral dos riscos de mortalidade na infância e essa mortalidade é substituída por riscos mais elevados a partir dos 15 anos de idade em decorrência da entrada na vida adulta (SILVA et al., 2016). Portanto, compreender a mortalidade adulta, sua distribuição no espaço e a sua relação com características sociais e econômicas é essencial para o entendimento dos fatores contextuais e socioeconômicos, suas causas e consequências na geração de desigualdades em saúde no Brasil (RENTERÍA PÉREZ; TURRA, 2008; PEREIRA; QUEIROZ, 2016). Analisar o nível e padrão da mortalidade é de grande importância no planejamento de políticas públicas e tais análises só são possíveis com informações de boa qualidade como boa cobertura dos óbitos e das causas de morte. Queiroz et al. (2017) afirmam que no período de 1980 e 2010 houve uma grande melhora da cobertura de mortes ao analisar a qualidade dos dados em relação a mortalidade adulta.

Os níveis em pequenas áreas, definidas como microrregiões são considerados os melhores para se evitar viés criado pela heterogeneidade nos níveis de mortalidade e características socioeconômicas dos municípios e para detectar padrões geográficos de mortalidade que às vezes não ficam evidentes usando áreas maiores (RICHARDSON et al., 2004).

Portanto, procurar-se-á identificar os determinantes da mortalidade adulta nas microrregiões brasileiras e responder as seguintes questões: i) Quais são os fatores que mais caracterizam mudanças na probabilidade de morte adulta nas microrregiões brasileiras em 2010? ii) Será que decisões governamentais em nível da localidade são capazes de prover a estrutura e desenvolvimento necessário para manter uma qualidade de vida a fim de evitar a mortalidade nas idades entre 15 a 60 anos?

A identificação das características domiciliares e regionais com impacto no risco de mortalidade adulta pode ajudar no desenvolvimento de políticas e ações direcionadas a tratar esse problema de saúde e socioeconômico. Aumentos na esperança de vida serão menos adquiridos por maiores reduções somente na mortalidade infantil (QUEIROZ et al., 2017). Acredita-se que progressos futuros na esperança de vida irão depender da mortalidade adulta em algum grau (VALLIN; MESLÉ, 2004).

Diante do exposto, a hipótese da presente pesquisa é de que a mortalidade adulta nas microrregiões do Brasil é um fenômeno complexo, que envolve interações de diversas características e que necessita da combinação de um grande volume de atributos para o seu total entendimento. Sendo assim, é provável que a probabilidade de morte adulta seja uma combinação de fatores socioeconômicos, de saúde e de infraestrutura e não de fatores isolados. E microrregiões com um melhor conjunto de infraestrutura e com programas de políticas públicas bem estruturadas teriam uma mortalidade adulta mais baixa.

Para tanto, os modelos de regressão tradicionais podem não ser suficientes para compreensão destas características em conjunto, uma vez que os pressupostos da modelagem paramétrica são irrealistas para investigações de natureza mais ampla. Métodos de aprendizado de máquina possuem um grande potencial para este tipo de análise, uma vez que permitem um melhor entendimento das interações entre os diferentes fatores de forma conjunta. No entanto, são raramente utilizados em pesquisas demográficas e em estudos de mortalidade adulta no Brasil. Logo, a aplicação de técnicas de aprendizado de máquina a este contexto é inovadora à realidade brasileira. Com modelos que tornem tais análises mais eficientes e eficazes, além de prevenir mortes adultas, pretende-se também melhorar a tomada de decisões para criação de políticas públicas voltadas para melhorar as condições de vida deste grupo populacional.

O objetivo principal do presente estudo será investigar as relações entre os fatores socioeconômicos, estruturais, contextuais e de saúde e a probabilidade de morte adulta nas microrregiões brasileiras no ano de 2010. Além disso, pretende-se discutir diferentes algoritmos de aprendizado de máquina a fim de entender àqueles com melhor performance e que melhor atingiram o objetivo principal.

Essa dissertação será composta por três capítulos, além desta introdução. O primeiro deles discorre sobre as transformações ocorridas na mortalidade, em que um declínio acentuado no nível da mortalidade foi acompanhado pela diferenciação das causas e formação um padrão epidemiológico. Adicionalmente, mostrou-se um panorama geral das desigualdades socioeconômicas no Brasil e suas principais tendências e, por último, se

discorre sobre os principais fatores associados à mortalidade em geral e, em particular, o que se tem sobre mortalidade adulta.

No segundo capítulo, apresentam-se as variáveis utilizadas, sua distribuição e estatísticas básicas, o método de correção dos óbitos escolhido para o cálculo da probabilidade de morte adulta, a explicação do funcionamento dos algoritmos *Random Forest, Extreme Gradient Boosted Trees, Naive Bayes* e *Support Vector Machine* e do processo necessário para testá-los de forma a conseguir a melhor performance dos mesmos.

No terceiro capítulo se apresenta os principais resultados, performances dos algoritmos e, por fim, se discute à luz das principais tendências na mortalidade e epidemiológicas os resultados encontrados, além de comparações entre os algoritmos testados no presente estudo. Por último, obtêm-se as principais conclusões tanto em relação ao método utilizado quanto em relação aos principais resultados encontrados para definir os fatores que mais se relacionam à mortalidade adulta.

CAPÍTULO 1 – TRANSIÇÃO DA MORTALIDADE E FATORES DETERMINANTES

O declínio quase universal nas taxas de mortalidade registradas em diferentes épocas no mundo levou os demógrafos a criar o termo “transição da mortalidade” para se referir à passagem da alta mortalidade, associada em grande parte à prevalência de doenças infecciosas e parasitárias para baixas taxas de mortalidade, decorrente do controle de tais doenças transmissíveis (OMRAN, 1971; ELO, 2007; AKSAN; CHACKRABORTY, 2014).

A transição da mortalidade foi paralelamente acompanhada pela transição epidemiológica. Como a distribuição dos óbitos mudou tanto em termos de causa quanto de estrutura etária, a esperança de vida ao nascer passou de cerca de 40 para cerca de 60 anos devido às mudanças na predominância das diferentes causas de morte. Nos estágios iniciais da transição de alta para a baixa mortalidade, enfatizou-se especialmente o declínio do papel das doenças transmissíveis e a crescente participação da mortalidade por doenças não transmissíveis, principalmente doença cardiovascular e câncer. Nos últimos anos, a análise da experiência de diferentes grupos de países, sob a perspectiva da transição epidemiológica, revelou que novos declínios foram possíveis devido à redução da mortalidade causada por doenças cardiovasculares (OMRAN, 1971; AKSAN; CHACKRABORTY, 2014).

Reconhecendo a importância das desigualdades na mortalidade surgiu o termo transição de saúde com o objetivo de chamar a atenção para os fatores sociais e comportamentais subjacentes à mortalidade e às transições epidemiológicas (LERNER, 1972; ELO, 2007).

1.1 Transição da mortalidade

O declínio da mortalidade é uma das maiores conquistas da civilização. Há cem anos na Europa, a expectativa de vida era por volta de 40 anos. Aqueles que sobreviviam até os cinco anos de idade tinham uma expectativa de vida de 50 anos. Atualmente, na Europa, a expectativa de vida atingiu a idade de 79 anos em média, 76 anos para os homens e 82 para as mulheres. O declínio da mortalidade começou por volta de 1800 na França e no final do século XIX nos outros países europeus. Esse declínio ocorreu seguindo padrões particulares (OEPPEN; VAUPEL, 2002; WILLEKENS, 2014).

O aumento da esperança de vida foi estável, isto é, desde 1840, aumentou, em média, cerca de 2,5 anos a cada década para mulheres e 2,2 anos para homens nos países mais desenvolvidos (OEPPEN; VAUPEL, 2002; WILLEKENS, 2014).

Embora os dados de mortalidade em nível nacional não tenham sido disponibilizados para os Estados Unidos até 1933, as evidências existentes sugerem que o declínio da mortalidade apresentado pelo país foi semelhante ao da Inglaterra. A maior esperança de vida foi registrada no Japão, um país desenvolvido onde as melhorias na saúde no início do século XX ficaram para trás em relação aos países europeus, mas onde os declínios de mortalidade têm sido particularmente impressionantes desde os anos 50 (ELO, 2007, OEPPEN; VAUPEL, 2002).

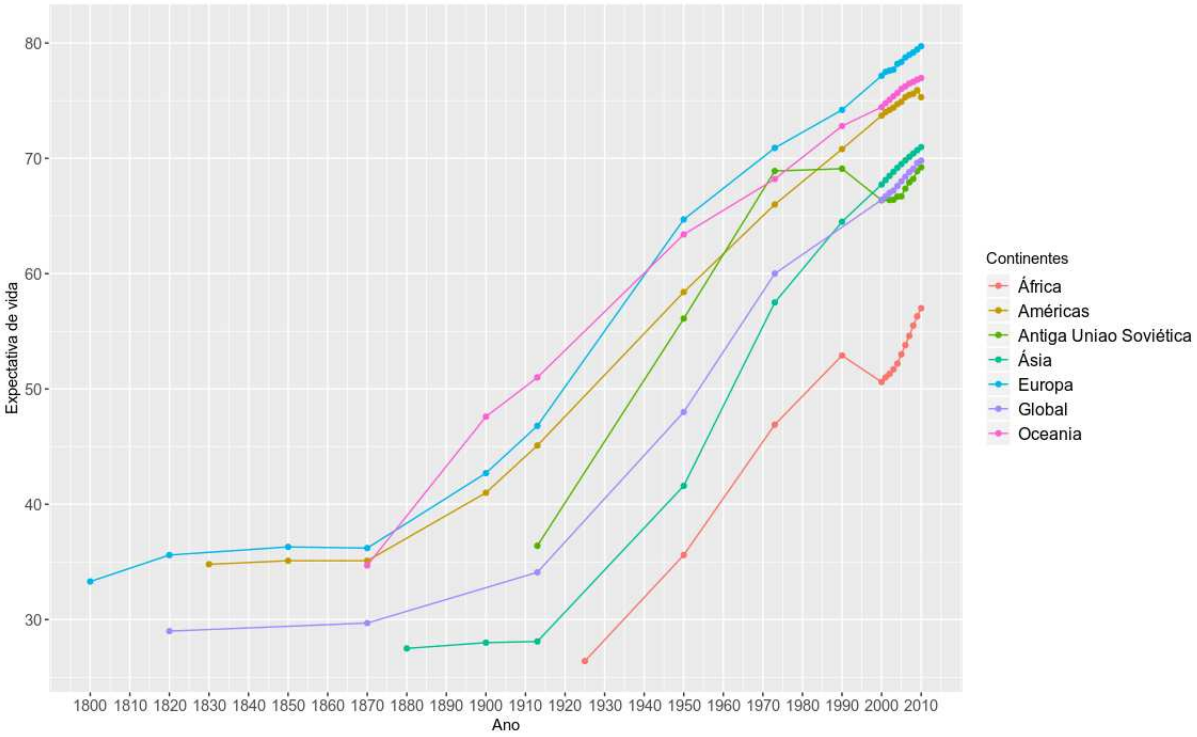
A esperança de vida ao nascer no Japão alcançou 84,6 anos para mulheres e 77,6 anos para homens até o ano 2000. Além disso, as estimativas das Nações Unidas mostram uma expectativa média de vida de 74,8 anos nas regiões mais desenvolvidas do mundo, com 56% dos países industrializados com esperanças de vida de mais de 75 anos no período entre 1995 e 2000. As menores esperanças de vida dentre os países industrializados são encontradas na Europa Oriental e na antiga União Soviética, onde as condições de saúde estagnaram durante o final do século XX, particularmente para homens adultos (ELO, 2007).

O declínio da mortalidade nos países em desenvolvimento ocorreu em tempos e ritmos diferentes se comparado com os países desenvolvidos uma vez que a natureza do declínio da mortalidade nos países da América Latina (AL) foi diferente do declínio da mortalidade nas sociedades mais desenvolvidas (Gráfico 1). A transição da mortalidade nesses países possui algumas peculiaridades como seu caráter comprimido e a evolução de um regime em que condições crônicas coexistem com doenças infecciosas (PALLONI; PINTO-AGUIRRE, 2011).

Em 1900, os países da América Latina apresentavam um nível surpreendente alto de mortalidade. Países como Bolívia, Brasil, Chile, Paraguai e Costa Rica apresentavam expectativas de vida ao nascer entre 24 a 30 anos. O declínio foi lento até 1930 e na década posterior registrou as maiores taxas de melhorias nas taxas mortalidade (ARRIAGA; DAVIS, 1969).

Arriaga e Davis (1969) dividem os países da América Latina em dois grupos, os que tiveram um declínio da mortalidade e atingiram uma expectativa de vida entre 25 e 35 anos antes de 1930 e os que só atingiram este nível de expectativa de vida de 30 a 35 anos após 1930. No primeiro grupo estão os seguintes países: Brasil, Chile, Colômbia, Costa Rica e Panamá e no segundo grupo estão: República Dominicana, Guatemala e Nicarágua. Após 1930, estes países começam a se mover para os níveis de expectativa de vida entre 40 e 50 anos e a taxa de declínio da mortalidade começou a ser quase a mesma.

GRÁFICO 1 – Evolução da Esperança de Vida segundo continentes – 1800 a 2010



Fonte: IISH (1800-1949); ONU (1950-2010).

O declínio da mortalidade na América Latina após 1930 ocorreu devido, principalmente, à queda de mortalidade infantil e medidas de saúde pública proporcionando ganhos na esperança de vida ao nascer (ARRIAGA; DAVIS, 1969).

De acordo Arriaga e Davis (1969), o declínio da mortalidade estava relacionado à situação econômica do país, pois os países do primeiro grupo eram mais economicamente desenvolvidos nesta época comparativamente aos países do segundo grupo. Já após 1930, a relação é quase independente do desenvolvimento econômico, ou seja, neste caso, outros fatores passaram a afetar o declínio da mortalidade. Isto é, as medidas de saúde pública exerceram uma forte influência nas taxas de mortalidade independentemente do desenvolvimento econômico após esse período (ARRIAGA; DAVIS, 1969; SOARES, 2007).

Em outras palavras, aumentos na expectativa de vida após 1930 são atribuídos à importação de medidas de saúde pública dos países desenvolvidos, incluindo infraestrutura de grande escala (sistemas de purificação de água, água encanada, sistema de esgoto) e campanhas financiadas internacionalmente para erradicar doenças transmitidas por vetores (por exemplo, malária, dengue e febre amarela) (PALLONI; PINTO-AGUIRRE, 2011). Em relação à mortalidade infantil, em 1970, Cuba apresentava uma taxa de 36 por mil nascidos

vivos, na Argentina 59 por mil, Costa Rica 62 por mil e 79 por mil no Chile (MARTINE; CARVALHO; ALFONSO, 1994).

Assim como ocorreu nos países da América Latina em geral, no Brasil, houve um ganho em anos de vida expressivo após a década de 1930. A redução dos níveis de mortalidade foi bem mais rápida do que aquela experimentada anteriormente pelas populações europeias. Como dito anteriormente, o declínio da mortalidade nos países desenvolvidos ocorreu de forma gradual durante séculos. Por sua vez, no Brasil a importação de tecnologias médicas permitiu avanços mais rápidos e em um curto período de tempo (MARTINE; CARVALHO; ALFONSO, 1994).

Durante as primeiras décadas do século XX houve pouca oscilação nas taxas de mortalidade no Brasil. Estimativas revelaram que o coeficiente de mortalidade declinou pouco, variando de 29,1 por mil habitantes em 1900 para 24,4 por mil em 1940. Reflexo disso foi à mudança inexpressiva da esperança de vida no período, passando de 33 anos para 37 para homens e de 34 para 39 anos no caso das mulheres (CHAIMOWICZ, 1997).

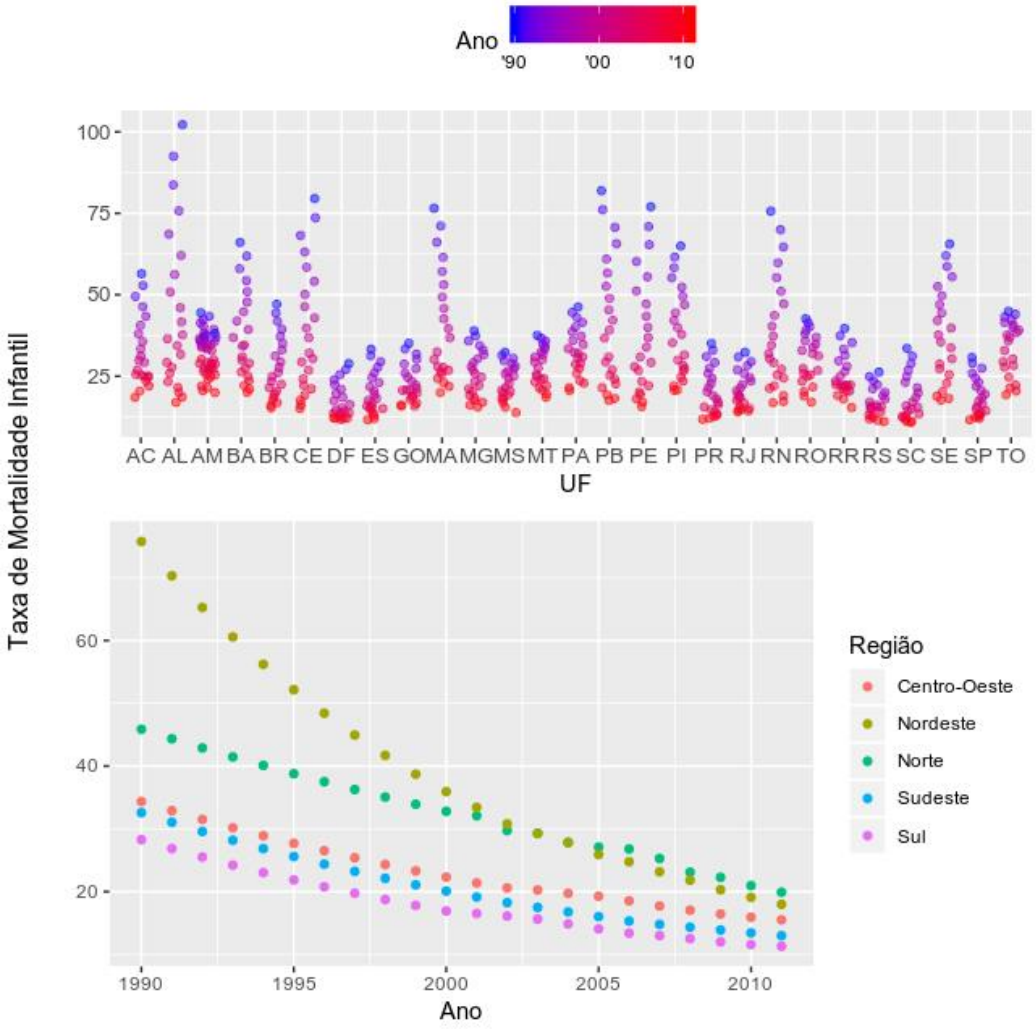
No período entre 1940 e 1970 se observou um rápido declínio no Brasil. Na década de 1940, a taxa de mortalidade caiu 13%, enquanto que nas quatro décadas anteriores houve o declínio de somente 16%. Somente na década de 1940, a esperança de vida aumentou 4 anos para os homens e, aproximadamente, 7 anos para mulheres, alcançando quase 54 anos em 1970 (CHAIMOWICZ, 1997). Esses ganhos foram desiguais entre as regiões: o Nordeste apresentou o maior aumento relativo na expectativa de vida no período, porém permaneceu com um dos níveis mais altos de mortalidade. Foram necessários 30 anos para o Nordeste atingir em 1970 o nível médio de mortalidade que o Brasil atingiu em 1940 (MARTINE; CARVALHO; ALFONSO, 1994).

A mortalidade infantil decresceu de 160 por mil nascidos vivos em 1940 para 85 por mil em 1980 (PRATA, 1992). Ademais, após 1970, manteve-se a tendência de declínio da mortalidade em todo o país, com um ganho médio de seis anos na esperança de vida entre 1970 e 1980 e apesar das desigualdades regionais, notou-se uma tendência de convergência dos níveis de mortalidade regionalmente (MARTINE; CARVALHO; ALFONSO, 1994). Em 1980, a esperança de vida ultrapassou os 60 anos de idade.

A maior contribuição para os ganhos em esperança de vida se deu devido a queda da mortalidade infantil. A Figura 1 ilustra as principais tendências e mostra que a mortalidade infantil caiu pela metade atingindo 45 óbitos por mil nascidos vivos e a esperança de vida atingiu os quase 66 anos. Entre 1991 e 2010, a taxa de mortalidade infantil caiu para 16,2

óbitos por mil nascidos vivos e a expectativa de vida atingiu 73 anos em 2010 (VASCONCELOS; GOMES, 2012).

FIGURA 1 – Mortalidade infantil por mil. UF, Grandes Regiões – Brasil, 1990-2010



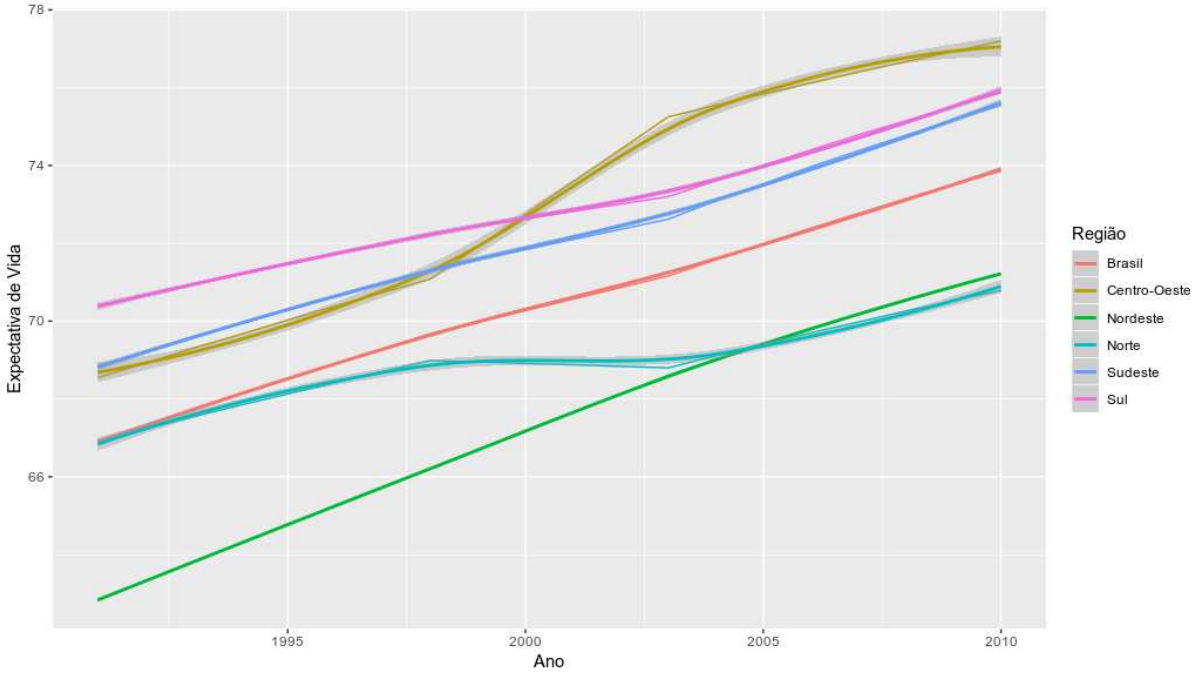
Fonte: IBGE (2010a); Brasil (2010).

Essas transformações não ocorreram simultaneamente em todas as regiões (Gráfico 2):

> Em 1970, os indicadores de mortalidade para as regiões Sudeste, Sul e Centro-Oeste evidenciavam um processo de transição já iniciado, enquanto nas regiões Norte e Nordeste, os elevados níveis de mortalidade infantil caracterizavam um momento de pré-transição. Após 1970, todas as regiões encontravam-se em processo de transição, a queda dos níveis de mortalidade, especialmente da mortalidade infantil, e o consequente aumento da esperança de vida eram generalizados. Todas as Regiões, à exceção do Nordeste, alcançavam ou superavam os 60 anos de vida média, que apenas a região Sul havia atingido na década anterior (VASCONCELOS; GOMES, p. 545, 2012).

No nível municipal apesar da tendência de queda de 1991, 2000 e 2010, alguns municípios apresentaram taxas altas de mortalidade infantil em relação à média brasileira, assim como alguns municípios apresentaram baixas esperanças de vida se comparado à média do país. Por exemplo, o município de Fátima, pertencente ao estado da Bahia, em 1991 apresentava uma taxa de mortalidade infantil de 120 óbitos por mil nascidos vivos, sendo muito acima da média brasileira e continuou uma das mais altas em 2010 com 40,7 óbitos por mil nascidos vivos.

GRÁFICO 2 – Evolução da Expectativa de Vida – Brasil e Grandes Regiões



Fonte: IBGE (2010a); Brasil (2010).

As mais baixas taxas de mortalidade infantil e mais altas esperanças de vida se encontram nos municípios da região do Sul, por exemplo, Santa Catarina (SC) e Rio Grande de Sul (RS). O município de Antônio Carlos (SC) em 1991 apresentava uma expectativa de vida de 72,5 anos e em 2010 chegou a 78,4 anos muito acima da média brasileira já em 1991. A mesma tendência e nível ocorrem em Joinville (SC) e Taquara (RS). Em relação à mortalidade infantil, esses municípios possuem taxas mais baixas que a média brasileira. Observou-se em Antônio Carlos, em 1991, uma taxa de 18,2 óbitos por mil nascidos vivos, alcançando em 2010 uma taxa de 9,3 óbitos por mil nascidos vivos. Outro exemplo está no município de Taquara, com uma taxa de mortalidade infantil de 14,6 óbitos por mil nascidos

vivos em 1991 (já abaixo da média brasileira em 2010), alcançando 9,1 óbitos por mil nascidos vivos.

Nesse processo de transição da mortalidade no Brasil, pode-se destacar a redução da mortalidade por doenças infecciosas e parasitárias e o aumento da importância de doenças crônico-degenerativas, que tem muitos dos seus fatores de risco associados às condições de vida urbanas, como hábitos alimentares, vida sedentária e estresse. Além disso, o maior acesso da população a rede de esgoto e água, maior atenção à saúde e campanhas de vacinação contribuíram de forma expressiva na redução dos níveis de mortalidade no país e no aumento da expectativa de vida no período apresentado.

Essa redução da predominância das doenças infecciosas e um aumento das doenças crônico-degenerativas ficou conhecida como transição epidemiológica. O entendimento desta transição é fundamental para se compreender as tendências na mortalidade e a mudança no padrão etário da mesma (ELO, 2007; OMRAN, 1971).

1.2 Transição epidemiológica

A queda nas taxas de mortalidade por doenças infecciosas levou a melhorias significativas nas chances de sobrevivência dos nascidos vivos e crianças sendo amplamente responsável pelo aumento da expectativa de vida no final do século XIX e início do século XX nos países industrializados (ELO, 2007). Ou seja, a queda da mortalidade está relacionada às mudanças no perfil das principais causas de morte e na estrutura etária da mortalidade (OMRAN, 1971).

Omran (1971) foi o primeiro a definir os principais estágios da transição epidemiológica levando em consideração diferentes padrões econômicos e de transição demográfica apresentada pelos países. São eles: a era da peste e da fome, em que a mortalidade foi alta e flutuante e a esperança de vida ficava entre 20 e 40 anos; o segundo estágio refere-se à era da regressão das pandemias, em que a taxa de mortalidade é menos variável e cai progressivamente, sendo que a esperança de vida aumentou para patamares entre 30 e 50 anos; o terceiro estágio diz respeito às doenças degenerativas e causadas pelo homem, em que a taxa de mortalidade continua a cair estabilizando em um nível baixo e a esperança de vida ultrapassa os 50 anos.

De acordo com Omran (1971), os ganhos nos níveis de saúde dos países desenvolvidos estavam associados às transformações sociais ligadas à Revolução Industrial, decorrente de melhorias na oferta de alimentos, condições de habitação e em medidas de saneamento básico. Sendo assim, as mudanças nos padrões de saúde e doença que

caracterizaram a transição epidemiológica estão estreitamente relacionadas à transição demográfica e socioeconômica que constituem o processo de modernização.

As primeiras mudanças nos padrões saúde-doença ocorreram com maior intensidade nas crianças e mulheres jovens, e para Omran (1971), as variações no padrão, ritmo, determinantes e consequências, se diferem em três modelos de transição epidemiológica. O primeiro deles é o modelo clássico, ou seja, uma transição gradual e progressiva de alta mortalidade e fecundidade que acompanharam o processo de modernização em alguns países da Europa Ocidental. O segundo modelo, o acelerado, aconteceu no Japão, com a transição da mortalidade ocorrendo de forma acelerada, e a mortalidade flutuante da Era da peste e fome e fase da regressão das pandemias seguindo um padrão parecido com o primeiro modelo, porém tardiamente. O terceiro modelo, o contemporâneo ou atrasado, refere-se à transição relativamente recente ou em curso nos países em desenvolvimento e apesar dos ganhos de sobrevida na mulher e na criança, a mortalidade infantil permanece alta.

Olshansky e Ault (1986), analisando as tendências nos padrões de morbidade e mortalidade nos Estados Unidos, sugeriram o início de um quarto estágio, complementando o quadro exposto por Omran (1971). O quarto estágio ficou definido como sendo "A Era do Retardamento das Doenças Degenerativas", em que há um declínio da mortalidade nas idades mais avançadas consequência de um retardamento nas mortes causadas pelas doenças crônicas e degenerativas.

Horiuchi (1999) apresenta um quadro em que mudanças históricas do padrão de mortalidade no passado e no futuro são sintetizadas em cinco transições epidemiológicas. Das lesões externas para as doenças infecciosas; das doenças infecciosas para as doenças degenerativas e mais recentemente, o declínio da mortalidade por doenças cardiovasculares. Para o futuro, o autor afirma que se podem esperar mais duas outras transições, o declínio da mortalidade por câncer e a desaceleração da senescência.

O declínio da mortalidade devido a segunda e terceira transição definidas por Horiuchi (1999) mudou a distribuição etária de mortes para idades mais avançadas. O padrão do declínio por idade difere marcadamente entre essas duas transições. A tendência da esperança de vida entre as idades de 10 e 60 anos foi diferente da encontrada entre 0 a 10 anos.

O diferencial de tendências da esperança de vida, usando como exemplo a Suécia, sugere que nos estágios iniciais da segunda transição, a redução da mortalidade foi pronunciada para as crianças, mas não necessariamente para os jovens e jovens adultos; nos estágios finais da segunda transição, o nível de mortalidade decaiu substancialmente entre os

adultos em idade reprodutiva, principalmente devido à redução da mortalidade por tuberculose e na terceira transição, um marcado declínio da mortalidade foi observado entre os idosos principalmente devido à redução da mortalidade por doenças cardiovasculares, apesar de uma notável diferença entre os sexos (HORIUCHI, 1999).

Alguns autores (CASELLI; MESLE; VALLIN, 2002) afirmam que nem todas as sociedades experimentaram as três eras no processo de modernização sugerida por Omran (1971). Adicionalmente, para eles, Omran (1971) não conseguiu prever que o surgimento de doenças causadas pelo homem seria contido por políticas eficientes, e especialmente, que uma revolução ocorreria no tratamento de doenças cardiovasculares.

Nos países em desenvolvimento, em que o início das tentativas de desenvolvimento econômico ocorreu tardiamente se comparado com os países desenvolvidos que se beneficiaram positivamente da Revolução Industrial, verificaram-se mudanças importantes no padrão de morbimortalidade. Para os países menos desenvolvidos, a transição epidemiológica não só ocorreu mais tarde como está seguindo uma rota diferente. Esses países em desenvolvimento que alcançaram alta expectativa de vida são caracterizados pela relativamente alta mortalidade infantil e baixa mortalidade adulta quando comparados com os países desenvolvidos (VALLIN; MESLÉ, 2004).

A qualidade de vida melhorou de forma irregular na população brasileira. O declínio das doenças infecciosas ocorreu lentamente e só se acentuou após o surgimento dos antibióticos e outros avanços em tecnologias na saúde (ARAÚJO, 2012). Assim como nos países em desenvolvimento, as doenças crônico-degenerativas como doenças cardiovasculares, neoplasias e causas externas passaram a ter maior predominância como causas de morte. Porém, apesar disso, e diferentemente do que ocorreu nos países industrializados, persiste ainda nos países em desenvolvimento, taxas altas de morbidade e mortalidade devido a doenças infecciosas e parasitárias (ARAÚJO, 2012).

Seguindo esse ponto de vista, Frenk; Lozano e Bobadilla (1994) estudou as particularidades dos países da América Latina e defenderam a existência de um modelo polarizado de transição epidemiológica latino-americana, caracterizado por sobreposição de etapas (incidência alta e concomitante das doenças pré e pós transição); contra-transição (ressurgimento de algumas doenças infecciosas que já haviam sido controladas); casos de transição prolongada (processos de transição que ainda não foram concluído) e polarização epidemiológica (níveis diferenciados de transição e entre grupos sociais dentro do mesmo país).

Os níveis de mortalidade são menores, porém sua composição por causas de morte é mais complexa do que nos países centrais. Em relação à estrutura da mortalidade por idade, todos os países da América Latina passaram por mudanças, ou seja, no começo da transição epidemiológica, a maioria das mortes correspondeu às crianças menores de 15 anos. Durante a transição, a maior parte das mortes se move para os grupos etários mais velhos. Este processo ocorre, pois, as doenças infecciosas e parasitárias e os problemas de má nutrição afetam principalmente as crianças, sendo assim, com o combate dessas doenças a sobrevivência dessas aumenta mais rapidamente que dos adultos e idosos (FRENK; LOZANO; BOBADILLA, 1994).

Os mecanismos mais importantes que interferem na transição epidemiológica na América Latina são as mudanças nos fatores de risco que atuam principalmente sobre a probabilidade de contrair uma doença, ou seja, nas taxas de incidência; e as melhorias na tecnologia e atenção à saúde que modificam as taxas de letalidade. As mudanças nos fatores de risco se relacionam com o modelo de desenvolvimento incluindo transformações das sociedades em que se predomina a produção agrícola para sociedades com predomínio de serviços e mais industrializada; mudanças na distribuição da população, de áreas rurais para áreas urbanas, com a consequente concentração econômica, levando a um aumento da desigualdade e empobrecimento de grupos populacionais específicos (FRENK; LOZANO; BOBADILLA, 1994).

Em relação às melhorias na área da saúde e seus impactos sobre a taxa de letalidade, Frenk; Lozano e Bobadilla, (1994) afirmam que as mudanças na distribuição, organização e aumento das tecnologias dos serviços de saúde contribuíram para a transição epidemiológica. Parte importante da sobrevivência deve-se a diminuição das taxas de letalidade alcançada mediante a aplicação de tecnologias eficazes de diagnóstico e tratamento. Os efeitos dessas melhorias contribuíram na redução de mortes causadas pelas doenças infecciosas e parasitárias nas etapas iniciais da transição epidemiológica reduzindo a incidência de algumas doenças não transmissíveis.

Em 1930, no Brasil, o perfil epidemiológico consistia em doenças infecciosas como a primeira causa de morte e poucas causas de morte por doenças do aparelho circulatório e causas externas (PRATA, 1992). Já em 1985, as doenças do aparelho circulatório passam a ser as primeiras causas de morte em todas as regiões brasileiras. Porém, esta transição não se deu de forma uniforme, coexistindo regiões com uma grande incidência de mortes por doenças infecciosas ao mesmo tempo em que se aumenta o número de óbitos por doenças cardiovasculares (PRATA, 1992). Adiciona-se ao aumento das mortes por

doenças crônico-degenerativas, o crescimento dos óbitos por causas externas (MELLO JORGE; GAWRYSZEWSKI; LATORRE, 1997).

Sendo assim, o Brasil se depara com um perfil conjunto em que há permanência ou até mesmo ressurgimento de doenças infectocontagiosas ao mesmo tempo em que emerge o aumento da morbidade e mortalidade pelas doenças crônico-degenerativas. Apesar do acentuado declínio na mortalidade proporcional, de 45,7% em 1930 para 7,97% em 1986, quando se considera a mortalidade por 100 mil habitantes, o Brasil possui uma mortalidade devido às doenças infecciosas de 33/100.000 habitantes, considerada elevada em comparação com as taxas de outros países da América Latina, sendo, por exemplo, mais que o dobro da mortalidade observada no Chile (ARAÚJO, 2012).

O Brasil é um exemplo de um dos países que possui polarização epidemiológica, ou seja, em um mesmo período, o país conviveu com elevadas taxas de morbidade e mortalidade por doenças crônico-degenerativas com altas incidências e prevalências de doenças infecciosas e parasitárias, e a prolongada persistência de níveis diferenciados de transição entre grupos sociais distintos e entre regiões brasileiras.

As doenças infecciosas eram a primeira causa de morte em todas as regiões, porém no Nordeste correspondia a 60% e no Sudeste, 43% na década de 30. Já em 1985, as doenças do aparelho circulatório se tornaram a primeira causa de morte em todas as regiões, porém correspondiam a 37% no Centro-Oeste, Sul e Sudeste e 28% no Nordeste e 22% no Norte enquanto que as doenças infecciosas e parasitárias correspondiam a 16% na região Norte, 14% no Nordeste, 8% no Centro-Oeste, e apenas 5% no Sul e 4% no Sudeste (PRATA, 1992; ARAÚJO, 2012). No Brasil, tanto a adoção de medidas preventivas de saúde quanto o desenvolvimento econômico ocorreram simultaneamente e foram responsáveis por essas mudanças das doenças infecciosas para as doenças crônico-degenerativas (PRATA, 1992).

Durante a década de 90, a mortalidade manteve a tendência histórica de queda, principalmente a infanto-juvenil levando a uma maior diferenciação e complexidade na estrutura por causas de morte (SIMÕES, 2002). As causas relacionadas às neoplasias, doenças circulatórias e respiratórias, além das causas externas, afetam, em sua maioria, pessoas idosas e os jovens do sexo masculino, respectivamente. As causas externas estão assumindo cada vez mais importância na estrutura geral das causas de morte, tanto em termos absolutos quanto relativos (SIMÕES, 2002; PEREIRA; QUEIROZ, 2016).

A mortalidade por causas externas é a principal causa de morte para os jovens adultos de 15 a 29 anos no Brasil (GAWRYSZEWSKI; HIDALGO; VALENICH, 2005; 2006; BRASIL, 2008; PEREIRA; QUEIROZ, 2016). A taxa de mortalidade por causas externas

aumentou de 69,7 por cem mil habitantes em 2000 para 72,9 em 2009, um aumento de 3,9%, correspondendo 12,6% das mortes totais. Entre os tipos de ocorrência, acidentes de transporte afetam mais homens entre 20 e 39 anos (MOURA et al., 2016). Na outra ponta da estrutura etária, as pessoas idosas (60 anos ou mais de idade) estão mais expostas a problemas do aparelho circulatório e neoplasias (SIMÕES, 2002).

O aumento da mortalidade por causas externas afeta os jovens adultos e possui influência direta na variabilidade da idade ao morrer e aumento da esperança de vida. Esta variabilidade é um dos principais indicadores do processo de compressão da mortalidade. Nos países desenvolvidos, o declínio da mortalidade levou a uma redução nesta variabilidade, ou seja, a medida que os níveis de mortalidade declinam as mortes começam a se concentrar cada vez mais em um intervalo de idade restrito (FRIES, 1980).

Na América Latina, os países não seguiram as mesmas tendências de declínio de mortalidade e de perfil epidemiológico se comparado com os países desenvolvidos. Gonzaga; Queiroz e Lima (2018) mostraram que houve uma estagnação ou reversão da variabilidade nas últimas décadas, acompanhada por aumentos contínuos na esperança de vida ao nascer indicando que a distribuição das mortes está mudando para as idades mais avançadas, isto é, houve uma redução da mortalidade infantil responsável pelos principais aumentos da esperança de vida. Ao mesmo tempo houve um aumento ou estagnação das mortes prematuras nos países latino-americanos, incluindo o Brasil, devido à violência e acidentes de trânsito, responsável pela desaceleração da compressão da mortalidade (GONZAGA; QUEIROZ; LIMA, 2018). Nesse sentido, as mortes por causas externas estão contribuindo de forma negativa para a evolução da esperança de vida e redução da variabilidade da idade à morte.

Como se pôde observar, os processos de transição da mortalidade e morbidade são incorporados em um contexto e região, sendo assim, estão situados em tempos e espaços distintos. O significado dos contextos históricos, culturais e políticos para as transições ilustra a importância da inserção em uma determinada região e os processos descritos interagem com o ambiente em que se encontram, ou seja, influenciam o ambiente e são influenciados por ele e mostram assim, as desigualdades existentes (WILLEKENS, 2014).

1.3 Desigualdades socioeconômicas no Brasil

Por desigualdade social entende-se:

> [...] distribuição, dentro de uma sociedade, de bens e serviços materiais e não-materiais escassos; o termo abrange a distribuição de renda, habitação, educação, nutrição, serviços públicos (água, esgoto), bem como acesso a empregos e recursos produtivos tal como à terra no campo (WOOD; CARVALHO, 1994, p. 17).

A medida mais utilizada no que tange a medição da desigualdade é o coeficiente de Gini. A partir dele, observou-se que entre 2001 e 2005, o grau da desigualdade de renda no Brasil diminuiu de forma acelerada, declinou 4,6%, passando de 0,593 para 0,566. Apesar da acentuada queda, o país ainda se encontra, comparando-se com outros países, entre os mais desiguais do mundo. Ou seja, apesar do declínio, cerca de 90% do mundo ainda apresenta distribuições de renda menos concentradas que a do Brasil (BARROS et al., 2007).

Nos últimos 25 anos, algumas dinâmicas em relação a desigualdade de renda foram encontradas. Entre 1981 e 1993 houve um contínuo crescimento do grau de desigualdade e o principal fator que desencadeou isso foi o processo de aceleração inflacionária que levou a um aumento da desigualdade de rendimentos e assim, da desigualdade da renda familiar per capita. Além da inflação, outro fator que contribuiu para um aumento de desigualdade foi a expansão do nível de educação da força de trabalho que elevou a desigualdade entre diferentes níveis educacionais (FERREIRA et al., 2007; BARROS et al., 2007).

O período de 1993 a 2005 foi marcado por um persistente declínio do grau de desigualdade, consequência da redução nas diferenças entre áreas urbanas e rurais e aumento no volume das políticas de transferência do governo, principalmente, no período de 2001 a 2005 (FERREIRA et al., 2007; BARROS et al., 2007).

O desenvolvimento econômico brasileiro não foi capaz de atenuar a elevada concentração de renda e assim, o Brasil convive com uma aguda desigualdade socioeconômica até hoje. Em 1980, 47,9% da renda estavam nas mãos dos 10% mais ricos da população, enquanto que somente 1% da renda estava com os 10% mais pobres (WOOD; CARVALHO, 1994). Essa concentração se manteve no ano de 2005, apesar da queda da desigualdade de renda, ou seja, os 10% mais ricos se apropriando de mais de 40% da renda e os 40% mais pobres se apropriando de menos de 10% (BARROS et al., 2007).

Regionalmente, a concentração de renda também se apresenta de forma significativa. A participação do Nordeste na renda nacional foi de 14,9% em 1949 e 12,2% em

1970, mas possui 30% da população. No Sudeste ocorre o contrário, detinha 42,7% da população em 1970, porém concentrava 64,5% da renda nacional (WOOD & CARVALHO, 1994).

Em 1970, alguns poucos municípios do Sul e Sudeste apresentaram os mais altos valores de Produto Interno Bruto (PIB) per capita, ao passo que os municípios com menor PIB estavam concentrados na região Nordeste. Dez anos depois, o Nordeste manteve com os menores valores de PIB enquanto que no Sul e Sudeste mais municípios passaram a ter maiores valores. A desigualdade espacial se manteve nos três últimos Censos (1991, 2000 e 2010). As regiões Sul e Sudeste possuem os municípios mais ricos e as regiões Norte e Centro-Oeste alguns municípios ricos cercados por municípios com riqueza intermediária. Já os municípios da região Nordeste se encontram entre os mais pobres se comparados aos demais (ARRETCHE, 2015). Para mostras essas disparidades espaciais, o Mapa temático na Figura 2 exemplifica as desigualdades socioeconômicas em termos de taxa de desemprego e pib per capita observado nas microrregiões em 2010.

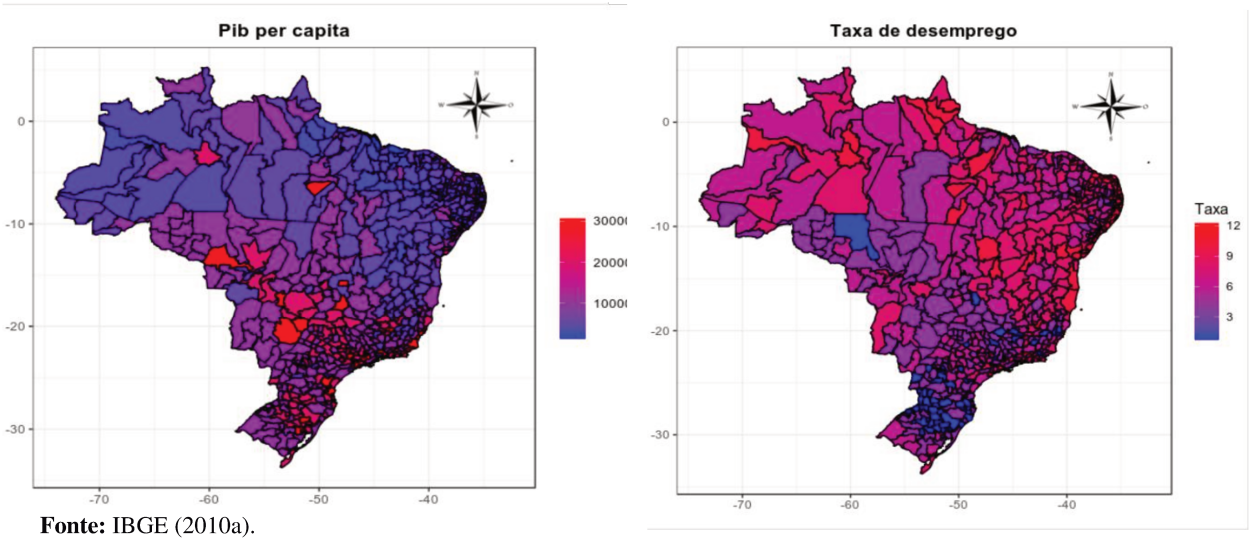
A oferta de serviços essenciais no Brasil também é capaz de dar um panorama da dimensão da desigualdade regional. Em 1970, a cobertura era pequena no Brasil como um todo, municípios em que mais da metade da população tinha acesso à rede elétrica eram raros e concentravam-se em São Paulo e nos estados do Sul. A escassez era maior na cobertura de água e esgoto (WOOD; CARVALHO, 1994; ARRETCHE, 2015).

O abastecimento de água e à ligação de rede elétrica aumentou em 1980, mas de forma desigual. As cidades do Sul e Sudeste caminham para a universalização. Porém os municípios das regiões Norte, Nordeste e Centro-Oeste não chegam à metade da cobertura. Em relação à oferta de esgoto, a ampliação de cobertura ocorreu apenas nos municípios das regiões Sul e Sudeste (WOOD; CARVALHO, 1994; ARRETCHE, 2015).

Em 2010, o acesso à energia elétrica era universal independente da renda per capita do município devido ao Programa Nacional da Universalização e Uso de Energias Elétricas. Em relação ao abastecimento de água, a cobertura dos municípios do Norte e Nordeste aumentou chegando aos 75%. Entretanto, a coleta de esgoto permaneceu praticamente inalterada (ARRETCHE, 2015).

Sendo a educação também um fator importante para a desigualdade, observou-se que, a distribuição da educação e seus retornos explicam cerca de 40% da distribuição de salários, ou seja, tem uma grande influência na renda das famílias. O período entre 1977 e 2004 mostra que a força de trabalho brasileira passou de 4,66 anos de estudos em média, para 7,61 (ARRETCHE, 2015).

FIGURA 2 – Taxa de desemprego e PIB per capita – Microrregião, Brasil – 2010



Fonte: IBGE (2010a).

Entre 1995 e 2005 houve uma expansão educacional acelerada no Brasil (MENEZES-FILHO; FERNANDES; PICHETTI, 2007). Este progresso foi mais de duas vezes o observado nos dez anos anteriores:

> A proporção de pessoas com ensino primário incompleto perfazia mais de 90% da população nascida em 1910, e sua parcela diminuiu paulatinamente, até cair de modo acelerado a partir da coorte de 1940. Os demais grupos educacionais aumentaram de participação, de forma relativamente homogênea, até 1950, quando então a parcela com ensino superior estagnou-se, enquanto os grupos intermediários continuaram a aumentar até o dos nascidos em 1970. Entre 1970 e 1982, a proporção de pessoas com nível médio aumentou significativamente, o grupo com ensino fundamental completo declinou, enquanto o grupo com nível superior passou a aumentar depois de décadas de estagnação (MENEZES-FILHO; FERNANDES; PICHETTI, 2007, p. 291).

Neste processo de aumento da frequência educacional, na década de 1980, houve um pequeno aumento da frequência escolar no ensino médio e a desigualdade educacional aumentou. A partir do final da década de 1990, quando a proporção de indivíduos no ensino médio aumentou, a desigualdade educacional começou a se reduzir (MENEZES-FILHO; FERNANDES; PICHETTI, 2007).

Os diferentes âmbitos das desigualdades que o Brasil apresenta, regionalmente, entre grupos populacionais, entre rendas, entre níveis educacionais, por exemplo mostra que o país ainda se encontra longe de atingir os níveis baixos de desigualdade encontrados em outros países. Barros et al. (2007) coloca que ainda seriam necessários mais 20 anos para que o país alcance um nível similar ao da média dos países com o mesmo grau de desenvolvimento que o Brasil.

Como colocado por Wood e Carvalho (1994), o comportamento da mortalidade em uma população está ligado às desigualdades sociais e econômicas devido ao fato de que as condições materiais dos indivíduos exercem forte influência sobre o padrão e nível dos óbitos e nesse sentido, torna-se necessário entender como as desigualdades estão associadas ao nível de mortalidade da população brasileira.

1.4 Fatores associados à mortalidade

Para se entender melhor as tendências de mortalidade e os determinantes que levam ao seu declínio ou aumento, é necessário identificar os fatores socioeconômicos e outros fatores associados. Esses fatores não interagem de forma aleatória e raramente operam sozinhos. Os riscos de mortalidade são determinados por correntes causais que consistem de uma série de fatores operando sucessivamente. Conhecer os mecanismos que estão por trás das tendências e variações da mortalidade é necessário para tentar se entender as suas tendências futuras (KUNST; WOLLESWINKEL-VAN DEN BOSCH; MACKENBACH, 1999; WILLEKENS, 1990).

A mortalidade é influenciada tanto por determinantes diretos e indiretos da doença. A exposição aos fatores de saúde ocorre durante a vida e podem ser específicos do indivíduo (fatores biológicos e comportamentais) ou compartilhados por aquele de uma mesma coorte ou período (desenvolvimento econômico, sistema educacional, assistência à saúde). Neste sentido, a mortalidade pode ser explicada por mecanismos causais por meio dos quais está diretamente ou indiretamente relacionada. Por exemplo, o *status* socioeconômico de um indivíduo não afeta diretamente sua saúde, mas atua sobre os determinantes próximos (hábitos alimentares, fumar, consumo de álcool, etc) (KUNST; WOLLESWINKEL-VAN DEN BOSCH; MACKENBACH, 1999; WILLEKENS, 1990).

No caso da mortalidade infantil, Mosley e Chen (1984) afirmam que é a identificação de um conjunto de determinantes próximos, ou variáveis intermediárias, que influenciam diretamente no risco de morbidade e mortalidade. Os determinantes sociais e econômicos devem operar por meio dessas variáveis para afetar a sobrevivência infantil. Estes determinantes socioeconômicos podem ser divididos em três grandes categorias de variáveis.

Em primeiro lugar, as variáveis de nível individual: produtividade dos pais; tradições, normas e atitudes. Neste sentido, o nível educacional dos pais, por exemplo, possui impacto na sobrevivência infantil. Em relação as tradições, normas e atitudes, os fatores modificam as escolhas econômicas e as práticas relacionadas à saúde de acordo com as tradições e normas culturais da sociedade em que vivem (MOSLEY; CHEN, 1984).

Em segundo, as variáveis de nível doméstico como renda/riqueza, ou seja, uma variedade de bens, serviços e ativos no nível doméstico operam na saúde e mortalidade infantil por meio dos determinantes próximos. Algumas das principais formas pelas quais os efeitos da renda influenciam a saúde da criança são por meio da disponibilidade de alimentos e nutrientes adequados, abastecimento de água encanada, habitação, higiene e cuidados preventivos, cuidados de saúde e informação (sobre nutrição, higiene e imunizações adequadas) (MOSLEY; CHEN, 1984).

E por último, as variáveis a nível comunitário como o ambiente, a economia do local, política e sistema de saúde. As variáveis relacionadas à organização da produção podem determinar a distribuição de recursos e as variáveis ligadas à infraestrutura física como eletricidade, água e esgoto podem influenciar a saúde das crianças (MOSLEY; CHEN, 1984).

Uma das formas encontradas no Brasil de melhorar a saúde das famílias e das crianças ao nível comunitário foi através do sistema de Atenção Básica à Saúde (ABS) visando ações preventivas e de proteção, nutrição, planejamento familiar, imunização, fornecimento de medicamentos essenciais, entre outras. A criação do Programa Saúde da Família (componente da ABS) se deu para melhorar o modelo assistencial e centrado na promoção da qualidade de vida focando na realidade de cada local e levando em conta os princípios que regem o Sistema Único de Saúde (SUS) de integralidade, universalidade e equidade promovendo ações básicas de modo a prevenir que as pessoas fiquem doentes (SOUSA; HAMANN, 2009).

Nos municípios que possuem implementado o Programa Saúde da Família se observou maior redução da mortalidade infantil em comparação com os que não possuíam (MACINKO; GUANAIS; SOUZA, 2006) e os países orientados pelo modelo de Atenção Básica possuem melhores indicadores na saúde como a detecção precoce de cânceres, menor mortalidade evitável devido a causas preveníveis e maior expectativa de vida (SHI et al, 2004).

Em alguns aspectos, as tendências e diferenciais geográficos na mortalidade adulta podem ser diferentes daqueles que influenciam a mortalidade infantil (TIMAEUS; CHACKIEL; RUZICKA, 1996). Para a mortalidade adulta, os fatores contextuais podem instigar uma mudança direta nos indicadores de risco próximo como, por exemplo, condições macroeconômicas, fatores culturais, sistema de saúde e fatores ecológicos.

A condição macroeconômica é considerada um fator contextual e se refere aos indicadores econômicos de uma área geográfica como um todo e não de indivíduos. Os fatores macroeconômicos podem ter um efeito direto ou indireto sobre os indicadores de risco imediatos que estão diretamente associados à saúde. Por exemplo, um aumento nos gastos do

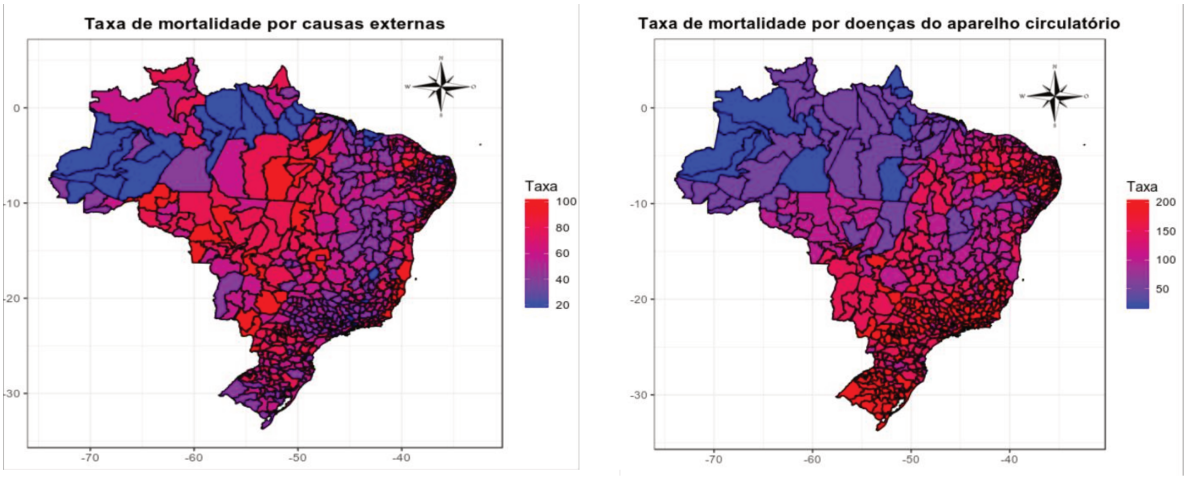
governo com a saúde pública levará a uma população mais saudável (VAN POPPEL, 1991). Um Produto Interno Bruto (PIB) mais alto, por exemplo, resulta em vidas mais longas e saudáveis, pois o crescimento da renda gerado pelo crescimento regional possui influência sobre a saúde e os níveis de mortalidade e recursos adicionais permitem que mais necessidades básicas sejam compradas, especialmente alimentos e busca por melhor habitação. Uma explicação que foi dada para a associação entre maior igualdade de renda e melhor saúde foi a sua tendência em melhorar a coesão social e reduzir as divisões sociais, enquanto, inversamente, a pobreza está associada à exclusão social (SPIJKER, 2004).

Kaplan et al. (1996) mostraram que os estados dos EUA com maior desigualdade de renda apresentaram taxas mais altas de violência, maiores incapacidades, mais pessoas sem plano de saúde, menos investimento em educação e alfabetização e resultados educacionais mais pobres. Adicionalmente, de acordo com o estudo de Duarte et al. (2002), o risco de mortalidade da alguma forma é afetado pelas transformações relacionadas à urbanização e ao aumento da escolarização da população e, assim, o risco de mortalidade precoce está associado às características socioeconômicas do local de residência. Como fator contextual, diferenças geográficas e sociais no acesso aos cuidados de saúde podem afetar a mortalidade adulta (MCKEE, 1999).

No Brasil, essas diferenças geográficas são observadas em diferentes aspectos. Na Figura 3, observa-se diferenças significativas nas taxas de mortalidade por causas externas e por doenças do aparelho circulatório por cem mil habitantes nas microrregiões brasileiras, em que regiões que apresentam taxas de mortalidade adulta mais alta, em sua maioria, apresentam taxas de mortalidade por doenças do aparelho circulatório mais baixas. As maiores taxas de mortalidade por doenças do aparelho circulatório podem estar relacionadas às regiões com maior índice de envelhecimento, localizadas nas áreas Sul e Sudeste, em que a transição demográfica encontra-se em fases mais avançadas (CHAIMOWICZ, 1997). Enquanto que as taxas de mortalidade por causas externas estão associadas ao grupo mais jovem de 15 a 39 anos e pode estar associado às regiões em fases intermediárias da transição demográfica.

Outro aspecto que diferencia significativamente os níveis de mortalidade entre regiões e grupos populacionais é a variável de raça/cor (ARAÚJO, 2007; TINEU; BORGES, 2016). Na perspectiva socioeconômica e social, os negros residem em áreas com menor infraestrutura básicas e com maiores restrições no acesso aos serviços de saúde (ARAÚJO, 2007). Williams (2012) afirma que tanto nos Estados Unidos como globalmente, os grupos raciais desfavorecidos experimentam taxas mais elevadas de doenças, incapacidades e mortalidade do que a média.

FIGURA 3 – Taxa de mortalidade: causas externas e aparelho circulatório – Microrregiões, Brasil, 2010



Fonte: IBGE (2010a); Brasil (2010).

O estudo de Araújo (2007) sobre a cidade de Salvador, no Brasil ao tomar a "variável raça/cor enquanto construto social determinante de diferenciais em saúde" (ARAÚJO, 2007, p. 126) argumenta que áreas em que se predomina negros possuem condições socioeconômicas desfavoráveis e mostra que uma região que possui maior proporção de negros de 15 a 49 anos apresentam maiores riscos de morte por causas externas, principalmente homicídios.

Alguns trabalhos já mostraram que melhores indicadores socioeconômicos como altas taxas de alfabetização, maior grau de urbanização e maior PIB per capita do local estão relacionados a menores taxas de mortalidade adulta ou esperança de vida ao nascer maior, no contexto de taxas de mortalidade específicas, como taxa de mortalidade por doenças cardiovasculares (DUARTE et al, 2002; ISHITANI, 2006).

Wood e Carvalho (1994) observaram que os diferenciais regionais de mortalidade ocorrem, em alguma magnitude, devido a variações na distribuição de renda nas diversas regiões do Brasil. Para eles, existem três conjuntos de fatores que impactam o bem-estar da população e por sua vez os níveis de mortalidade: os serviços de saúde, os serviços que levam a uma melhoria no nível de saúde como disponibilidade de água potável e, por último, as características individuais como renda e moradia e estão associados a como os indivíduos respondem as adversidades na saúde e no ambiente.

Indicadores de pobreza, desigualdade educacional e de oportunidades como a taxa de trabalho infantil estão fortemente associados a maiores riscos na saúde das crianças e na vida adulta e possuem níveis desiguais regionalmente:

> No curto prazo, o impacto econômico mais óbvio do trabalho infantil no nível da família é o aumento da renda familiar. A longo prazo, a subacumulação de capital humano causada por baixa frequência escolar e saúde precária é uma séria consequência negativa do trabalho infantil, representando uma oportunidade perdida de aumentar a produtividade e a capacidade de ganhos futuros da próxima geração. Trabalhadores infantis crescem para serem adultos com salários baixos. Como resultado, seus filhos também serão obrigados a trabalhar para complementar a renda da família. Desta forma, a pobreza e o trabalho infantil são transmitidos de geração em geração (ROGGERO et al., 2007, p. 271, Tradução própria[1]).

A queda dos níveis de mortalidade no Brasil não ocorreu de forma homogênea e sua queda agregada não eliminou as diferenças de mortalidade por região e por estrato socioeconômico (WOOD; CARVALHO, 1994). Coexiste um padrão de exposição a riscos de mortalidade semelhante ao de países desenvolvidos juntamente com um Brasil que ainda impera a exposição aos riscos típicos de países menos desenvolvidos (DUARTE et al., 2002).

Observa-se na Figura 4 que as microrregiões localizadas na região Norte do Brasil, em azul, se destacam por estar em estados considerados menos desenvolvidos em que se esperaria maior mortalidade adulta, porém possuem mortalidade adulta baixa similar às microrregiões do Sul do país. No Estado do Amazonas, as microrregiões que se destacam em azul são: Japurá, Juruá, Parintins, Purus e Madeira. No Acre é a microrregião de Brasiléia e no Pará é a região de Óbidos.

Essas microrregiões possuem baixa cobertura de água encanada, de lixo e esgoto, baixa cobertura do programa saúde da família, altas taxas de desemprego, baixo grau de urbanização e baixas taxas de mortalidade por causas externas. Era de se esperar altas taxas de mortalidade adulta, porém pode-se perceber que são regiões com baixo grau de urbanização, o que poderia determinar baixas taxas de mortalidade por causas externas, apesar de condições socioeconômicas desfavoráveis e baixa probabilidade de morte adulta nestas regiões (SANTANA et al., 2015).

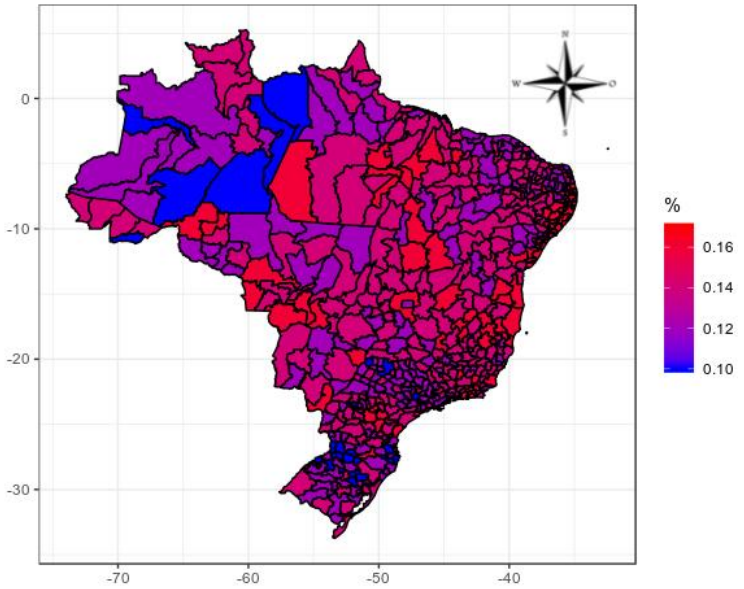
As desigualdades socioeconômicas em relação à mortalidade adulta são mais pronunciadas em áreas mais urbanizadas, em que as populações desfavorecidas e pobres estão concentradas em bairros marginalizados e que as áreas urbanas têm certas características especiais que podem influenciar a saúde da população (BORRELL et al., 2014).

Países em desenvolvimento, como Brasil, tiveram uma urbanização acelerada, porém de forma não planejada levando a regiões com pouca infraestrutura e aos paradoxos em

1 "Short term, the most obvious economic impact of child labor at the family level is an increase in household income. Long term, the under accumulation of human capital caused by low school attendance and poor health is a serious negative consequence of child labor, representing a missed opportunity to enhance the productivity and future earnings capacity of the next generation. Child laborers grow up to be low-wage–earning adults; as a result, their offspring will also be compelled to work to supplement the family's income. In this way, poverty and child labor is passed from generation to generation".

que as regiões urbanas oferecem mais oportunidades de emprego, maiores possibilidades de acesso à saúde e simultaneamente a urbanização cria um ambiente que concentra riscos e novos perigos como aumento da violência, acidentes, dietas não saudáveis e falta de atividade física (WHO, 2010).

FIGURA 4 – Probabilidade de morte adulta corrigida ($_{45}q_{15}$) – Microrregiões, Brasil – 2010



Fonte: Adaptado de Schmertmann e Gonzaga (2018).

Por sua vez, um estudo sobre mortalidade adulta entre os 50 estados do Estados Unidos demonstrou que os níveis gerais de mortalidade adulta diferem em nível geográfico, sendo que os estados variam em questões de políticas, recursos e estruturas de oportunidade afetando em algumas maneiras a saúde da população (MONTEZ et al., 2019).

Os fatores contextuais são importantes e determinantes cruciais da mortalidade adulta, isto é, após levar em consideração características individuais como, por exemplo, sexo, idade, raça e educação, o Estado possui um importante papel para o aumento ou declínio da mortalidade deste grupo etário. Em termos de escolaridade, o estudo mostrou que as políticas públicas acabam impactando mais adultos de menor renda e baixa escolaridade, argumentando que adultos com maior nível educacional podem melhorar sua saúde e longevidade com recursos próprios independente das políticas públicas de desemprego, bem-estar, salário mínimo oferecido pelo estado (MONTEZ et al., 2019).

Belon; Barros e Marín-León (2012) ao estudar diferenciais socioeconômicos em um município brasileiro encontraram que as maiores taxas de mortalidade entre adultos estavam concentradas em áreas com condições de vida precárias com um aumento gradiente

da mortalidade à medida que diminui o nível socioeconômico do município. Além disso, as desigualdades sociais na mortalidade foram maiores entre a população jovem e adulta ao se comparar com a população idosa.

CAPÍTULO 2 – MATERIAL E MÉTODOS

Os dados foram oriundos do Censo Demográfico de 2010 realizado pelo Instituto Brasileiro de Geografia e Estatística (IBGE); do Sistema de Informação de Mortalidade, do Cadastro Nacional de Estabelecimentos de Saúde (CNES) e do Sistema de Informação da Atenção Básica (SIAB) organizados pelo Ministério da Saúde por meio do Departamento de Informática do Sistema Único de Saúde do Brasil (DATASUS) também para o ano de 2010.

Para o cálculo da variável probabilidade de morte adulta ($_{45}q_{15}$), utilizou-se os óbitos do Sistema de Informação de mortalidade (SIM) usando a abordagem de TOPALS com estimação bayesiana de correção proposta por Schmertmann e Gonzaga (2018) que lança mão da estimação bayesiana para a mortalidade específica por idade em pequenas áreas com registros vitais defeituosos incorporando ao modelo a incerteza em relação aos níveis de mortalidade.

Buscou-se combinar variáveis socioeconômicas, demográficas e de acesso a saúde em nível das microrregiões com o objetivo de entender as relações destas variáveis com a probabilidade de morte adulta por meio de algoritmos de aprendizado de máquina. A combinação das variáveis das diferentes fontes foi feita através do código da microrregião. Por microrregiões entende-se o agrupamento de municípios limítrofes com base em similaridades econômicas e sociais. O Brasil é dividido entre 558 microrregiões e consiste na amostra de análise deste estudo.

2.1 Fontes de dados

O Censo Demográfico agrega informações sobre a população brasileira como educação, renda, trabalho, habitação e vulnerabilidade com periodicidade de 10 anos. Essa pesquisa possui cobertura nacional e, teoricamente, todos os domicílios são recenseados. No total são 5.565 municípios e mais de 67 milhões de domicílios sendo estes agregados em mais de 300 mil setores censitários. O setor censitário é a unidade territorial criada para controle de coleta dos dados, estabelecido por áreas contíguas, respeitando-se o limite da divisão político-administrativa.

É a principal fonte de informação do Brasil pelo tamanho de sua cobertura e nível de complexidade, sendo possível desagregar as informações em nível municipal. Para este estudo foram usados os dados das 558 microrregiões referentes à malha municipal existente em 2010. A pesquisa é do tipo transversal e no caso do ano de 2010 sua data de referência foi 31 de julho de 2010.

No Censo Demográfico de 2010 foram aplicados o questionário básico (aplicado em todos os domicílios, exceto naqueles selecionados para a amostra, em que responderá um questionário maior que contém as informações básicas mais complementares), e o questionário da amostra (é aplicado para todas as unidades selecionadas para a amostra. Este abrange características do domicílio e pesquisa importantes informações sociais, econômicas e demográficas). Foram selecionados 6.192.332 domicílios para responder ao Questionário da Amostra, o que constitui uma fração amostral de 10,7% para o país como um todo. Os níveis geográficos utilizados e fornecidos pelo IBGE são todo o território nacional, Unidades de Federação, Grandes Regiões, Mesorregiões, Microrregiões, Municípios e Áreas de Ponderação.

Define-se área de ponderação como a unidade geográfica que agrupa setores censitários para que seja possível aplicar os procedimentos de calibração das estimativas. São 10.184 áreas de ponderação no Censo de 2010 (IBGE, 2010a).

As fontes CNES e SIAB são fornecidas pelo DATASUS, o banco de dados de estatísticas vitais do SUS. Essas fontes de informação juntam-se aos dois principais sistemas de informações: SINASC (Sistema de Informações sobre Nascidos Vivos) que são originados da Declaração de Nascido Vivo, o SIM (Sistema de Informações sobre Mortalidade) em que os dados são originados da Declaração de Óbitos. Ambos os sistemas são alimentados pelos dados do Registro Civil que acompanham as ocorrências de nascimentos, óbitos, migrações, nupcialidade, etc.

O Registro Civil deve registrar os eventos vitais e publicar informes e levantamentos estatísticos do registro destes dados, esses dados ficam disponíveis no site do IBGE, e possui informações sobre nascidos vivos, divórcios, separações judiciais, casamentos, óbitos e óbitos fetais para o Brasil.

Por sua vez, o CNES disponibiliza a infraestrutura dos serviços de saúde, tipo de atendimento prestado, serviços especializados, leitos e a capacidade instalada existente e disponível com o objetivo de auxiliar no planejamento em saúde. O CNES é o cadastro oficial do Ministério da Saúde no que diz respeito à capacidade instalada e mão-de-obra assistencial de saúde no Brasil em estabelecimentos de saúde público ou privados. Tem por finalidades principais cadastrar e manter atualizado as informações sobre os estabelecimentos de saúde em relação a recursos físicos e humano, disponibilizar tais informações para outros sistemas, disponibilizar informações à sociedade sobre os serviços disponíveis e suas formas de acesso e fornecer informações para tomada de decisão e planejamento.

Por meio do SIAB, obtêm-se informações sobre cadastros de famílias, condições de moradia e saneamento, situação de saúde, produção e composição das equipes de saúde. É o principal instrumento que monitora as ações do Programa Saúde da Família (PSF) e tem como objetivos monitorar e avaliar a atenção básica e consolidar a avaliação nas três instâncias de gestão do SUS. Acompanha as ações e resultados das atividades realizadas pelas equipes do PSF.

2.2 Variáveis utilizadas

A Tabela 1 apresenta as variáveis que foram utilizadas nos modelos, as fontes de onde foram retiradas e uma descrição referente a forma como as mesmas foram operacionalizadas.

As variáveis escolhidas apresentam diferentes ângulos de características socioeconômicas, demográficas e de saúde nas microrregiões brasileiras. A Coluna ‘VAR’ indica o nome das variáveis como elas foram usadas no banco de dados criado, a ‘NOME’ é o nome completo da variável, a fonte da qual foi retirada e a descrição determina o cálculo utilizado para construir cada uma delas.

TABELA 1 – Variáveis utilizadas no modelo

VAR	NOME	Fonte	Descrição
$_{45}q_{15}$	Probabilidade de morte adulta	-	Categorias: P0: $_{45}q_{15}$<0.1406; P50: 45q15>=0.1406
cob_PSF	Cobertura do Programa Saúde da Família	Datasus, 2010	Número de famílias atingidas pelo programa saúde da família/total de famílias
cob_agua	Cobertura de Água encanada	Censo, 2010	Razão entre a população que vive em domicílios particulares permanentes com água canalizada para um ou mais cômodos e a população total residente em domicílios particulares permanentes por 100. A água pode ser proveniente de rede geral, de poço, de nascente ou de reservatório abastecido por água das chuvas ou carro-pipa.
cob_lixo	Cobertura de coleta de lixo	Censo, 2010	Razão entre a população que vive em domicílios com coleta de lixo e a população total residente em domicílios particulares permanentes multiplicados por 100. Estão incluídas as situações em que a coleta de lixo é realizada diretamente por empresa pública ou privada, ou o lixo é depositado em caçamba, tanque ou depósito fora do domicílio, para posterior coleta pela prestadora de serviços. São considerados apenas os domicílios particulares permanentes localizados em área urbana
cob_esg	Cobertura de Esgoto	Censo, 2010	Razão entre a população residente em domicílios particulares permanentes servidos por rede coletora ou fossa séptica no domicilio e a população total residente em domicílios particulares permanentes x 100
cob_vacina	Cobertura de Vacinação	CNES, 2010	Número de crianças com esquema básico completo na idade alvo para determinado tipo de vacina sobre número de crianças na idade alvo x 100
n_leitos	Número de leitos	CNES, 2010	Razão da média anual do número mensal de leitos hospitalares existentes e a população total residente x 100
pib_percapita	PIB per capita	Censo, 2010	valor do PIB em moeda corrente, a preços de mercado/população total residente
tx_analf	Taxa de analfabetismo	Censo, 2010	Razão entre o número de pessoas residentes de 15 anos ou mais de idade que não sabem ler e escrever um bilhete simples e a população total residente desta faixa etária x 100
renda	Renda Média Domiciliar	Censo, 2010	Soma das rendas domiciliares per capita/ População residente em que a soma das rendas domiciliares per capita é a soma da renda dos moradores sobre o número de moradores no domicílio
tx_desemp	Taxa de Desemprego	Censo, 2010	Número de residentes com mais de 16 anos e economicamente ativos que se encontram desocupados e procurando trabalho, na semana de referência sobre o número de residentes economicamente ativos (PEA) com mais de 16 anos.
cursosup	% de Indivíduos do Curso Superior	Censo, 2010	Pessoas de 25 anos ou mais de idade, residentes em domicílios particulares (Pessoas) sobre o total da população de 25 anos ou mais, residentes em domicílios particulares.
tx_trabinf	Taxa de Trabalho Infantil	Censo, 2010	Número de crianças residentes de 10 a 15 anos de idade que se encontram trabalhando ou procurando emprego na semana de referência sobre a População total residente desta mesma faixa etária x 100
rz_sexo	Razão de Sexo	Censo, 2010	Número de residentes do sexo masculino sobre residentes do sexo feminino x 100
brancos	% de Indivíduos de cor branca	Censo, 2010	% residentes que se autodeclararam de cor branca sobre o total dos indivíduos
negros	% de Indivíduos de cor preta e parda	Censo, 2010	% residentes que se autodeclararam de cor preta ou parda sobre o total dos indivíduos
grau_urb	Grau de Urbanização	Censo, 2010	População urbana residente sobre população total residente x 100
prop_jovem	% de jovens de 15 a 29 anos	Censo, 2010	Proporção de jovens de 15 a 29 anos sobre a população total residente
tx_pobreza	Taxa de Pobreza	Censo, 2010	Domicílios particulares permanentes com menos de ½ salário mínimo médio mensal sobre o total de domicílios particulares residentes
ndice_envelhecimento	Índice de Envelhecimento	Censo, 2010	Número de pessoas residentes de 60 anos ou mais de idade sobre o número de pessoas residentes com menos de 15 anos de idade x 100
tx_causasexternas	Taxa de mortalidade por causas externas	SIM – Datasus	Número de óbitos de residentes por causas externas sobre população total residente ajustada ao meio do ano x 100.000
tx_aparelhocirc	Taxa de mortalidade por doenças do aparelho circulatório	SIM – Datasus	Número de óbitos de residentes por doenças do aparelho circulatório sobre a população total residente x 100.000

Fonte: Elaboração própria.

2.2.1 Probabilidades de morte adulta: correção dos óbitos

Para que fosse possível calcular a probabilidade de morte adulta de forma a refletir o número de óbitos reais de uma microrregião foi aplicada a técnica de TOPALS em conjunto com bayesiano empírico em pequenas áreas para a correção dos óbitos.

A estimativa da mortalidade em pequenas áreas possui o problema universal de populações pequenas e alta variabilidade amostral em mortes registradas e com baixas taxas de mortalidade e curtos períodos de exposição. As taxas de eventos / exposições observadas são muito instáveis e, assim, a estimativa dos padrões de mortalidade se torna difícil. Em tais situações, os modelos de correção de óbitos devem apresentar formas de preencher estas lacunas.

Além disso, no Brasil, os registros vitais são incompletos, e desta forma, algumas mortes não são registradas pelas estatísticas oficiais. Vários métodos baseados em equações de dinâmica populacional foram desenvolvidos para avaliar a cobertura de mortes relatadas em relação à população. Os métodos de distribuição de morte (DDM) são usados para estimar a mortalidade de adultos em uma população instável (TIMAEUS, 1991; HILL; YOU; CHOI, 2009). Existem três abordagens principais: o Método Geral de Equilíbrio de Crescimento (GGB) (HILL, 1987), o Método de Gerações Extintas (SEG) (BENNETH; HORIUCHI, 1981), e o método de Gerações Extintas Ajustado (SEG-ajustado) (HILL; YOU; CHOI, 2009). Estes métodos inicialmente propostos para estimar a cobertura dos registros de óbitos e ajustar a mortalidade dependem de uma estabilidade da composição do sexo e idade da população e que a área seja fechada a fluxos migratórios.

Eles não fornecem medidas de incerteza sobre a integridade dos registros de óbitos (MURRAY et al., 2010). A combinação de baixo risco e populações pequenas pode tornar a estimativa de pequenas áreas, como as microrregiões, uma tarefa difícil. Nessas circunstâncias, as taxas de eventos / exposições observadas são muitas vezes instáveis entre idades e tempos, e a estimativa dos padrões de mortalidade torna-se mais difícil (THATCHER; KANNISTO; ANDREEV, 2002; ASSUNÇÃO et al., 2005; DIVINO; EGIDI; SALVATORE, 2009).

Os padrões de mortalidade são menos claros nas pequenas áreas. No Brasil, estes padrões diferem significativamente em grandes regiões (AGOSTINHO, 2009). As diferenças se encontram na mortalidade por idade específica por causa do óbito e nos padrões de idade de mortalidade por todas as causas (PRATA, 1992; ARAÚJO, 2012). Neste sentido, para se trabalhar com probabilidade de morte adulta nas microrregiões, escolheu-se o modelo proposto por Gonzaga e Schmertmann (2016) e Schmertmann e Gonzaga (2018) usando

regressão TOPALS e estimação Bayesiana. De forma geral, os autores sugerem um método diferente da padronização indireta (PI), ou seja, um método comum para lidar com dados esparsos para populações pequenas. A PI assume um padrão específico de mortalidade relativa por idade e estima apenas o nível de mortalidade local a partir de dados de morte e exposição locais específicos por idade.

O padrão de idade assumido para uma pequena área é geralmente o padrão empírico observado em um agregado geográfico maior, como o estado ao qual a pequena área pertence (GONZAGA; SCHMERTMANN, 2016). A qualidade das estimativas de mortalidade depende do pressuposto fundamental da PI, isto é, que um padrão específico de taxas de mortalidade relativas por idade está correto.

A alternativa à PI, proposta pelos autores, permite estimar as mortes em pequenas áreas sem a imposição rígida sobre os padrões das taxas específicas de mortalidade para suavizar as taxas de mortalidade em áreas com registro vital defeituoso. Gonzaga e Schmertmann (2016) propõe um método de regressão de Poisson baseado em TOPALS, um modelo relacional desenvolvido por Beer (2012) para suavizar e projetar probabilidades de morte específicas por idade. TOPALS define relações completas de taxas específicas por idade por meio de ajustes matemáticos para uma tabela padrão especificada. Constrói um conjunto ajustado de logaritmos de taxas de mortalidade em idades de 0 a 99, adicionando uma função *spline* linear com sete parâmetros (α0 ... α1) a um conjunto padrão pré-especificado.

Schmmertman e Gonzaga (2018) disponibilizaram todo o modelo em uma linguagem chamada *Stan* (CARPENTER et al., 2017). Para ambos os sexos, pode-se retirar a informação estimada das distribuições posteriores dos parâmetros de mortalidade (α), tabelas completas do logaritmo da taxa de mortalidade para todos os 27 estados brasileiros e todas as 558 microrregiões.

A partir do logaritmo, obteve-se a taxa de mortalidade por idade simples de (0 a 99 anos) por microrregião e partir dela calculou-se a probabilidade de sobreviver da idade n até a idade x, $p_x({}_np_x=\exp(-{}_nm_x))$, número de sobreviventes a idade x na coorte ($l_{x+n} = l_x * {}_np_x$) e assim, calculou-se a probabilidade de morte adulta, ${}_{45}q_{15}$ (${}_{45}q_{15} = l_{59} / l_{15}$).

2.3 Métodos de Aprendizado de Máquina

O Aprendizado de Máquina é um subcampo da inteligência artificial que atraiu considerável interesse nas últimas décadas (ARPINO; LE MOGLIE; MENCARINI, 2018). O método de Aprendizado de Máquina é útil como um substituto ou um complemento à

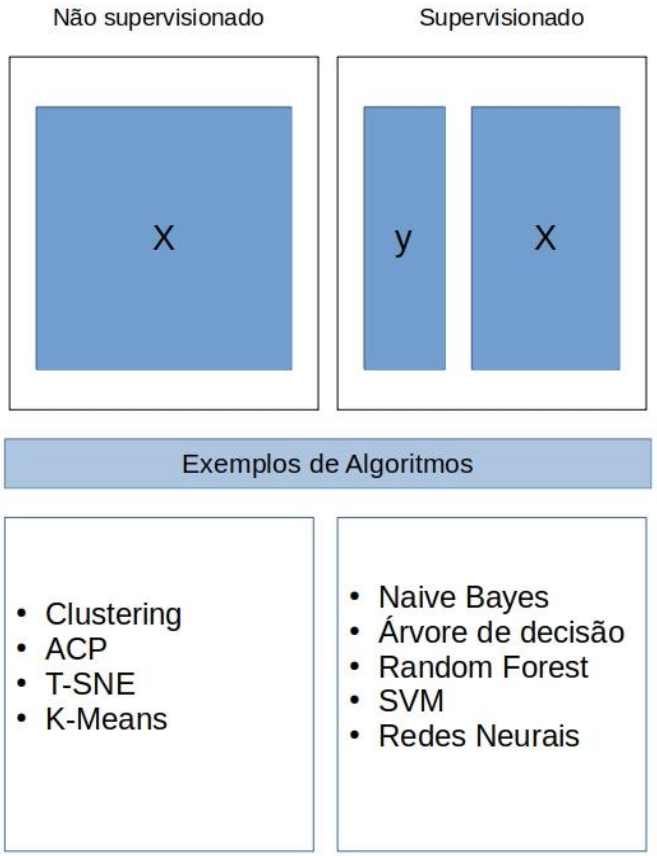
regressão paramétrica. Diferentemente das abordagens baseadas em regressão tradicional, o aprendizado de máquina não impõe um modelo paramétrico ligando uma variável dependente a um conjunto de variáveis independentes. A ideia chave é deixar o algoritmo encontrar o caminho para o resultado e ligações entre as variáveis independentes. Desta forma, é possível procurar automaticamente relações e interações entre as variáveis independentes. Além disso, colinearidade e violações de pressupostos não são preocupações importantes dependendo do algoritmo escolhido (DE ROSE; PALLARA, 1997; BILLARI; FÜRNKRANZ; PRSKAWETZ, 2006).

Existem duas grandes categorias de técnicas de Aprendizado de Máquina, aprendizagem "supervisionada" e a "não supervisionada". "Aprendizado não supervisionado" enfoca métodos para encontrar padrões em dados e para redução de dados (KUHN; JOHNSON, 2013). Algoritmos de aprendizagem não supervisionados foram desenvolvidos e aplicados a problemas como agrupamento ou classificação de imagens, vídeos e documentos de texto em grupos semelhantes. O aprendizado não supervisionado é muito utilizado para se agrupar dados não rotulados baseado na similaridade das características, enquanto o aprendizado supervisionado é adequado para a modelagem preditiva por meio da construção de algumas relações entre características socioeconômicas (como entradas) e o resultado de interesse (como resultado, a probabilidade de morte adulta). Estes dois tipos de aprendizagem e alguns exemplos de algoritmos que se encaixam nessas categorias são ilustrados na Figura 5.

As técnicas supervisionadas de Aprendizado de Máquina são algoritmos iterativos para aproximação de funções. Esses métodos focam principalmente em problemas de previsão: dado um "conjunto de dados de treinamento" com dados sobre um determinado resultado Y, que pode ser categórico, discreto ou contínuo, e algumas covariáveis X, o objetivo é estimar um modelo para prever resultados em um novo conjunto de dados (um conjunto de dados de "teste") em função de X.

Apesar do objetivo primário dessas técnicas ser construir um modelo preditivo, esses métodos podem ser proveitosamente usados para examinar como um conjunto (potencialmente grande) de variáveis independentes está vinculado a um resultado. Portanto, essas técnicas podem ser usadas como uma alternativa não paramétrica às abordagens do tipo regressão ao estudar as relações entre um conjunto de variáveis independentes e uma variável dependente (KUHN; JOHNSON, 2013).

FIGURA 5 – Exemplos de algoritmos de aprendizado supervisionado e não supervisionado



Fonte: Adaptado de Jiang et al. (2017).
Obs: ACP: Análise de Componentes Principais; SVM: *Support Vector Machine.*

Uma característica individual dos algoritmos supervisionados é que o modelo de construção é orientado por dados, de modo que eles podem se ajustar às relações complexas de maneira automática, superando principalmente os esforços de seleção de variáveis e construção de modelos. Mais especificamente, os algoritmos de Aprendizagem de Máquina podem detectar automaticamente não-linearidades e não-aditivos. Estes algoritmos podem ser úteis para melhorar a análise de dados devido à sua flexibilidade, em particular quando se lida com conjuntos de dados grandes (em termos de tamanho de amostra e número de covariáveis).

Com a finalidade de analisar os principais determinantes da probabilidade de morte adulta nas microrregiões brasileiras em 2010, o presente estudo utilizou os modelos de Aprendizado de Máquina. Quatro algoritmos foram testados e analisados: *Random Forest, Extreme Gradient Boosted Trees, Support Vector Machine e Naive Bayes*. Os algoritmos utilizados foram capazes de classificar os determinantes da probabilidade de morte adulta nas microrregiões, destacando os determinantes mais poderosos.

Na presente pesquisa, os algoritmos não supervisionados de agrupamento foram utilizados para verificar a interação entre as variáveis de entrada (variáveis explicativas, preditores), ou seja, foi implementado a Análise de Componentes Principais (ACP) e T-SNE (Incorporação de Vizinhos Estocásticos Distribuídos) juntamente com o algoritmo K-Means.

A partir destes agrupamentos foi possível entender como essas variáveis se agrupam e interagem. Além disso, todas as variáveis foram analisadas em relação a probabilidade de morte adulta, histogramas e box-plot para entender as linearidade e não-linearidades.

Para a modelagem, foram testados alguns tipos de algoritmos supervisionados e teve como objetivos: comparar a performance de quatro algoritmos de aprendizado de máquina e usar os dois melhores modelos para fazer uma discussão sobre as variáveis que mais possuem poder explicativo em relação à mortalidade adulta.

2.3.1 Pré-processamento

Transformações de variáveis preditoras podem ser necessárias. Algumas técnicas de modelagem podem ter requisitos estritos, como os preditores possuírem uma escala comum. A maioria dos conjuntos de dados exige algum grau de pré-processamento para expandir o universo de possíveis modelos preditivos e otimizar o desempenho preditivo de cada modelo (KUHN; JOHNSON, 2013).

Como as variáveis selecionadas possuem escalas muito diferentes entre si como, por exemplo, cobertura de vacina que é uma proporção (%) e PIB per capita que vai de valores de 2000 até 20.000, optou-se por realizar algumas transformações nas variáveis do estudo (KUHN; JOHNSON, 2013).

Duas transformações foram conduzidas: centralização e transformação em mesma escala. Para centralizar uma variável preditora, o valor médio do preditor é subtraído de todos os valores. Como resultado da centralização, o preditor tem uma média zero. Da mesma forma, para os dados terem a mesma escala, cada valor da variável preditora é dividido pelo seu desvio padrão. Colocar os dados na mesma escala força os valores a terem um desvio padrão comuns de um. Essas manipulações são geralmente usadas para melhorar a estabilidade numérica de alguns algoritmos de Aprendizado de Máquina. Alguns modelos se beneficiam do fato de os preditores estarem em uma escala comum (KUHN; JOHNSON, 2013).

Além destas transformações, a variável desfecho (probabilidade de morte adulta) inicialmente uma variável contínua foi transformada em uma variável de duas classes usando o valor médio da probabilidade de morte adulta, ou seja, as microrregiões que apresentavam

probabilidade de morte adulta abaixo de 0,1406 (ponto médio) foram classificadas como classe P0 e o restante como classe P50. Esta abordagem foi escolhida para melhorar a performance dos algoritmos de aprendizado de máquina como *Support Vector Machine* e *Random Forests* que possuem alta robustez à alta dimensionalidade dos dados. O estudo de Lustgarten et al. (2008) mostra que a transformação da variável contínua em duas classes pode ajudar a melhorar significativamente o desempenho de classificação desses algoritmos, bem como algoritmos como *Naïve Bayes*, que são sensíveis à dimensionalidade dos dados.

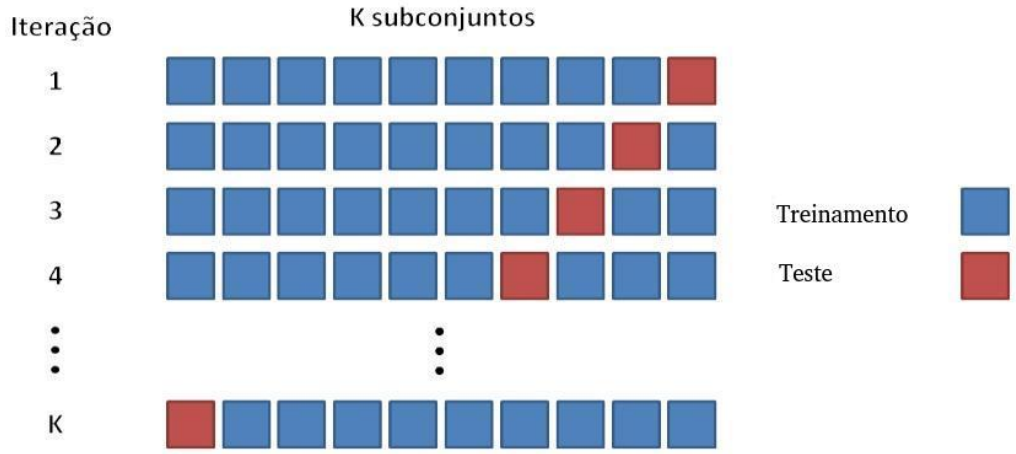
2.3.2 Ajustes do modelo e reamostragem

Existem alguns tipos de técnicas de reamostragem. De forma geral, possuem uma ideia em comum: um subconjunto de amostras é usado para ajustar um modelo e as amostras restantes são usadas para estimar a eficácia do modelo. Esse processo é repetido várias vezes e os resultados são agregados e resumidos. As diferenças nas técnicas geralmente se concentram em torno do método em que as subamostras são escolhidas. Dentre as técnicas existentes, tem-se: Validação Cruzada Generalizada, Validação Cruzada *Leave One Out* (deixa um de fora), Validação Cruzada por k-subconjuntos, *Bootstrap,* entre outras (KUHN; JOHSON, 2013; JAMES et al., 2017).

De acordo com Kuhn e Johnson (2013), nenhum método de reamostragem é uniformemente melhor que o outro; a escolha deve ser feita considerando vários fatores. Se o tamanho das amostras é pequeno, a validação cruzada k vezes é recomendada por várias razões: as propriedades de viés e de variância são boas e, dado o tamanho da amostra, os custos computacionais não são grandes.

Pelos motivos apresentados acima, usou-se a validação cruzada em k-subconjuntos para avaliar os modelos. A validação cruzada por k-subconjuntos é um procedimento no qual a amostra é dividida em k subamostras de tamanho igual. Os modelos são treinados em subconjuntos de k – 1 (o conjunto de treinamento) e então testados por meio do conjunto que sobra (o conjunto de testes) para avaliar o desempenho de previsão do modelo, iterando por meio de cada uma das k amostras (Figura 7). A escolha de k é geralmente 5 ou 10. Foi escolhido para os quatro algoritmos analisados neste trabalho k=10 (KUHN; JOHSON, 2013; JAMES et al., 2017).

FIGURA 6 – Esquema de validação cruzada com k-subconjuntos



Fonte: Adaptado Kuhn e Johnson (2013).

Foi gerado para cada algoritmo um modelo com 10 reamostragens e a partir delas se agregou os resultados chegando a um único valor de acurácia, precisão, *recall*, etc para cada um dos quatro algoritmos que serão analisados no próximo capítulo.

2.3.3 Principais métricas utilizadas

A partir da matriz de confusão, como ilustrado pela Tabela 2, é possível calcular as principais medidas que serão usadas para a comparação dos algoritmos.

TABELA 2 – Exemplo de matriz de confusão para um problema de classificação

		OBSERVADO	
		CLASSE 1	**CLASSE 2**
PREDITO	**CLASSE 1**	Verdadeiro Positivo (a)	Falso Positivo (b)
	CLASSE 2	Falso Negativo (c)	Verdadeiro Negativo (d)

Fonte: Elaboração própria.

As linhas são as classes preditas e as colunas são as classes reais e os valores da diagonal principal são os valores que o modelo gerou como prognósticos corretos. Ou seja, os valores do primeiro quadrante (a) significam que o ponto observado do dado era da Classe 1 e o valor predito também foi da Classe 1, e os valores do último quadrante (d) seguem a mesma lógica, porém com a Classe 2.

Os valores de *b* indicam que o ponto observado era da Classe 2, porém o modelo fez a predição que era da Classe 1, enquanto que os valores de *c* significam que os pontos observados eram da Classe 1 e o modelo classificou como sendo da Classe 2.

A partir dos valores da matriz de confusão foram calculadas algumas métricas. São elas:

$$\text{Acurácia} = \frac{(a+d)}{(a+b+c+d)} \quad (1)$$

$$\text{Precisão} = \frac{a}{(a+b)} \quad (2)$$

$$\text{Sensibilidade} = \frac{a}{(a+c)} \quad (3)$$

$$\text{Especificidade} = \frac{d}{(b+d)} \quad (4)$$

$$\text{F1} = \frac{2 \text{ x Precisão x Sensibilidade}}{\text{Precisão + Sensibilidade}} \quad (5)$$

$$\text{Kappa} = \frac{(\text{acurácia observada} - \text{acurácia esperada})}{(1 - \text{acurácia esperada})} \quad (6)$$

A Acurácia (Equação 1) determina os números de predições feitas corretamente pelo modelo sobre todas as predições feitas. Já a medida de Precisão (Equação 2) nos diz qual a proporção de observações da Classe 1 que foram classificadas como tal que realmente eram da Classe 1 e Sensibilidade (Equação 3) calcula qual a proporção de observações que realmente eram da Classe 1 que foram diagnosticados pelo algoritmo como sendo da Classe 1, ou seja, é a proporção de verdadeiros positivos (AWAD; KHANNA, 2015). F1 é uma medida da média ponderada das métricas de Precisão e Sensibilidade. Portanto, essa pontuação leva em conta tanto os falsos positivos quanto os falsos negativos (AWAD; KHANNA, 2015).

A Especificidade (Equação 4) demonstra as observações definidas pelo algoritmo como sendo da Classe 2 que realmente eram da Classe 2 (AWAD; KHANNA, 2015). A Estatística Kappa é uma medida que compara a acurácia observada com a esperada (chance aleatória). Ela é usada não apenas para avaliar um único classificador, mas também para avaliar os classificadores entre si (AWAD; KHANNA, 2015).

2.4 Funcionamento dos algoritmos

Para a implementação dos algoritmos se utilizou as funções TrainControl e Train do pacote *caret* usando o *Software* RStudio versão 3.5[2]. Esse pacote permite implementar inúmeros algoritmos de aprendizado de máquina e as principais técnicas de reamostragem incluindo validação cruzada por k-subconjuntos explicada anteriormente.

2.4.1 Análises de Componentes Principais (ACP), *T-distributed Stochastic Neighbor Embedding (T-SNE)* e *K-means*

A análise de componentes principais (ACP) se refere ao processo pelo qual os componentes principais são gerados e usados no entendimento dos dados. A ACP é uma abordagem não supervisionada, pois envolve apenas um conjunto de recursos X_1, X_2, [...], Xp e nenhuma resposta associada Y. A ACP serve como uma ferramenta para visualização de dados (visualização das observações ou visualização das variáveis) (JAMES et al., 2017). No presente estudo foi utilizada como forma de entender e visualizar a interação das variáveis preditoras.

Os componentes principais permitem reduzir um conjunto de variáveis correlacionadas em um número menor de variáveis representativas que explicam coletivamente a maior parte da variabilidade no conjunto original (KUHN; JOHSON, 2013; JAMES et al., 2017).

Para conjuntos de dados com muitas variáveis preditoras, é possível estabelecer quantos componentes devem ser retidos por meio de um gráfico de *scree* que contém o número do componente ordenado (eixo x) e a quantidade de variabilidade resumida (eixo y). Para a maioria dos conjuntos de dados, os primeiros componentes principais resumem a maior parte da variabilidade e o gráfico mostra uma descida íngreme; a variação será reduzida para os componentes restantes (KUHN; JOHSON, 2013).

Além de gerar os componentes principais, a partir da análise é possível caracterizar os preditores associados a cada componente. Cada componente é uma combinação linear (Equação 7) dos preditores e o coeficiente de cada preditor é chamado de *Loading*. Ou seja, P é o número de preditores e os coeficientes aj_1, aj_2, [...], aj_P são chamados de pesos do componente e auxiliam no entendimento de quais preditores são mais importantes para cada componente principal. L*oadings* próximos de zero mostram que a variável preditora não contribuiu de forma significativa para o componente em questão (KUHN; JOHSON, 2013; JAMES et al. 2017).

[2] Consultar anexo.

$$\mathrm{PC}_j = (a_{j1} \times \text{Predictor 1}) + (a_{j2} \times \text{Predictor 2}) + \cdots + (a_{jP} \times \text{Predictor } P) \quad (7)$$

Juntamente com a apresentação de algumas análises feitas com a Análise de Componentes Principais, a Incorporação de Vizinhos Estocásticos Distribuídos em T (t-SNE) foi apresentada. O t-SNE constitui uma técnica não linear para redução de dimensionalidade que é particularmente adequada para a visualização de conjuntos de dados de alta dimensão (MAATEN; HINTON, 2008).

Em termos simples, esta técnica (t-SNE) minimiza a divergência entre duas distribuições: uma distribuição que mede semelhanças entre pares dos objetos de entrada e uma distribuição que mede semelhanças entre pares dos correspondentes pontos de baixa dimensionalidade na incorporação (MAATEN; HINTON, 2008).

Dessa forma, o t-SNE mapeia os dados multidimensionais para um espaço dimensional menor e tenta encontrar padrões nos dados, identificando agrupamentos observados com base na similaridade de pontos de dados com vários preditores. Foi usada nas análises principalmente como uma técnica de exploração e visualização de dados (MAATEN; HINTON, 2008).

As saídas do algoritmo t-SNE foram usadas para a construção de grupos usando outro algoritmo, o *K-Means* que facilita a análise visual dos dados e cria agrupamentos levando em consideração atributos similares entre os dados (JAMES et al., 2017). O algoritmo *K-means* particiona um conjunto de dados em K agrupamentos distintos e não sobrepostos. Para executá-lo, primeiro se especificou o número desejado de agrupamentos K; então o algoritmo K-Means atribuirá a cada observação a exatamente um dos *clusters* pré-definidos (JAMES et al., 2017).

2.4.2 Support Vector Machine

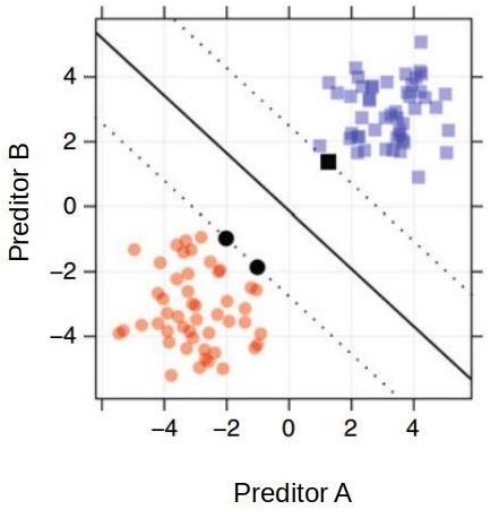
Support Vector Machine (SVM) representa as instâncias como um conjunto de pontos de 2 tipos no lugar N dimensional e gera um hiperplano dimensional (N – 1) para separar esses pontos em 2 grupos. O SVM tenta encontrar um hiperplano de separação que divide esses pontos em dois tipos e está o mais distante possível de todos esses pontos (SAKR et al., 2017).

A métrica principal do algoritmo chama-se margem que pode ser definida como a distância entre o limite de classificação e o ponto de ajuste mais próximo do dado de treinamento. A Figura 7 abaixo mostra um possível limite de classificação como uma linha sólida. As linhas tracejadas em ambos os lados do limite estão na distância máxima da linha

até os dados do conjunto de treinamento mais próximo. Neste exemplo, os três pontos de dados estão igualmente próximos do limite de classificação e são destacados com símbolos pretos sólidos. A margem definida por esses pontos de dados pode ser quantificada e usada para avaliar possíveis modelos (KUHN; JOHSON, 2013).

Em linhas gerais, o algoritmo tenta criar uma fronteira que melhor separa os dados calculando os vetores de suporte (pontos pretos na figura abaixo) e eles são assim definidos, pois são os pontos mais próximos das linhas, calculando a distância entre o ponto quadrado e ponto redondo e é chamada de margem.

FIGURA 7 – Esquema geral de funcionamento do algoritmo SVM



Fonte: Adaptado de Kuhn e Johnson (2013).

No presente estudo, o método utilizado foi o Support Vector Machine Radial, ou seja, o modelo foi ajustado usando uma transformação não linear chamada função de base Radial permitindo que o algoritmo produza limites de decisão extremamente flexíveis e, assim, foi possível melhorar a performance e números de acertos.

2.4.3 Naive Bayes

Naive Bayes examina a distribuição condicional das variáveis preditoras em cada classe para determinar se há diferenças entre as classes, tratando todos os preditores como independentes (PAN et al., 2017; SAKR et al., 2017).

Neste sentido, o método supõe que um preditor não tem nenhuma relação com algum dos outros preditores do modelo. Este é baseado no Teorema de Bayes em que tenta responder a seguinte pergunta: “baseado nos preditores X que foram observados, qual a

probabilidade da variável desfecho Y ser de determinada Classe Cl?" (KUHN; JOHNSON, 2013, p. 354,).

Ou seja, o algoritmo estima $Pr[Y=Cl|X_n]$, isto é, dado um determinado valor de uma variável preditora X_n, qual a probabilidade da variável desfecho ser da Classe Cl) (KUHN; JOHSON, 2013; JAMES et al., 2017).

Na análise proposta, significa dizer que dado uma variável preditora listada na Tabela 1 qual a probabilidade de ser da Classe P0 ou de ser da Classe P50? As probabilidades de cada classe da variável são criadas e a classe prevista é aquela associada à maior probabilidade da classe. A essência do modelo é a determinação das probabilidades condicionais e incondicionais associadas aos preditores (KUHN; JOHSON, 2013; JAMES et al., 2017).

2.4.4 Árvores de decisão: Random Forest e Extreme Gradient Boosting

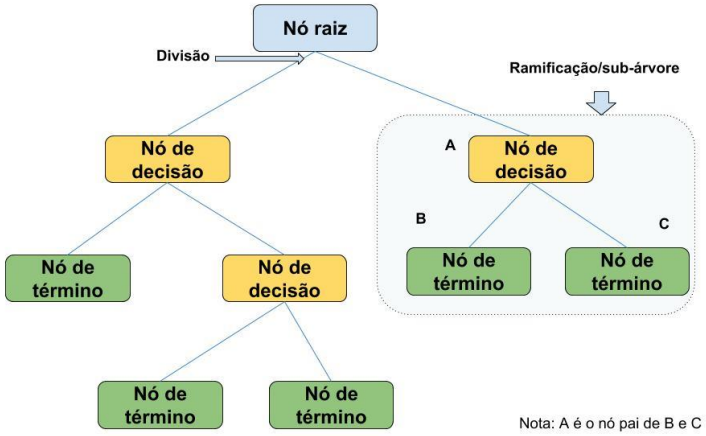
As árvores de decisão, de forma geral como mostra a Figura 8, constroem várias árvores com base nos dados e usa as decisões de cada árvore para alcançar uma única (PAN et al. 2017; NGUYEN, 2016).

No método *Random Forest*, as árvores de decisão são construídas tomando partições aleatoriamente reamostradas dos dados e usando um subconjunto aleatório de variáveis potenciais para informar cada divisão dos dados. Ao criar amostras aleatórias com as observações e determinando possíveis variáveis de divisão, esse método introduz um elemento de aleatoriedade que impede o sobreajuste do modelo. Depois de treinar todas as árvores, elas são combinadas com base nas previsões de cada árvore individual (JAMES et al., 2017).

É necessário, no algoritmo *Random Forest*, avaliar a importância de cada preditor para que seja possível identificar as variáveis relevantes na construção de cada árvore. O parâmetro de ajuste deste algoritmo é o número de preditores selecionados aleatoriamente, k, para escolher em cada divisão e é comumente referido como *mtry*. A escolha das divisões ótimas de cada nó de decisão usualmente é feita usando o critério chamado de Índice de Gini usado para avaliar a distribuição das classes da variável desfecho em cada nó. Essa divisão é necessária para que os nós filhos sejam mais puros. O critério de melhoria aqui usado, o índice de Gini, é agregado em todo o conjunto para gerar uma medida geral de importância das variáveis (KUHN; JOHSON, 2013).

Uma amostra nova de k preditores é obtida a cada divisão e, normalmente, escolhemos k ≈√ p – ou seja, o número de preditores considerados em cada divisão é aproximadamente igual à raiz quadrada do número total de preditores (KUHN; JOHSON, 2013; JAMES et al., 2017).

FIGURA 8 – Esquema geral de árvore de decisão



Fonte: Adaptado de Kuhn e Johnson (2013).

Escolheu-se para o ajuste do modelo de Random Forest testado neste estudo, o número de 500 árvores e o parâmetro *mtry* igual a 3.

Os algoritmos chamados de Gradient Boosting também são baseados em árvores de conjunto semelhantes à Random Forest, porém os XGB começam com árvores de aprendizagem fracas e constroem árvores mais confiáveis com base nos resíduos das previsões de cada árvore anterior. Isso difere na abordagem de RF, em que as árvores são independentes umas das outras e dependem da aleatoriedade na seleção de variáveis e observação para produzir estimativas imparciais (NGUYEN, 2016).

No ajuste do modelo de XGB usando as variáveis preditoras apresentadas na Tabela 1, os parâmetros utilizados foram: *nrounds =150; subsample= (0,50; 0,75;1,00); eta = (0,3; 0,4); max-depth = (1,2,3).* O parâmetro *eta* é usado para evitar o sobreajuste do modelo. O max-depth define o máximo de profundidade de uma árvore. Já o *subsample* relaciona a subamostra em cada treinamento, por exemplo, uma subamostra definida como 0,5 significa que o modelo usa como amostra aleatoriamente metade dos dados de treinamento antes de criar as árvores e assim evitar sob reajuste.

Sendo assim, a presente pesquisa aplicou os quatro algoritmos explicados anteriormente para determinar o grau de importância de cada variável preditora em relação a probabilidade de morte adulta, ou seja, as análises partem da hipótese de que as condições de

vida nas microrregiões influenciam no risco de morte de indivíduos de 15 a 59 anos observando quais as variáveis que mais afetam esta mortalidade dividida em duas classes (P0 e P50), que significam, menor e maior probabilidade de morte, respectivamente.

Pretende-se entender e analisar como as variáveis de cobertura do Programa Saúde da Família, de água encanada, coleta de lixo, esgoto, vacinação, número de leitos, PIB per capita, taxa de analfabetismo, renda média domiciliar, taxa de desemprego, proporção de indivíduos com curso superior, taxa de trabalho infantil, razão de sexo, proporção de indivíduos de cor branca, de cor preta e parda, grau de urbanização, proporção de jovens de 15 a 29 anos, taxa de pobreza, índice de envelhecimento, taxa de mortalidade por causas externas e por doenças do aparelho circulatório determinam mortalidade adulta mais alta (P50) e mais baixa (P0).

CAPÍTULO 3 – RESULTADOS E DISCUSSÃO

3.1. Análise descritiva

As variáveis escolhidas refletem características socioeconômicas das microrregiões. Entender a distribuição delas ajuda a entender o problema de forma ampla e torna-se necessário para as primeiras etapas de modelagem para os algoritmos. Ou seja, entender a distribuição ajuda a uma melhor decisão das formas de particionar os dados (KUHN; JOHNSON, 2013; JAMES et al., 2017).

O Teorema do "Sem Almoço Grátis" (WOLPERT, 1996) argumenta que, sem ter informações substanciais sobre o problema de modelagem, não existe um modelo único que sempre será melhor que qualquer outro modelo. Por causa disso, um forte argumento pode ser feito para tentar uma ampla variedade de técnicas e, em seguida, determinar qual modelo focar (KUHN; JOHNSON, 2013).

As variáveis preditoras podem seguir uma distribuição simétrica ou assimétrica (para preditores contínuos). Os preditores dentro de um conjunto de dados podem ou não ter um relacionamento subjacente com a resposta. Diferentes tipos de algoritmos lidam com a assimetria das variáveis preditoras de diferentes formas (KUHN; JOHNSON, 2013; JAMES et al., 2017).

É necessário se ter uma compreensão detalhada dos preditores e da variável de interesse de qualquer conjunto de dados antes de tentar criar um modelo. A falta de compreensão pode levar a dificuldades computacionais e desempenho inferior ao modelo ideal. Além disso, a maioria dos conjuntos de dados exigirá algum grau de pré-processamento para expandir o universo de possíveis modelos preditivos e otimizar o desempenho de cada modelo (KUHN; JOHNSON, 2013; JAMES et al., 2017).

O Brasil é um país de grande heterogeneidade regional e desigualdades em que convivem diferentes níveis ao que concerne aos indicadores de saúde, demográficos e socioeconômicos. A maioria das microrregiões brasileiras possuem uma cobertura de água e esgoto com mais de 90% dos domicílios, sendo que algumas microrregiões apresentam baixas coberturas (Tabela 3).

Quando nos referimos os indicadores de escolaridade como taxa de analfabetismo e proporção de indivíduos com curso superior se apreende que o país possui grande diversidade de níveis, em que se convive microrregiões com altas taxas de analfabetismo e baixa proporção de indivíduos com curso superior (Tabela 3). A média de indivíduos que cursaram pelo menos o curso superior em 2010 foi de 8,3%, sendo que a Região Sudeste

apresentou os maiores níveis de instrução (IBGE, 2010a). Em relação à taxa de analfabetismo, o país possui uma média de 9,6%, em 2010, porém distribuído de forma desigual, em que a Região Nordeste apresenta 28% de pessoas analfabetas contra 5% nas regiões Sudeste e Sul (IBGE, 2010a).

Os indicadores socioeconômicos – taxa de pobreza, taxa de trabalho infantil, renda média domiciliar per capita, PIB per capita e taxa de desemprego - também apresentam diferenças regionais importantes, sendo que a média do Brasil em relação a taxa de pobreza foi 7% chegando a 30% na microrregião de Traipu, no estado de Alagoas, localizada na Região Nordeste do Brasil enquanto que as menores taxas estão localizadas nas microrregiões da Região Sul e Sudeste.

A taxa de desemprego possui um valor mínimo de 1,3% da população economicamente ativa que se encontrava sem emprego, com uma média de 6,93% e um valor máximo de 20,47%. As regiões Norte e Nordeste apresentaram a maior concentração de microrregiões com maiores taxas de desemprego, apesar da região Sudeste, Sul e Centro-Oeste apresentarem algumas microrregiões com alta taxa, porém em menor proporção que na Região Norte e Nordeste. A variável PIB per capita (indica o nível de produção econômica de uma determinada microrregião brasileira em 2010) possui um valor mínimo de R$ 2.867,00 per capita, média de R$ 14.105,00 per capita e valor máximo de R$ 77.872,00 per capita.

O PIB do estado de São Paulo, Minas Gerais e o Rio de Janeiro correspondiam em 2010 há 50% da participação do PIB do país, enquanto que dezoito estados juntos respondiam por 22,2% do PIB (SANTOS; PALES; RODRIGUES, 2014). Além disso, as menores rendas médias per capita se encontravam nos estados das regiões Norte e Nordeste, abaixo de R$ 600,00.

Na parte relacionada à saúde, as maiorias das microrregiões possuem cobertura do Programa Saúde da Família acima de 70% e o número de leitos acima 2 por mil habitantes. A Organização Mundial de Saúde (OMS) define como o número ideal de leitos disponíveis como sendo entre 3 a 5, ou seja, a média brasileira está abaixo da orientação da OMS para a maioria das microrregiões. Sendo que as poucas microrregiões que possuem acima de 4 leitos por mil habitantes estão todas localizadas nos estados de Minas Gerais, São Paulo, Rio de Janeiro, Paraná e Rio Grande do Sul.

Em relação as variáveis demográficas, o grau de urbanização possui uma média acima de 70%, a proporção de jovens de 15 a 29 anos fica entre 20% e 30% em todas as microrregiões e o índice de envelhecimento possui uma distribuição ampla, sendo que o

Brasil apresentou, em média, um índice de envelhecimento aproximadamente de 45 pessoas com mais de 60 anos para cada 100 com menos de 15 anos (CLOSS; SCHWANKE, 2012).

TABELA 3 – Resumo das estatísticas das variáveis preditoras

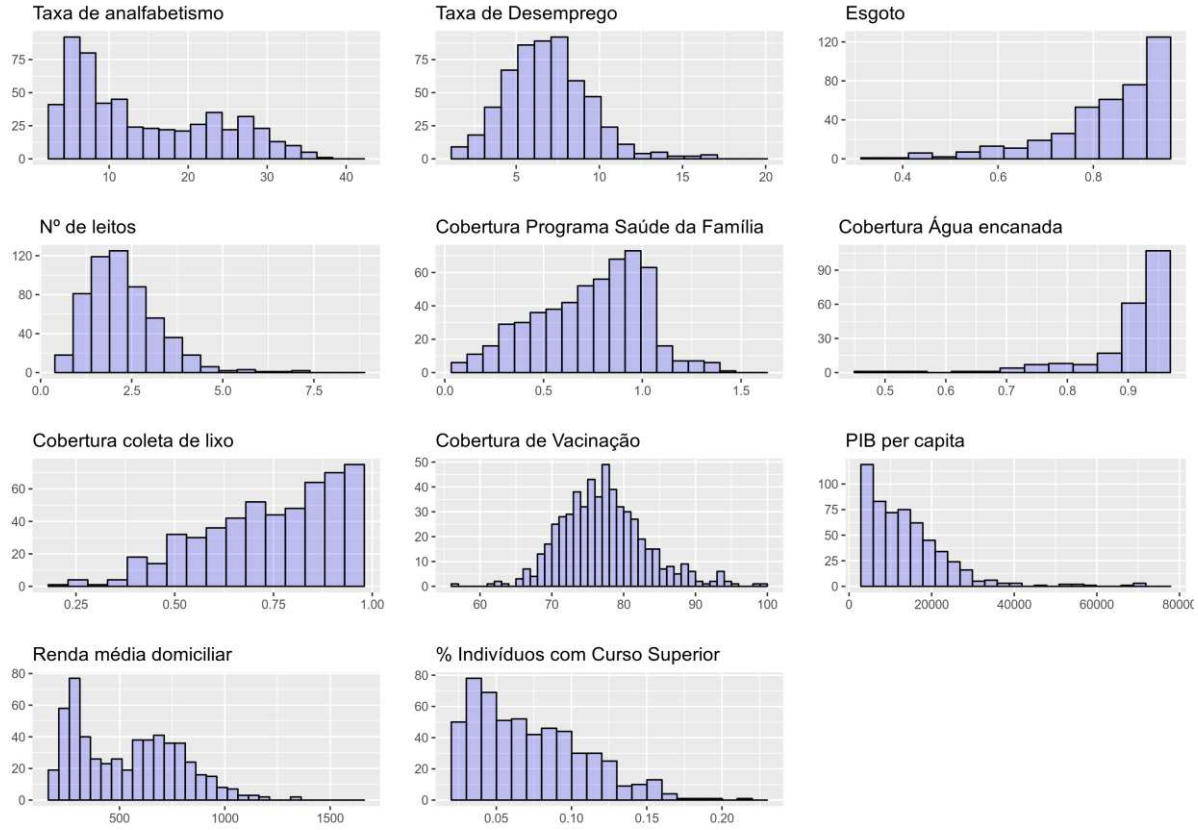
	Valores					
Variáveis	**Mínimo**	**1º Quartil**	**Mediana**	**Média**	**3º Quartil**	**Máximo**
cob_PSF	0,03	0,55	0,77	0,74	0,96	1,69
cob_agua	0,45	0,94	0,99	0,96	1	1
cob_lixo	0,18	0,63	0,79	0,76	0,9	1
cob_esg	0,31	0,81	0,91	0,87	0,97	1
cob_vacina	56	73,25	76,95	77,3	80,53	100,24
n_leitos	0,39	1,57	2,12	2,29	2,82	9,34
pib_percapita	2867	6565	12105	14105	18247	77872
tx_analf	2,3	6,43	11,25	14,43	22,48	42,4
renda	162	302	542	542	729	1665
tx_desemp	1,1	5,15	6,74	6,93	8,37	20,47
cursosup	0,02	0,04	0,07	0,07	0,1	0,24
tx_trabinf	3,31	8,02	10,79	11,51	14,15	31,27
rz_sexo	88	97	99	100	102	114
negros	0,07	0,35	0,58	0,53	0,71	0,87
brancos	0,08	0,26	0,4	0,45	0,64	0,92
grau_urb	22,34	58,48	72,72	72,01	86,32	100
prop_jovem	0,21	0,25	0,27	0,27	0,28	0,32
tx_pobreza	0	0,02	0,04	0,07	0,13	0,3
indice_envelhecimento	8,47	30,89	41,82	42,42	53,37	87,76
tx_causasexternas	0	56,55	68,99	70,88	83,13	154,87
tx_aparelhocirc	8,05	135,39	171,45	169,78	204,59	342,55

Fonte: IBGE (2010a); Brasil (2010).

Por sua vez, os histogramas apresentados nos Gráficos 3 e 4 ilustram que as variáveis são, em sua maioria, assimétricas. Uma distribuição não distorcida é aquela que é aproximadamente simétrica. Isso significa que a probabilidade de cair em ambos os lados da média da distribuição é aproximadamente igual. Uma distribuição assimétrica à direita tem um grande número de pontos no lado esquerdo da distribuição (valores menores) do que no lado direito (valores maiores).

No Gráfico 3 observa-se que as variáveis Taxam de analfabetismo, PIB per capita, proporção de indivíduos com Curso Superior e número de leitos por mil habitantes possuem distribuição assimétrica à esquerda, ou seja, a maioria das microrregiões possuem valores menores concentrados à esquerda e algumas microrregiões possuem valores maiores. A variável cobertura de esgoto, água encanada e lixo possuem distribuição concentrada à direita e há uma maior concentração de microrregiões com valores altos, isto é, com cobertura maior do que 50% nos domicílios das microrregiões.

GRÁFICO 3 – Histogramas das variáveis – Microrregiões, Brasil, 2010



Fonte: IBGE (2010a); Brasil (2010).

Pode-se perceber que a variável Renda média domiciliar per capita possui dois grandes grupos, um com microrregiões que possuem renda entre R$ 162,00 e R$ 500,00 e outro grupo com microrregiões entre R$ 500,00 a R$ 1000,00.

Já cobertura de vacinação possui uma distribuição razoavelmente simétrica, assim como as variáveis Razões de Sexo, Índice de envelhecimento, proporção de jovens de 15 a 29 anos, taxa de mortalidade por causas externas, por doenças do aparelho circulatório e a variável de interesse ${}_{45}q_{15}$ apresentadas no Gráfico 4.

O Brasil pode ser considerado um país envelhecido, de acordo com o definido por Shryock & Siegel. Para eles, um valor de índice de envelhecimento menor que 15 somos indicativos de uma população jovem, de 15 a 30 intermediários e acima de 30 uma população idosa. Este processo não ocorre de forma igual para todas as regiões (CHAIMOWICZ, 1997; CLOSS; SCHWANKE, 2012). Há microrregiões que possuem valor menor que 15, portanto possuem uma população jovem; um grupo de microrregiões que possui entre 20 e 40 de índice e, na Região Sudeste e Sul, em sua maioria possui microrregiões com acima de 30.

Portanto, nas regiões consideradas mais desenvolvidas há um índice de envelhecimento maior e estão em uma fase mais avançada da transição demográfica.

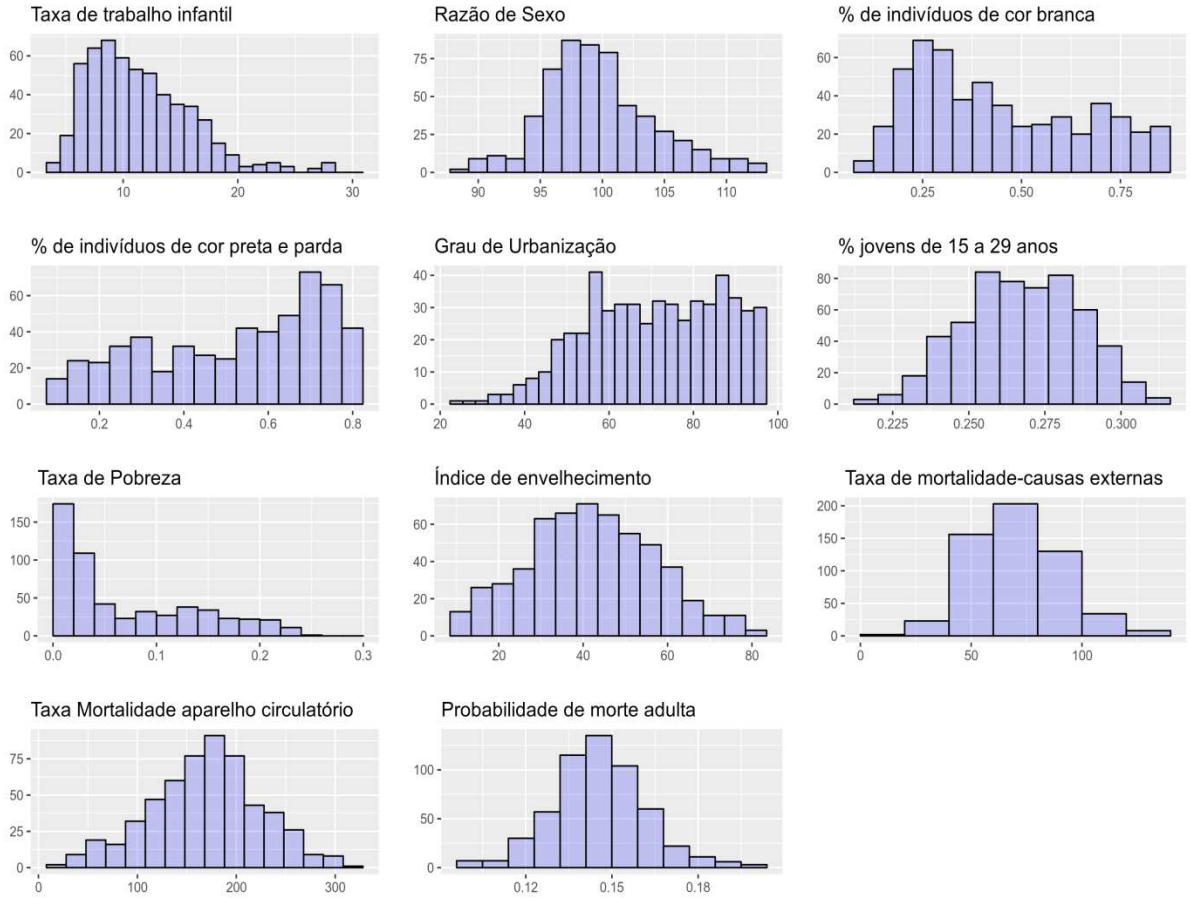
Ademais, as mortes relacionadas às doenças do aparelho circulatório são a primeira causa de morte ao se considerar todo o grupo etário de 15 a 59 anos no Brasil, apesar de quando se olha somente a faixa de 15 a 39, a primeira causa de morte entre os mais jovens se concentra nas causas externas (CESSE et al., 2009; MEDEIROS; MENEGHEL; GERHARDT, 2012).

A variável Taxa de trabalho infantil e taxa de pobreza têm distribuição assimétrica à esquerda, ou seja, a maioria das microrregiões possuem valores pequenos destas taxas. Por último, temos a proporção de indivíduos de cor branca, de cor preta e parta e grau de urbanização que são assimétricas e bimodais, isto é, se identifica dois grupos de valores.

De forma geral, observa-se a partir dos dados apresentados, que o Brasil possui um conjunto de microrregiões menos urbanizadas e outro conjunto com maior urbanização. O grau de urbanização total, em 2010, no Brasil era de 84% e a concentração era mais acentuada nas Regiões Sudeste, Sul e Centro-Oeste (MOURA; OLIVEIRA; PÊGO, 2018).

Saliente-se ainda que, a proporção de indivíduos que se autodeclaram pretos e pardos no país constitui a metade da população, 50,7% (IBGE, 2010b; IPEA, 2011). Entretanto, a proporção de pretos e pardos não está distribuída de forma igual no país. Por exemplo, os estados da Bahia, Maranhão e Pará são os estados com maior proporção de pretos e pardos, 76% da população total em média. De forma contrária se encontram os estados de São Paulo, Minas Gerais e Rio de Janeiro que possuem uma proporção de pretos e pardos de apenas, em média, 30% (IPEA, 2011).

GRÁFICO 4 – Histogramas das variáveis – Microrregiões, Brasil, 2010



Fonte: IBGE (2010a); Brasil (2010).

Observa-se no Gráfico 5 e 6 que a maioria das variáveis preditoras possuem uma relação não linear com a probabilidade de morte adulta. A Taxa de Mortalidade por Causas Externas ilustrada no Gráfico 5 apresenta uma relação linear com a probabilidade de morte adulta, ou seja, microrregiões com baixas taxas de mortalidade por causas externas apresentam baixa probabilidade de morte adulta e no outro extremo, as microrregiões que apresentam alta probabilidade de morte adulta exibem uma maior taxa de mortalidade por causas externas. A Taxa de Desemprego, da mesma forma, tem uma relação linear com a probabilidade de morte adulta. Em outras palavras, quanto maior a taxa de desemprego de uma determinada microrregião maior se apresenta a probabilidade de morte adulta (Gráfico 6).

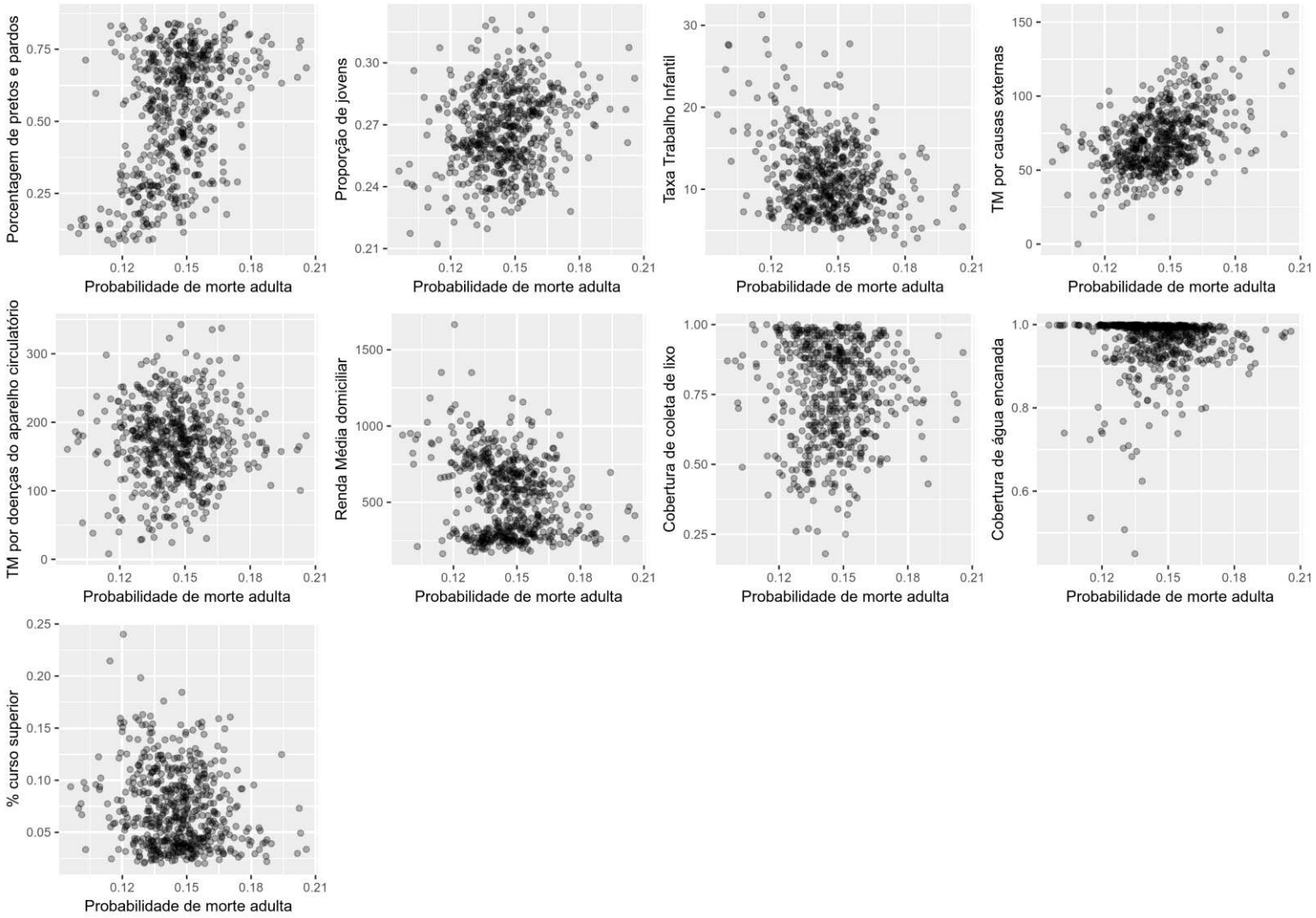
Observa-se também comportamentos diferenciados da variável de cobertura de água encanada e taxa de pobreza. Em relação a cobertura de água, as microrregiões estão concentradas nos valores mais próximos do máximo de cobertura, porém apresentam diferentes níveis de probabilidade de morte adulta (Gráfico 5). Neste mesmo sentido, a relação da taxa de pobreza (Gráfico 6) mostra que as microrregiões possuem valores de taxa

de pobreza concentradas em valores mais baixos, mas apresentam probabilidade de morte adulta com valores bem diferentes. Ou seja, microrregiões com valores baixos de taxa de pobreza, algumas apresentam probabilidade de morte adulta baixa e outras, mortalidade adulta alta.

Com exceção destas, todas as outras possuem uma distribuição aleatória com nenhuma relação aparente com a variável probabilidade de morte adulta. Esse resultado reforçou a decisão da não realização de ajustes de modelos lineares, pois a maioria das variáveis preditoras não possuíam uma relação linear com a variável probabilidade.

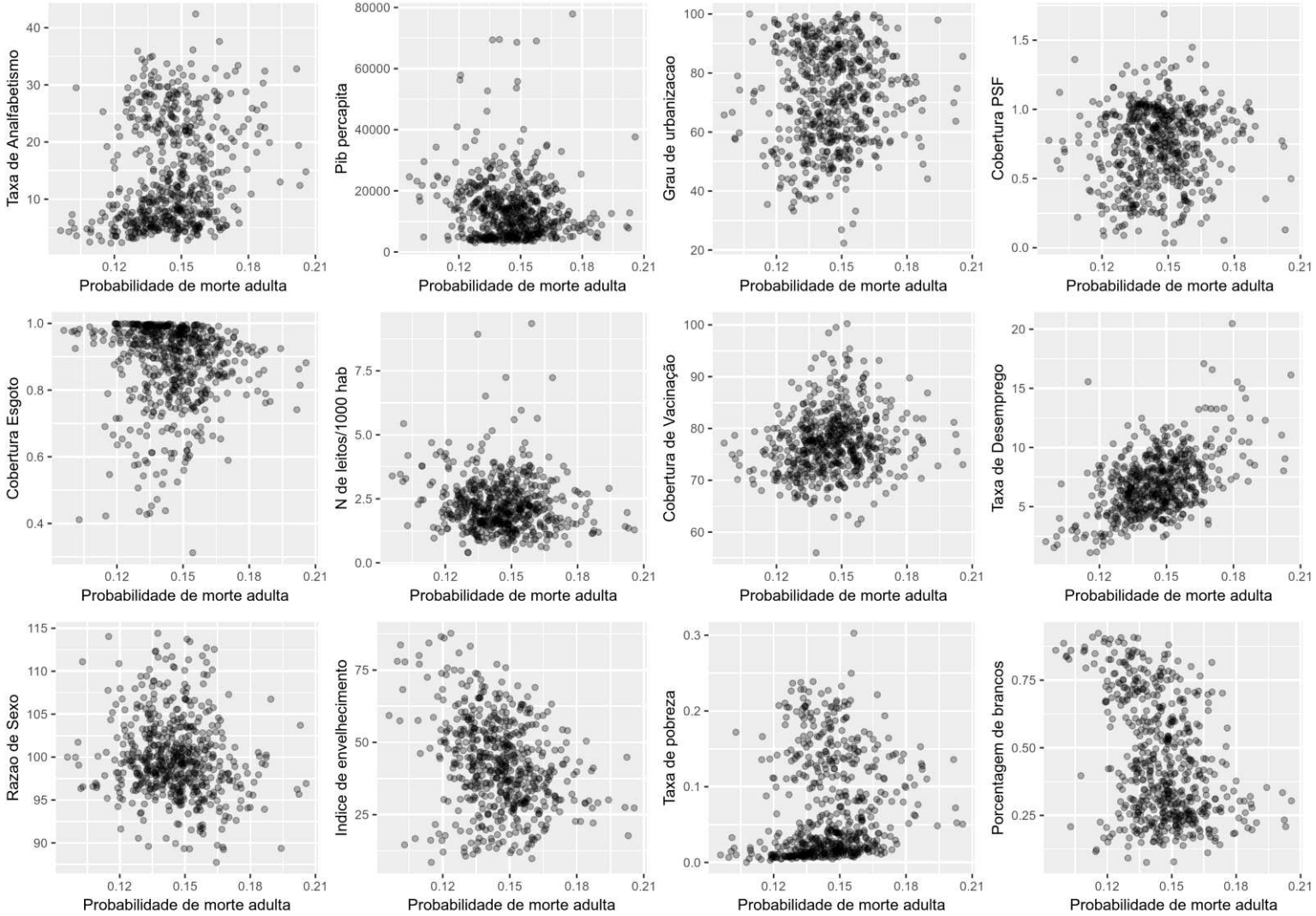
É importante destacar que todas as variáveis preditoras apresentadas foram centralizadas e transformadas para terem a mesma escala para que fosse possível extrair o melhor desempenho possível dos modelos analisados.

GRÁFICO 5 – Relação das variáveis preditoras com a variável desfecho – Brasil, 2010



Fonte: IBGE (2010a); Brasil (2010).

GRÁFICO 6 – Relação das variáveis preditoras com a variável desfecho – Brasil, 2010



Fonte: IBGE (2010a); Brasil (2010).

Além da investigação da relação da probabilidade de morte adulta com cada uma das variáveis individualmente, foi analisada a correlação entre cada uma das variáveis preditoras entre si. Colinearidade é o termo técnico para a situação em que um par de variáveis preditoras tem uma correlação substancial. Também é possível ter relações entre múltiplos preditores de uma vez (chamados multicolinearidade). A Figura 9 exemplifica a matriz de correlação com cada uma das variáveis preditoras apresentadas em pares. Cada correlação é colorida de acordo com sua magnitude.

A visualização dos dados da Figura 9 é simétrica, ou seja, as diagonais superior e inferior mostram informações idênticas. As cores azul-escuras indicam fortes correlações positivas, vermelho-escuro é usado para fortes correlações negativas e quanto mais próximo do branco significa que não há relação entre os preditores.

Existem alguns pequenos blocos de correlações fortes (azul e vermelho escuro) próximas da diagonal principal. Um dos blocos se ilustra no canto esquerdo onde estão variáveis ligadas às características socioeconômicas das microrregiões. Analisa-se que a proporção de indivíduos com curso superior está correlacionada positivamente com renda média domiciliar, cobertura de coleta de lixo e grau de urbanização. Além destas, taxa de pobreza e taxa de analfabetismo possuem correlação positiva entre si e taxa de mortalidade por doenças do aparelho circulatório e índice de envelhecimento também.

Entre as variáveis que possuem correlação negativa (vermelho escuro) estão a variável proporção de pretos e pardos e a variável proporção de brancos. Taxa de analfabetismo possui correlação negativa com a renda média domiciliar como também a taxa de pobreza.

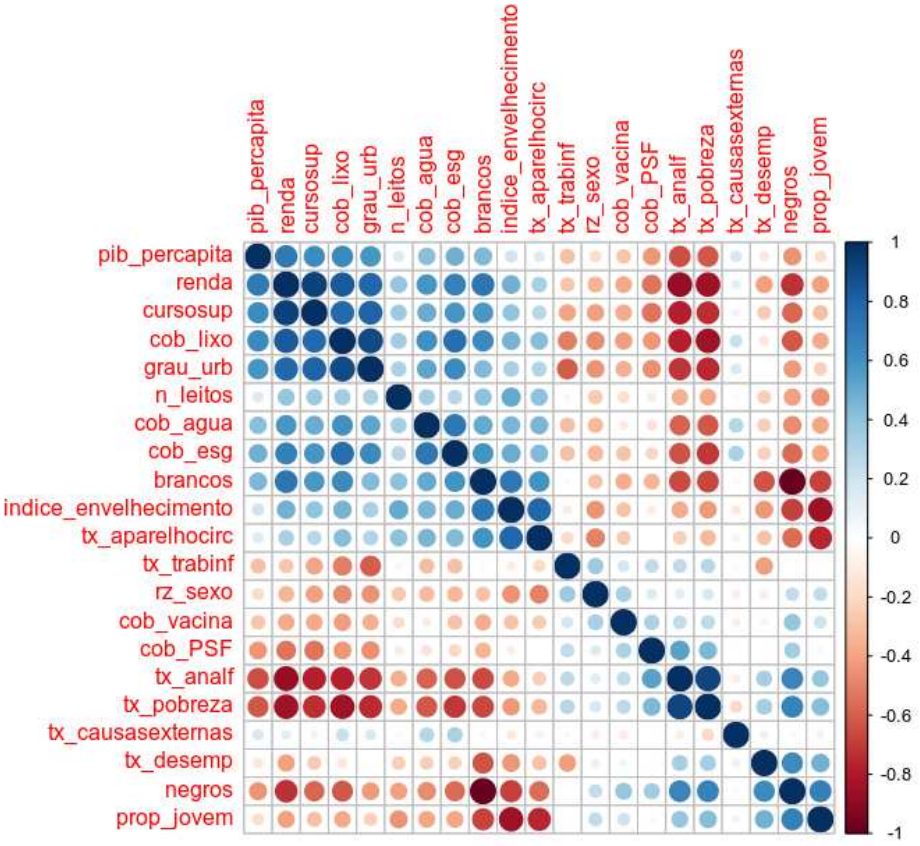
A relação entre a probabilidade de morte adulta e uma dada variável pode também depender dos valores das outras variáveis e neste sentido escolheu-se não remover nenhuma das variáveis inicialmente propostas.

De acordo com a matriz de correlação apresentada pela Figura 9, alguns blocos de correlação entre as variáveis são observáveis e neste sentido, para se entender de forma mais simplificada o problema proposto se aplicou uma Análise de Componentes Principais para compreender a interação entre essas variáveis. O resultado indica que a variável proporção de indivíduos de cor preta e parda possui alta correlação com a renda, taxa de analfabetismo e taxa de pobreza. Por sua vez, taxa de analfabetismo e taxa de pobreza apresentam, entre si, uma alta correlação positiva.

Do mesmo modo, a variável renda média per capita apresenta alta correlação negativa com as variáveis taxa de analfabetismo e taxa de pobreza e positiva com proporção

de indivíduos com curso superior, cobertura de lixo e grau de urbanização. Destaca-se que, a variável proporção de brancos se correlaciona de forma negativa com a proporção de jovens na população. Soma-se a este fato, a correlação positiva da proporção de jovens e a proporção de negros (pretos e pardos). De forma similar, o índice de envelhecimento possui correlação negativa com proporção de negros e positiva com proporção de brancos. Estas correlações entre jovens, envelhecimento e proporção por cor/raça pode estar associado ao fato de que a proporção de brancos acima de 65 anos no Brasil é de 57% enquanto que a proporção de negros no mesmo grupo etário é de 42% (IBGE, 2010a; OLIVEIRA; THOMAZ; SILVA, 2014). Além disso, a variável cobertura de lixo apresenta uma correlação média positiva com renda, proporção de indivíduos com curso superior, grau de urbanização e negativamente correlacionada com taxa de analfabetismo e taxa de pobreza.

FIGURA 9 – Matriz de correlação entre as variáveis preditoras



Fonte: IBGE (2010a); Brasil (2010).

A Análise de Componentes Principais (ACP) é um método utilizado para auxiliar no entendimento das relações mais significativas entre os preditores, sendo possível por meio dele analisar como cada preditor está associado com cada componente para identificar esses relacionamentos (KUHN; JOHSON, 2013; JAMES et al., 2017).

3.1.1 Interações entre as variáveis

A Tabela 4 mostra os autovalores, a porcentagem de variância que cada um dos componentes representa e o acumulado destas porcentagens. O primeiro componente representa 9,68% da variância amostral, ao se dividir a soma de todas as variâncias, obtém-se a porcentagem de variância explicada pelo primeiro componente em relação ao total, 46,10%. O segundo representa 2,83% da variância amostral e 13,47% da variância em relação ao total. Sendo assim, os dois primeiros componentes representam aproximadamente 60% da variância total.

TABELA 4 – Autovalores e porcentagens de variância explicada dos componentes principais

Componente	Autovalores	% Variância	% Acumulada
1	9,68	46,10	46,10
2	2,83	13,47	59,57
3	1,83	8,70	68,27
4	1,33	6,34	74,62
5	0,97	4,63	79,25
6	0,76	3,63	82,88
7	0,58	2,78	85,66
8	0,54	2,58	88,24
9	0,47	2,23	90,46
10	0,39	1,83	92,30
11	0,33	1,57	93,87
12	0,30	1,42	95,30
13	0,24	1,12	96,42
14	0,20	0,93	97,35
15	0,16	0,78	98,12
16	0,13	0,63	98,76
17	0,11	0,50	99,26
18	0,07	0,32	99,58
19	0,05	0,22	99,80
20	0,03	0,16	99,95
21	0,01	0,05	100,00

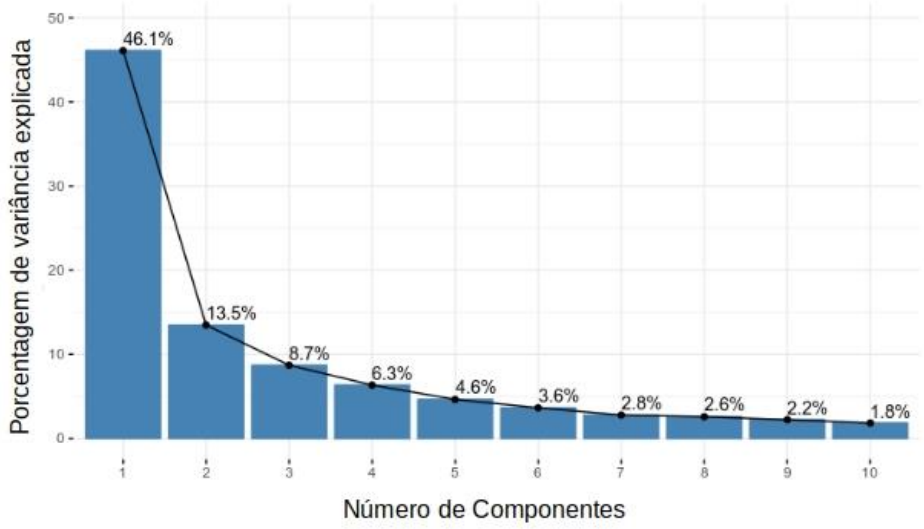
Fonte: IBGE (2010a); Brasil (2010).

O Gráfico 7 auxilia na escolha do número ótimo de componentes principais que representam a maior porcentagem de variância. Para melhor visualização, usa-se o critério de Kaiser que considera que o número de componentes principais deve ser todos os que possuem autovalores maiores que 1 (CARVALHO, 2013). Neste sentido, preservarão quatro componentes principais, os quais ilustram aproximadamente 75% da variância total.

Quando confrontados com um grande conjunto de variáveis correlacionadas, os componentes principais nos permitem resumir esse conjunto com um número menor de

variáveis representativas que explicam coletivamente a maior parte da variabilidade no conjunto original.

GRÁFICO 7 – Porcentagem da variância total explicada por cada componente



Fonte: IBGE (2010a); Brasil (2010).

Tendo em vista a escolha dos 4 componentes principais, a Tabela 5 apresenta o quanto cada variável se associa à determinado componente principal. Ou seja, em valores absolutos, o peso fatorial indica a associação de cada variável com o componente principal em questão.

Discriminam-se os pesos fatoriais em valores absolutos maiores que 0,25 para a escolha das variáveis para cada componente principal. Analisando cada componente principal, temos que:

- O primeiro é formado por 9 variáveis relacionadas, principalmente, a vulnerabilidade socioeconômica na região: cobertura de lixo, cobertura de esgoto, renda média domiciliar per capita, proporção de indivíduos com curso superior, proporção de indivíduos de cor preta e parda, proporção de brancos, grau de urbanização, taxa de analfabetismo e taxa de pobreza.
- O segundo é formado por 6 variáveis demográficas: taxa de desemprego, taxa de trabalho infantil, grau de urbanização, proporção de jovens, índice de envelhecimento e taxa de mortalidade por doenças do aparelho circulatório.

- O terceiro componente é formado por 5 variáveis que podem ser consideradas *proxy* da condição econômica e social da população: taxa de analfabetismo, taxa de desemprego, taxa de trabalho infantil, razão de sexo e taxa de mortalidade por doenças do aparelho circulatório.
- O quarto e último componente é formado principalmente por variáveis relacionadas ao acesso a saúde e fatores ambientais que afetam à saúde: cobertura do programa saúde da família, cobertura de água encanada, cobertura de esgoto, cobertura de vacinação e taxa de mortalidade por causas externas.

O primeiro componente representa quase metade da variância e pode-se inferir que as variáveis cobertura de lixo, cobertura de esgoto, curso superior, renda e proporção de brancos, que possuem sinal negativo, contribuem para o componente de forma contrária às variáveis de valor positivo, no caso, taxa de analfabetismo, taxa de pobreza e proporção de negros, ou seja, quanto menores os valores das variáveis com pesos fatoriais negativos e maiores os valores das variáveis com pesos positivos, maior será a vulnerabilidade caracterizada por este, e como resultado se tem uma maior probabilidade de mortalidade adulta nas microrregiões que apresentam essas características.

A maior associação destas variáveis com o componente principal 1 corrobora o que foi encontra na matriz de correlação (Figura 9), em que a variável proporção de negros está altamente correlacionada com a renda, taxa de analfabetismo e taxa de pobreza.

TABELA 5 – Pesos fatoriais rotacionados

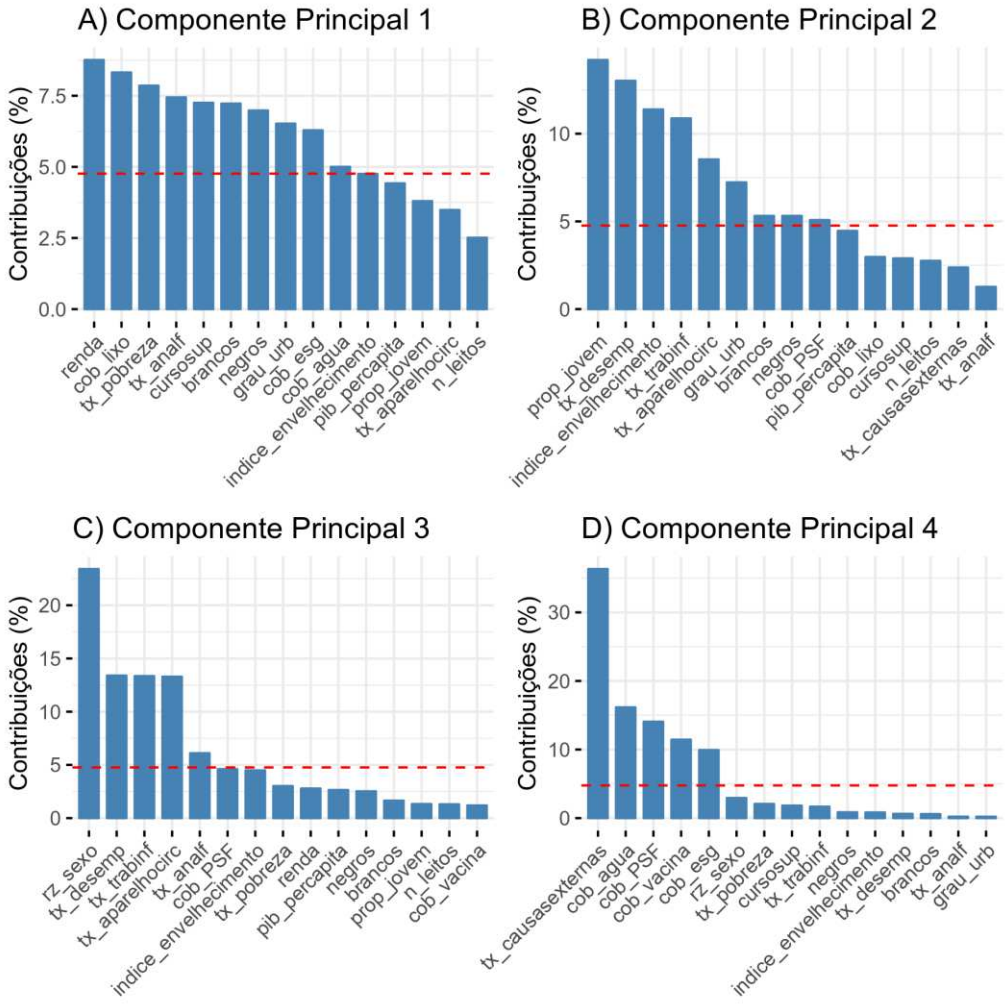
	Componentes			
Variáveis	**1**	**2**	**3**	**4**
cob_PSF	0,15	0,23	-0,21	-0,37
cob_agua	-0,22	-0,01	-0,11	-0,40
cob_lixo	-0,29	-0,17	-0,08	-0,03
cob_esg	-0,25	-0,05	-0,07	-0,31
cob_vacina	0,14	0,04	0,11	-0,34
n_leitos	-0,16	0,17	-0,11	0,01
pib_percapita	-0,21	-0,21	0,16	-0,01
tx_analf	0,27	0,11	-0,25	0,05
renda	-0,30	-0,10	0,17	0,04
tx_desemp	0,13	-0,36	-0,37	0,08
cursosup	-0,27	-0,17	0,08	0,14
tx_trabinf	0,11	0,33	0,37	-0,13
rz_sexo	0,14	0,03	0,48	-0,17
negros	0,26	-0,23	-0,16	-0,09
brancos	-0,27	0,23	0,13	0,08
grau_urb	-0,26	-0,27	-0,10	0,04
prop_jovem	0,19	-0,38	0,11	-0,02
tx_pobreza	0,28	0,10	-0,17	0,14
indice_envelhecimento	-0,22	0,34	-0,21	0,09
tx_causasexternas	-0,04	-0,15	-0,10	-0,60
tx_aparelhocirc	-0,19	0,29	-0,36	-0,02

Fonte: IBGE (2010a); Brasil (2010).

A linha pontilhada vermelha apresentada no Gráfico 8 indica a contribuição média esperada (1/número de variáveis). As variáveis que estão acima da linha vermelha são consideradas importantes para o componente principal como descrito acima.

A variável de maior contribuição comparando todos os componentes principais foi a taxa de mortalidade por causas externas (Componente Principal 4) com um *score* de 60%. Esta variável em conjunto com as outras de seu componente demonstra determinados fatores ambientais e de saúde que contribuem para a mortalidade adulta. Alguns estudos indicaram que condições do domicílio, como saneamento (cobertura de lixo, água e esgoto) se associam a maiores taxas de causas externas, ou seja, os locais que apresentaram maiores taxas de mortalidade por causas externas eram os que apresentaram os piores indicadores de qualidade de vida e tal mortalidade estava concentrada no grupo etário jovem adulto de 20 a 39 anos (PERES; CARDIA; SANTOS, 2006; ARAÚJO et al.; 2009; PEREIRA, 2014).

GRÁFICO 8 – Contribuição das variáveis para cada componente principal

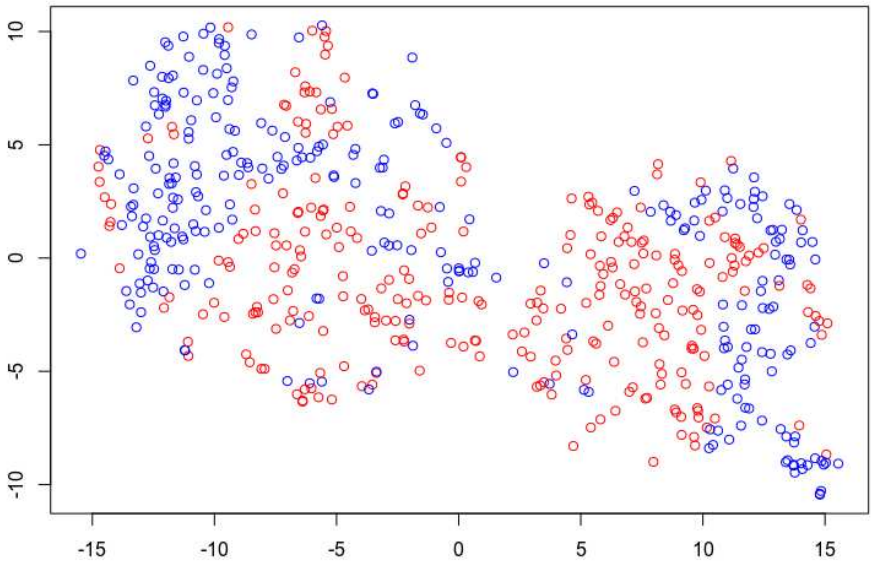


Fonte: IBGE (2010a); Brasil (2010).

Aplicou-se também o t-SNE (*t-distributed stochastic neighbor embedding*) e *K-Means*, a fim de exemplificar as duas classes criadas para a probabilidade de morte adulta que foram usadas nos modelos propostos.

O Gráfico 9 abaixo imprime a Dimensão 1 x Dimensão 2 geradas pelo modelo t-SNE, que tem como ideia principal converter um conjunto de dados de alta dimensão (por exemplo, várias variáveis independentes) em uma matriz de semelhanças entre pares. Cada ponto do gráfico representa uma observação, ou seja, uma microrregião. O t-SNE é capaz de capturar a estrutura dos dados de alta dimensão transformando em uma matriz de duas dimensões.

GRÁFICO 9 – Dados transformados em duas dimensões usando t-SNE

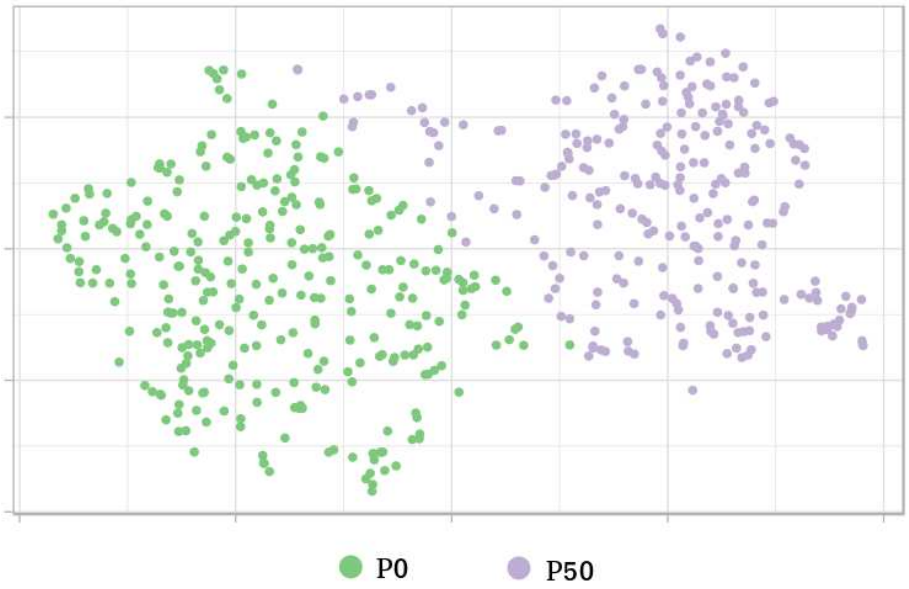


Fonte: IBGE (2010a); Brasil (2010).

Percebe-se que ao se ilustrar o conjunto de dados em duas dimensões que há dois agrupamentos um na parte esquerda e outro na parte direita. O *t-SNE* coloca casos semelhantes juntos, lidando bem com não linearidades de dados.

Para que fosse possível observar de forma mais clara esses agrupamentos aplicou-se o algoritmo *K-Means* nos resultados da matriz gerada pelo *t-SNE*, ou seja, o algoritmo gera as coordenas X-Y para cada observação e essas coordenadas foram usadas como entrada do algoritmo *K-Means* gerando dois grupos distintos de pontos como pode ser observado no Gráfico 10.

GRÁFICO 10 – Divisão dos dados em dois grupos, K= 2 usando K-Means



Fonte: IBGE (2010a); Brasil (2010).

A visualização dos dados por meio da Análise de Componentes Principais e do conjunto t-SNE com K-Means tornou possível o entendimento de forma global do banco construído e partir disso criou-se duas classes para a variável de interesse, a probabilidade de morte adulta, ou seja, a variável foi transformada, criando-se duas classes, P0 (agrupando valores abaixo de 0,1406 de ${}_{45}q_{15}$) e P50 (agrupando valores acima de 0,1406). Essa abordagem foi escolhida, pois, um padrão para a variável de interesse favorece e auxilia a análise em relação aos quatro algoritmos aplicados e neste sentido, tornam a interpretação mais tangível e acessível (DOUGHERTY; KOHAVI; SAHAMI, 1995; LUSTGARTEN et al., 2008).

3.2 Resultados e discussão

3.2.1 Análise da performance dos modelos

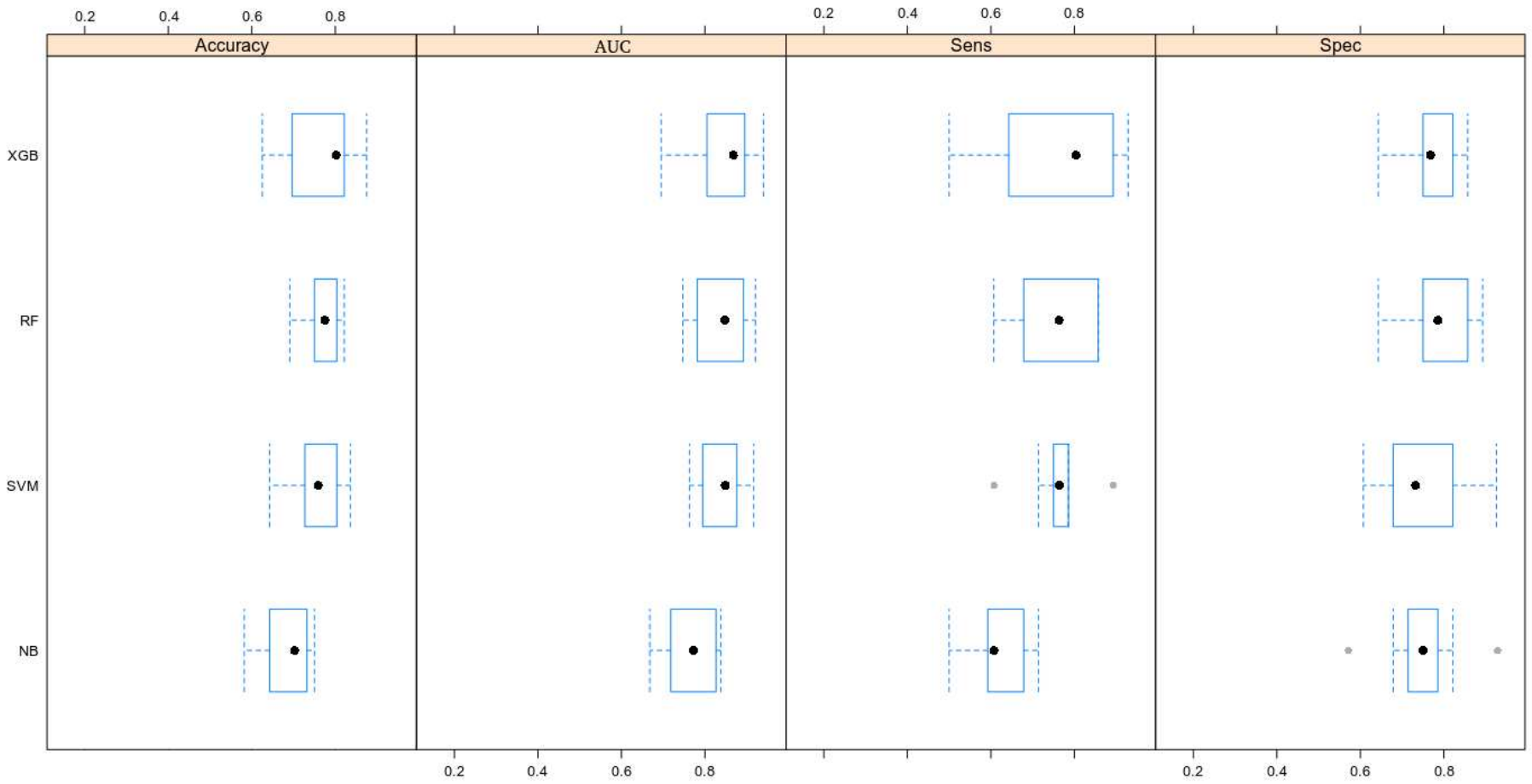
O desempenho dos modelos foi medido por meio do agregado das 10 reamostragens usando o método de validação cruzada dividindo os dados em 10 subconjuntos. A Figura 10 mostra as métricas de cada amostragem, ou seja, observa-se a variação das medidas de Acurácia, ROC (*Receiver Operating Characteristic*), Sensibilidade e Especificidade por subconjunto entre os quatro modelos: *Extreme Gradient Boosting* (XGB), *Random Forest* (RF), *Support Vector Machine* (SVM) e *Naive Bayes* (NB). As menores variações em relação à medida de Acurácia, ROC e Sensibilidade foram apresentadas pelo

modelo SVM, seguido do modelo RF. Os dois possuem variação similar em relação à Acurácia e ROC.

Em relação ao modelo XGB, a variação em relação à medida de Sensibilidade é um pouco maior em comparação com os outros modelos, a medida de Sensibilidade mede a proporção de verdadeiros positivos e, assim, avalia a capacidade de cada modelo em classificar uma microrregião como tendo probabilidade de morte adulta alta dado que realmente tenha, ou seja, houve maior variação ao se rodar os subconjuntos de reamostragem em relação ao grau de erro de classificação e a extensão do erro que conduz a predições erradas.

A medida AUC (Figura 10), representa a área abaixo da curva ROC (gráfico com sensibilidade em um eixo e especificidade no outro) e indica a escolha do ponto de corte da melhor combinação da medida de sensibilidade e da medida de especificidade, visto que classificar uma microrregião como uma determinada classe dado que ele é de outra classe (falso positivo) e classificar não sendo de uma determinada classe, quando ele é dessa classe (falso negativo) acarreta prejuízos na interpretação dos modelos.

FIGURA 10 – Métricas de desempenho por reamostragem



Fonte: IBGE (2010a); Brasil (2010).
Nota: Accuracy: Acurácia; Sens: Sensibilidade; Spec: Especificidade; XGB: Extreme Gradient Boosting; RF: Random Forest; SVM: Support Vector Machine; NB: Naive Bayes.

As médias das medidas dos quatro modelos são apresentadas na Tabela 6, ou seja, a média das reamostragens é fixada em um valor único para cada modelo. A medida de Acurácia foi similar entre os modelos de árvore de decisão (XGB e RF) e *Support Vector Machine* e mais baixa no modelo gerado pelo algoritmo *Naive Bayes*. Em relação à Sensibilidade, o modelo que se sobressai, novamente foi o SVM, seguido do XGB e do RF, sendo o pior desempenho apresentado, também pelo modelo NB, ou seja, calcula qual a proporção de observações que realmente eram da Classe 1 que foram diagnosticados pelo algoritmo como sendo da Classe 1, ou seja, é a proporção de verdadeiros positivos.

Nos modelos apresentados, portanto, o melhor desempenho geral foi apresentado pelo SVM seguido por RF e XGB, e o pior desempenho foi do modelo NB. Para as análises posteriores que serão feitas, a Acurácia é a medida mais importante seguido da medida de Precisão, e neste sentido demonstra o número de vezes que uma classe foi predita de forma correta em relação a todas as classes discriminadas. Esta última é uma métrica importante, pois os resultados devem conter menos falsos positivos com o objetivo de fornecer uma predição e interpretação mais eficazes.

A Acurácia, por sua vez, é uma medida que indica quão bem um modelo discrimina um dado nunca visto, ou seja, ela consegue mostrar qual foi o desempenho do modelo ao se classificar uma determinada observação como sendo de probabilidade de morte adulta baixa ou alta. Em resumo, o melhor desempenho geral foi do SVM seguido pelo RF.

TABELA 6 – Comparação do desempenho dos modelos

	XGB	SVM	RF	NB
Acurácia	75,82%	76,11%	75,81%	67,74%
Kappa	51,63%	52,21%	51,61%	35,48%
Sensibilidade	75,97%	76,82%	73,60%	61,29%
Especificidade	75,67%	75,39%	78,02%	74,19%
Precisão	75,74%	75,74%	77,00%	70,37%
F1	75,85%	76,28%	75,26%	65,52%
Taxa de detecção	37,98%	38,41%	36,80%	30,65%

Fonte: IBGE (2010a); Brasil (2010).

3.2.2 Análises das variáveis de maior importância para cada modelo

A medida de importância foi escalada para terem valores de 0 a 100 com o objetivo de comparar os modelos. A função do pacote utilizada que calcula a importância das variáveis mostra o efeito geral das variáveis preditoras em cada um dos modelos. A medida usada por cada um dos modelos é diferente, por isso a escala de 0 a 100. No caso dos modelos

de árvore (RF e XGB), o cálculo de importância é feito usando a acurácia de cada árvore e refeita permutando cada variável preditora. Nos modelos de SVM e NB foi analisada a curva ROC para cada variável preditora e assim a área abaixo da curva ROC (AUC) é então usada como medida da importância da variável. Os valores mais altos estão ligados às variáveis com poder preditivo elevado e as variáveis com menores valores possuem menor poder preditivo. Sendo que a variável com valor zero de importância foi excluída da visualização do gráfico.

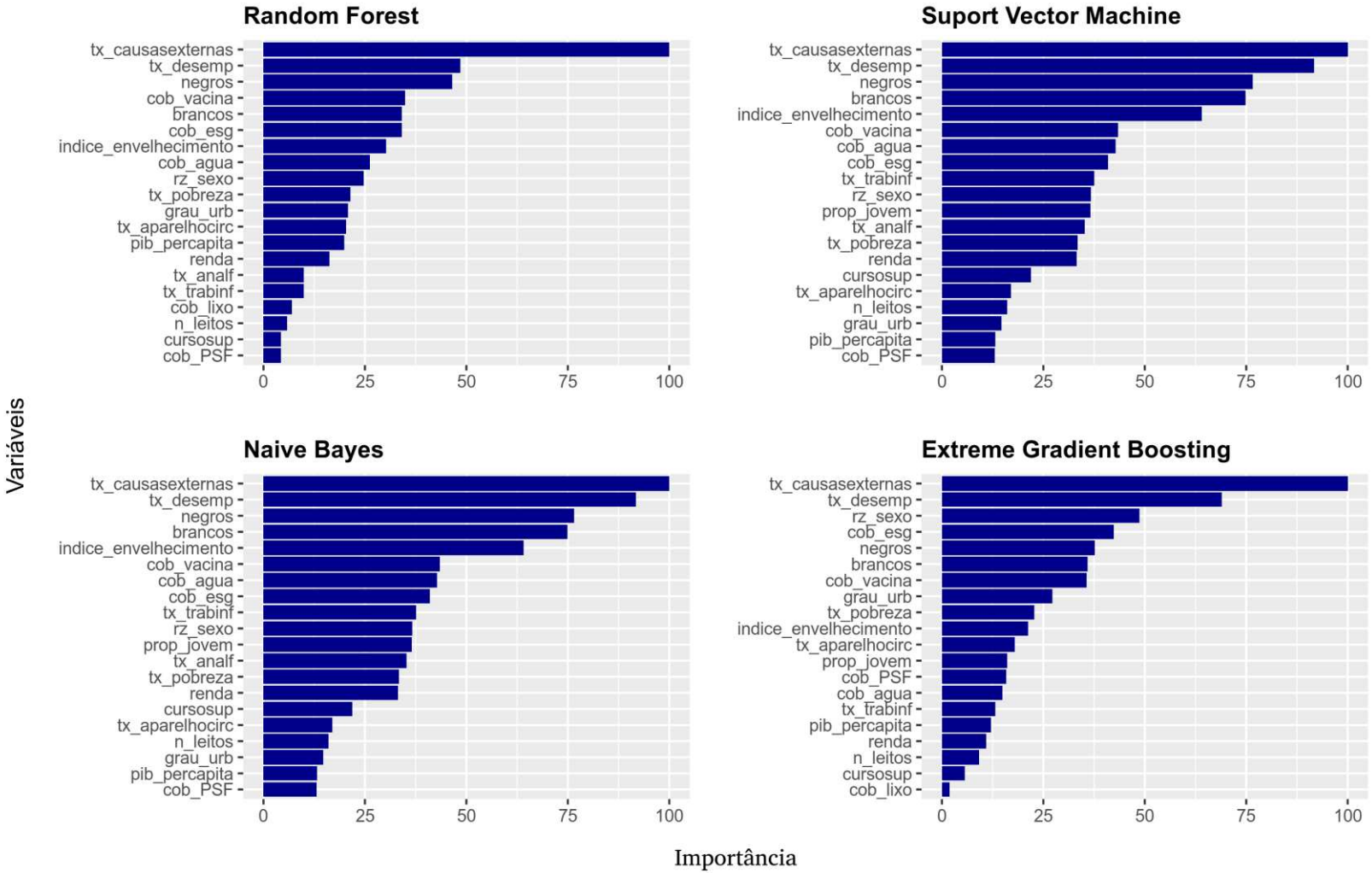
O Gráfico 11 ilustra a importância de cada determinante gerada pelos quatro modelos. Isto significa dizer que cada modelo, de forma geral, exemplifica o poder preditivo de cada uma das variáveis em relação à classificação de P50 (alta probabilidade de morte adulta).

A variável taxa de mortalidade por causas externas foi apontada como a variável mais importante nos quatro modelos seguido da variável taxa de desemprego como a segunda mais importante (Gráfico 11). Em terceiro lugar se encontra a variável proporção de negros na população para os modelos RF, SVM e NB, exceto o modelo XGB que classificou como terceira variável com maior preditivo a variável razão de sexo.

As variáveis com menor poder preditivo, ou seja, poder preditivo limitado foi, em alguma medida a variável proporção de indivíduos com curso superior, número de leitos por mil habitantes, cobertura do Programa Saúde da Família e Produto Interno Bruto per capita. O Grau de urbanização de uma microrregião foi uma característica que apresentou poder preditivo maior nos modelos de árvore de decisão RF e XGB, e não se apresentou como uma característica importante nos modelos NB e SVM. Diferenciais espaciais, em um estudo entre municípios de São Paulo revelaram que a mortalidade de 15 a 44 anos, nas áreas de moradia em piores condições sociais e ambientais foi o dobro comparado com as áreas de melhor condição (STEPHENS et al., 1994).

A variável proporção de jovens apresentou poder preditivo zero no algoritmo RF, ou seja, ele não teve nenhuma capacidade preditiva para este modelo. Para os modelos NB e SVM, a variável excluída foi à cobertura de lixo e para o modelo XGB, a variável com nenhum poder foi a taxa de analfabetismo.

GRÁFICO 11 – Comparação de importância das variáveis entre os modelos



Fonte: IBGE (2010a); Brasil (2010).

Observa-se que o fato das variáveis taxa de mortalidade por causas externas, proporção de indivíduos de cor preta e proporção de indivíduos de cor branca estar entre as principais variáveis de poder preditivo de mortalidade adulta podem corroborar o indicado por Araújo et al. (2009) que em seu estudo conclui que a cor da pele possui impacto nos anos potenciais de vida perdido por causas externas, ou seja, o número de anos de vida perdidos devido as mortes por causas externas foi 8,8 vezes maior para a população de cor preta e parda no Brasil.

O estudo de Campos et al (2015), investigou o perfil da carga das causas externas por faixa etária, sexo e regiões do Brasil e encontrou que em 2008 as causas externas representaram 10% da carga total de doença. Os homicídios e violências foram os principais responsáveis pelos anos de vida perdidos, com 38,8%, sendo que nas regiões Norte e Nordeste foram de 43,9% e 46,6%, respectivamente, seguido pelos acidentes de trânsito com 36% dos anos de vida perdidos.

Entre o período de 2000 a 2010 aumentou as taxas de homicídios nas Regiões Norte e Nordeste, sendo que no Sudeste, no mesmo período, houve declínio em relação aos homicídios (CAMPOS et al., 2015). O aumento nestas regiões pode estar associado ao processo de urbanização acelerada levando a novos pólos econômicos sem nenhuma estruturação de políticas de segurança pública e infraestrutura (WAISELFISZ, 2014; IPEA; FBSP, 2018). Outro motivo pode estar relacionado a estrutura etária, ou seja, um estudo realizado no Estado de São Paulo apontou que quase metade das reduções nos índices de homicídios ocorridos nos últimos anos poderia ser explicado pela aceleração do envelhecimento populacional, com declínio da população jovem, grupo mais vulnerável a sofrer os impactos da violência (MELLO; SCHNEIDER, 2007; BRASIL, 2015b). Além disso, o estudo aqui analisado coloca o índice de envelhecimento como variável de importância intermediária podendo corroborar o afirmado por Mello e Schneider (2007).

Ainda segundo Campos et al. (2015), os acidentes de trânsito foram a segunda causa com maior parcela de anos de vida perdidos por incapacidade em relação ao total das causas externas (29,1%). Além disso, no que dizem respeito aos grupos etários, os autores encontraram que o grupo de maior representação para os anos de vida perdido é o de 15 a 29 anos seguido pelo grupo de 30 a 44 anos. Soma-se a isto, a maior participação dos homens em todas as faixas etárias.

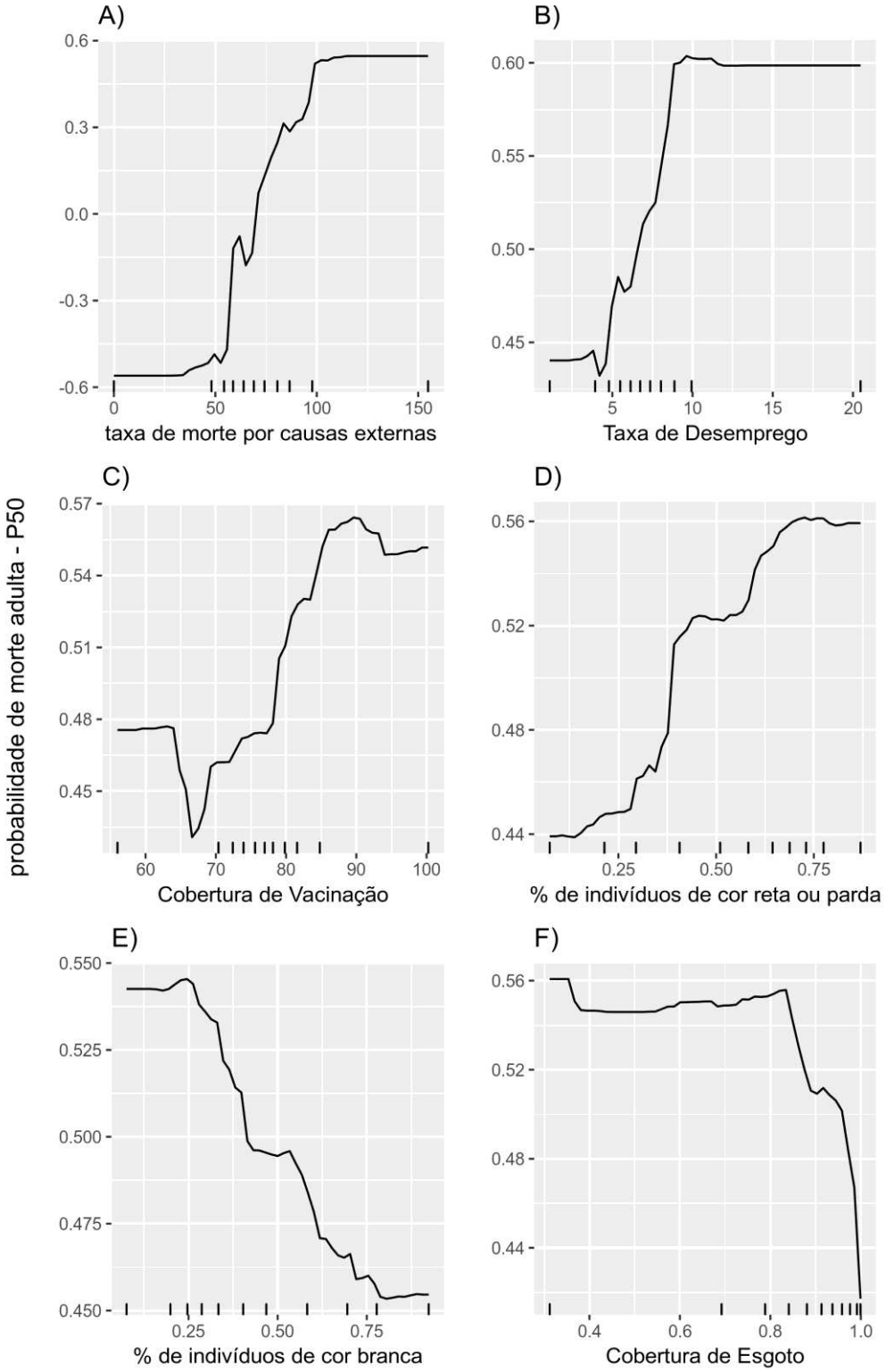
Além disso, apresenta-se a influência das variáveis consideradas mais importantes pelos dois modelos escolhidos RF e XGB apresentadas individualmente por meio de gráficos de dependência parcial para uma análise focada em cada uma dessas variáveis (Gráficos 12 e

13). Esse tipo de gráfico tem como principal objetivo mostrar o efeito marginal de uma característica no resultado previsto de um modelo previamente ajustado (MOLNAR, 2018). No caso do modelo de classificação proposto, em que são geradas probabilidades agrupadas em duas classes, P0 e P50, a função de dependência parcial exibe a probabilidade de uma determinada classe, dados valores diferentes para cada uma das variáveis selecionadas. Este tipo de gráfico leva em consideração todas as observações e faz uma suposição sobre o relacionamento entre uma variável e o resultado previsto de forma global (MOLNAR, 2018).

A escolha final dos melhores modelos depende do contexto de uso, em que um método mais simples com maior facilidade de interpretação pode superar os pequenos ganhos em desempenho e neste sentido, para uma análise individual sobre os principais determinantes de mortalidade adulta se escolheu os modelos de árvore, ou seja, RF e XGB. Tanto *Random Forest* quanto *Extreme Gradient Boosted Tree* possuem a vantagem de levarem em consideração a interação entre variáveis (NGUYEN, 2016).

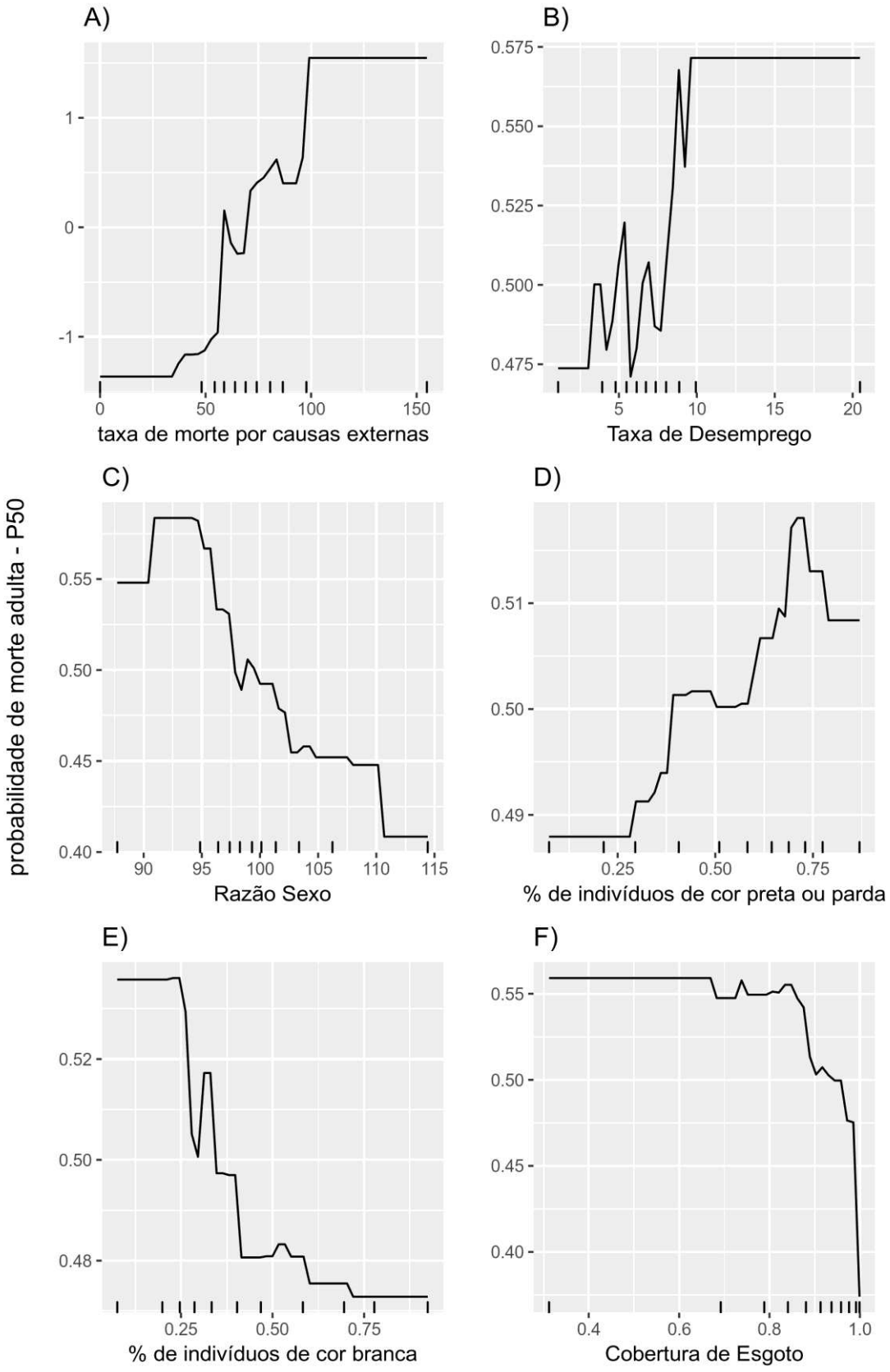
Tanto a proporção de indivíduos pretos e pardos quanto a proporção de indivíduos de cor branca estão entre as variáveis com maior poder preditivo em todos os modelos. Apesar da importância alta para as duas variáveis, os gráficos 12-D e 13-D comprovam que quanto maior a proporção de indivíduos de cor preta e parda na microrregião maior a probabilidade dos dois modelos classificarem esta microrregião como tendo uma probabilidade de morte adulta alta. De forma inversa se comportou a variável proporção de brancos, ou seja, quanto maior a proporção de brancos em uma microrregião, menor a probabilidade dos modelos em classificarem esta microrregião como tendo uma probabilidade de morte adulta alta. Embora raça e desigualdades socioeconômicas estejam claramente correlacionadas, grande parte da pesquisa sobre mortalidade os vê como características distintas, porque os vários indicadores socioeconômicos diferem dentro e entre os grupos raciais (MASTERS et al., 2014; WILLIAMS; PRIEST; ANDERSON, 2016; CUNNINGHAM et al., 2017).

GRÁFICO 12 – Dependência parcial das seis mais importantes variáveis – Modelo RF



Fonte: IBGE (2010a); Brasil (2010).

GRÁFICO 13 – Dependência parcial das seis mais importantes variáveis – Modelo XGB



Fonte: IBGE (2010a); Brasil (2010).

O estudo de Bosworth (2018) sobre desigualdades socioeconômicas nos Estados Unidos aponta que diferenciais raciais são fatores importantes nas disparidades nas taxas de mortalidade na maioria das idades, mas a interação do papel da raça e condições socioeconômicas na contabilização das diferenças é controversa. Em relação a estudos sobre desigualdade em saúde, Williams (2016) acrescenta que, nos Estados Unidos, há grandes diferenças raciais quanto se trata da qualidade do tratamento médico mesmo depois de ajustar por fatores de acesso, severidade das doenças e condições socioeconômicas.

Estudos tanto dos Estados Unidos quanto do Brasil mostraram que áreas onde a população é constituída por uma grande proporção de negros possuem indicadores socioeconômicos e de infraestrutura desfavoráveis. Esta segregação contribui para grandes diferenças em relação a níveis educacionais, emprego, renda, saúde delimitando as possíveis oportunidades de ascensão social e aumentando o número de homicídios e violência nesse grupo populacional (BRASIL, 2005; IPEA, 2011; NURU-JETER; LAVEIST, 2011; FIORIO et al., 2011; PETRUCELLI; SABOIA, 2013; OLIVEIRA THOMAZ; SILVA, 2014; CUNNINGHAM, et al., 2017). Dessa maneira, a predição dos algoritmos mostra que, assim como Lovell (1999) a raça tende a predizer riscos aumentados em saúde e em mortalidade independente da condição econômica.

Fiorio et al. (2011) afirmam que “negros e brancos ocupam lugares desiguais na sociedade e trazem consigo experiências também desiguais no nascer, viver, adoecer e morrer” (FIORIO et al., 2011, p. 529). O estudo do Ministério da Saúde (BRASIL, 2015b) aponta as principais desigualdades no perfil da mortalidade por raça/cor e mostra que entre a população negra, as maiores taxas de mortalidade foram devido a doenças cardiovasculares e diabetes mellitus. Por sua vez, entre a população branca, foi o infarto agudo do miocárdio e pneumonia. Além disso, os homicídios aparecem como uma das principais causas de morte entre a população preta e parda.

As mortes por causas externas ocorrem principalmente na população adolescente e vida adulta, ou seja, representa parte do grupo etário (15 a 29 anos) aqui em estudo. De acordo com os Gráficos 12-A e 13-A, quanto maior a taxa de mortalidade por causas externas maiores a probabilidade do modelo RF e XGB classificar uma determinada microrregião como sendo da Classe P50, ou seja, ter probabilidade de morte adulta alta.

A variável desemprego sendo a segunda variável de maior poder preditivo valida de forma geral uma extensa literatura que documentou e descreveu a forte associação entre desvantagem no mercado de trabalho, saúde e mortalidade (IVERSEN et al., 1987; MOSER et al., 1987; ROELFS et al., 2011, QUEIROZ et al., 2017). Altas taxas de desemprego estão

relacionadas à baixo crescimento e desenvolvimento econômico. Conforme Gerdtham e Johannesson (2003), o desemprego tem sido associado ao aumento da mortalidade possivelmente por meio de mediadores como mal estado psicológico (por exemplo, depressão) e comportamentos insalubres (aumento do abuso de substâncias como álcool e drogas). Por outro lado, o emprego tem sido associado à personalidade positiva e características como melhor auto-estima e independência.

Corrobora este fato que quanto maior a taxa de desemprego maior a probabilidade do modelo RF e XGB classificar uma determinada microrregião como sendo da classe P50 (Gráficos 12-B e Gráfico 13-B). Martikainen (1999) observa que o desemprego prolongado possui efeito negativo na saúde e aumenta o risco de morte prematura.

O estudo de Queiroz et al. (2017) encontrou que altas níveis de desemprego estavam relacionados a maior probabilidade de morte adulta no Brasil. Montgomery et al. (2013), por sua vez, mostraram por meio de um estudo longitudinal na Suécia que entre os homens de 34 a 38 anos e 45 a 49 anos, o risco de mortalidade relacionado ao desemprego é maior entre o segundo grupo etário mais velho e entre os mais qualificados. Com o aumento da idade, a experiência do desemprego pode ser inesperadamente prejudicial. A falta de adaptação as adversidades sociais, como por exemplo, perder o emprego, podem levar a aumento do risco de Acidente Vascular Cerebral. Isso destaca a maior vulnerabilidade dos trabalhadores mais velhos em tempos de recessão e maiores taxas de desemprego.

No outro extremo, o desemprego entre a população jovem e em transição para a vida adulta, de 18 a 24 anos se relaciona a maiores riscos de mortalidade geral, por homicídios e por todas as causas de morte depois de controlar pelo sexo, raça e educação (DAVILA et al., 2010). A literatura sobre taxas de lesões entre trabalhadores jovens e tendências de mortalidade nos EUA entre jovens de 15 a 24 anos, indica que adultos jovens do sexo masculino tinham uma maior probabilidade de morrer, o que pode ser devido a seus comportamentos mais agressivos e arriscados. Além disso, educação era protetora do risco de mortalidade em todas as causas de morte para adultos jovens, possivelmente relacionadas a ocupações menos perigosas disponíveis para jovens qualificados (SINGH; YU, 1996).

Clemens; Popham e Boyle (2015) observa que o desemprego é mais provável de ocorrer entre indivíduos de pior nível socioeconômico e pode ser que os efeitos prejudiciais à saúde associados à pobreza e desvantagem antes do desemprego possam ser responsáveis pelo aumento do risco de mortalidade e não pelos efeitos causados pelo próprio desemprego. Após a comparação com base na saúde, os resultados de Clemens; Popham e Boyle (2015) mostraram um aumento de 85% e 50% no risco de mortalidade para homens e mulheres,

respectivamente, que foram registrados como desempregados dez anos em relação aos que permaneceram no emprego.

O emprego que não é recuperado em um curto período de tempo resulta na perda de renda e pode rapidamente impor dificuldades econômicas subsequentes. O *stress* agudo e a privação material tendem a surgir entre muitos trabalhadores deslocados após a perda inicial de emprego à medida que aumentam as incertezas sobre as perspectivas de vida imediatas e futuras tendo consequências entre os trabalhadores vulneráveis, devido a circunstâncias sociais e de saúde (GARCY; VAGERO, 2012).

Outra variável de poder preditivo importante foi a cobertura de esgoto e ela reflete melhores condições de infraestrutura da região, além de ser uma *proxy* do desenvolvimento socioeconômico assim como a variável cobertura de água encanada (variável de poder preditivo intermediário nos modelos). Desta forma, a importância dessa variável comprova a influência para a determinação da mortalidade adulta. Santos e Noronha (2001), ao estudar os diferenciais socioeconômicos e os padrões socioespaciais de mortalidade no Rio Janeiro mostraram que os lugares com melhor infraestrutura, com saneamento e domicílios com esgoto ligado à rede geral apresentaram um perfil de mortalidade mais baixo que as localidades com menor infraestrutura. O Gráfico 12-F evidenciou que maiores coberturas de esgoto na microrregião significam para o modelo RF menor probabilidade de a região apresentar probabilidade de morte adulta alta (Classe P50). A relação do lugar de onde se vive e a condições de saúde da população está diretamente ligado a privação social e material da população menos favorecida levando a piores condições de saúde e a uma mortalidade mais alta. Inclui-se nesse processo de privação, o pior acesso a serviços de saúde.

Esperava-se que a razão de sexo apresentasse um comportamento diferente do ocorrido (Gráfico 13-C), pois a mortalidade é mais alta entre os homens deste grupo etário do que entre as mulheres. A microrregião com Razão de Sexo acima de 100 (predominância de homens na microrregião) apresentou a probabilidade de ser classificada como tendo baixa mortalidade adulta pelos modelos. Este resultado do gráfico de dependência parcial pode ser devido à escassez de observações acima de 105 homens para cada 100 mulheres.

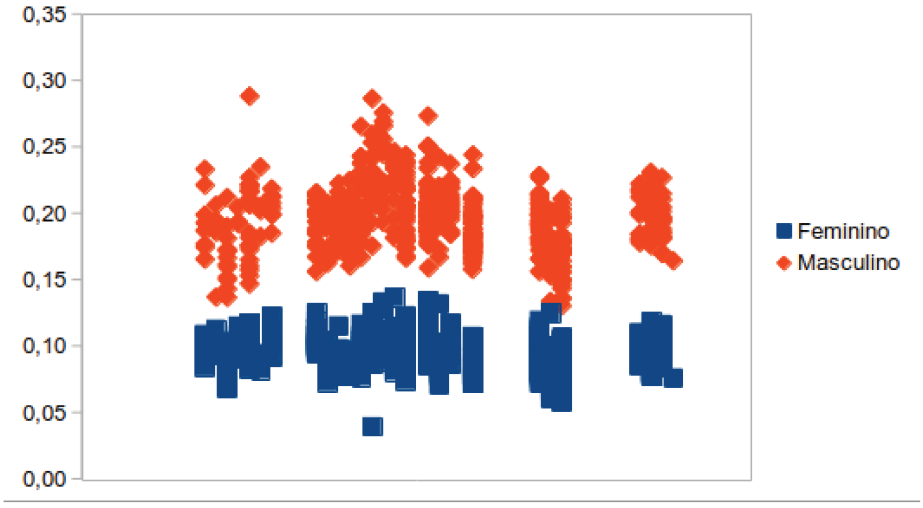
A probabilidade de morte adulta é mais alta para homens do que para mulheres como indica o Gráfico 14. A Razão de Sexo de uma determinada região pode ser influenciada pela migração e diferenciais de mortalidade por sexo. Razão de Sexo maior que 100, ou seja, uma população que possui um número maior de homens tem sido associada a taxas de criminalidade mais altas (HUDSON; BOER, 2002). A relação da razão de sexo pode estar associada ao fato de que maiores razões de sexo, ou seja, maior número de homens em

relação a um maior número de mulheres leva a probabilidade de morte adulta mais alta. O modelo XGB classificou esta variável como apresentando um poder preditivo alto.

A inserção de um indivíduo na sociedade está altamente ligada ao sexo, ou seja, a mortalidade, por exemplo, afeta de forma diferente homens e mulheres. A vivência do homem e da mulher é apreendida no contexto social de forma diferenciada, a partir de valores e ideais. Diferenciais de mortalidade por sexo são observados em todo as faixas de idade ao longo do ciclo de vida dos indivíduos (PEREIRA, 2014). O estudo de Wingard (1984) definiu dois motivos principais para as diferenças entre homens e mulheres nas condições de saúde, um deles está associado ao papel social no casamento, paternidade, emprego, entre outros e o outro está associado ao comportamento em relação ao consumo de álcool, drogas, conduta no trânsito e nos cuidados com a saúde. Estes motivos combinados levariam a aspectos diferenciados no estilo de vida e aos diferencias por sexo.

No grupo etário em estudo, a maior mortalidade masculina pode estar ligada aos fatores levantados acima se manifestando nas diferenças em valores, estilo de vida e ideais de masculinidade que levam a comportamentos de risco e se traduzem em maiores taxas de mortalidade adulta entre os homens em comparação às mulheres.

GRÁFICO 14 - Probabilidade de morte adulta por sexo – Microrregiões, Brasil, 2010



Fonte: IBGE (2010a); Brasil (2010).

A cobertura de vacinação de uma microrregião possui alto poder preditivo em todos os modelos, tendo o maior poder preditivo no modelo *Random Forest.* Entende-se a cobertura de vacinação como sendo uma intervenção de saúde pública e, logo, conclui-se que melhores coberturas de vacinas produzem melhores indicadores de saúde e menor mortalidade. Em geral, segmentos populacionais com piores indicadores socioeconômicos estão associados

a menores valores de cobertura de vacinação. Essa relação pode ser compreendida como um aspecto da condição de vida que dificulta o acesso a vacinação, portanto, piores condições sociais e econômicas podem levar a uma menor oferta de serviços e uma maior dificuldade de acessar determinada intervenção de saúde como campanhas de vacinação (TRAVASSOS; MARTINS, 2004; MORAES; RIBEIRO, 2008).

Ao se olhar está variável nos gráficos de dependência parcial (Gráficos 12 e 13), observa-se que está ela se comportou de forma não esperada, em que, regiões com valores maiores de cobertura de vacinação foram classificadas com maior probabilidade de pertencer à classe P50, isto é, apresentar mortalidade adulta mais alta. Esperava-se que maiores coberturas de vacinação mostrassem maior acesso aos serviços de saúde e, portanto, mortalidade adulta mais baixa. A escassez de pontos de dados (traços acima do eixo x) após os 80% de cobertura indica que o modelo não tinha muitos pontos de dados para aprender acima deste valor, o que implica que talvez não possamos confiar nas previsões do modelo de aprendizado de máquina para coberturas de vacinação maior de 80%.

No contexto do grupo etário adulto, o levantamento do Ministério da Saúde feito por meio do Programa Nacional de Imunização mostrou que a cobertura de vacinação na idade de 20 a 59 anos com as vacinas recomendada para adultos (Hepatite B, Febre Amarela, Tríplice Viral e a Dupla adulta) está muito abaixo do ideal (BRASIL, 2015a). O foco das principais campanhas de vacinação são as crianças, grávidas e idosos e, portanto, não há uma medida da cobertura para a população adulta. Logo, apesar do estudo em análise usar a cobertura de imunização das crianças, o método utilizado e os diferentes modelos foram capazes de achar uma relação entre cobertura de vacinação na microrregião e mortalidade adulta, podendo estar relacionado a insuficiência de vacinação no grupo adolescente e adulto mostrado pelo estudo do Brasil (2015a).

Um dos motivos para que as variáveis razão de sexo e cobertura de vacinação apresentassem comportamentos diferentes do esperado ao se analisar os gráficos de dependência parcial podem ser também devido a suposição de independência (MOLNAR, 2018). Isto é, para a dependência parcial da razão de sexo, por exemplo, assume-se que ela não está correlacionada com outras características, o que pode ser uma suposição errada e mostra que quando as variáveis são correlacionadas, coloca-se um peso em características que possuem um conjunto baixo de probabilidade real levando-se a fazer suposições irrealistas ou não esperadas. Nessa perspectiva, as variáveis razões de sexo e cobertura de vacinação possui alto poder preditivo como colocados pelos modelos, porém não devem ser avaliados de forma

independente sem levar em consideração outras características correlacionadas com essas variáveis.

As análises mostraram que todos os modelos podem discriminar de forma significativa as probabilidades de mortalidade adulta mais baixa (Classe P0) e probabilidades de mortalidade adulta mais alta (Classe P50) usando um conjunto de características socioeconômicas, demográficas e de saúde por microrregião, comprovando a importância destas características regionalmente como indicadores e determinantes da mortalidade no grupo de 15 a 59 anos.

A abordagem utilizada nesse estudo usando fatores contextuais permitiu apreender certas relações entre estrutura social e saúde que no nível individual pode não ser possível. Nessa perspectiva, buscou-se tentar entender por que em algumas microrregiões há uma mortalidade adulta mais alta que em outras e por que sua distribuição é desigual. Os recursos do local de domicílio podem determinar a qualidade de vida do indivíduo, e assim seu contexto social teria grande influência na determinação das condições de saúde.

Soma-se a isto, a importância das variáveis proporções de pretos e pardos e a proporção de brancos nas microrregiões na predição de mortalidade adulta refletindo a evidência de que as desigualdades em saúde da população brasileira estão fortemente relacionadas com a construção histórico-social que levou a sociedade brasileira a uma divisão por subgrupos populacionais por meio da raça/cor e assim apresentam uma falta de equidade determinada por essa variável. Ou seja, a maior proporção de pretos e pardos em uma determinada microrregião determina probabilidade de mortalidade adulta mais alta. Porém, alguns questionamentos surgem que não podem ser respondidos pelo presente estudo: ser branco em uma microrregião com maior proporção de negros aumenta a probabilidade de mortalidade adulta para este indivíduo branco? E de forma contrária, será que ser um indivíduo negro em uma microrregião com maior proporção de brancos diminui a probabilidade de morte adulta deste indivíduo?

Silva e Silva (2003) afirmam que o processo de exclusão e segregação é extremo no Brasil e que mesmo em indivíduos que moram em favelas localizadas em áreas com maior proporção de população branca e de melhor condição socioeconômica não se beneficiam desta localização onde existe maior proporção de brancos. A relação entre segregação socioespacial em Salvador e mortes por causas externas, principalmente violência mostrada por Araújo et al. (2010) mostrou que a mortalidade por todas as causas externas e por homicídios ficaram concentradas nas regiões de Salvador onde reside grande parte da população negra da capital. Os achados revelam que microrregiões com maior proporção de

negros possuem maiores riscos de ocorrência de mortes adulta e pode-se inferir que estas causas estejam associadas as causas externas (LIMA et al., 2005; ARAÚJO et al., 2010; MENDES et al., 2015).

Em relação a variável taxa de mortalidade por causas externas, os modelos a classificaram como a primeira de maior poder preditivo no contexto de mortalidade adulta, ou seja, a taxa de mortalidade por causas externas é de maior preditivo na classificação de mortalidade alta ou baixa em uma microrregião. O homicídio é a principal causa de óbitos a faixa etária de 15 a 29 anos sendo 60% das mortes por causas externas, da faixa etária de 30 a 44 anos e de 45 a 59 anos começa a ter a maior proporção de mortalidade por causas externas devido a acidentes de trânsito e diminui às taxas de homicídio (FERREIRA; ARAÚJO, 2006).

A mortalidade adulta e as taxas de mortalidade por causas externas são altamente correlacionadas (ver Gráfico 5, Seção 3.1) porém, é necessário observar que não somente esta variável, mas a proporção de negros e brancos, grau de urbanização, razão de sexo e cobertura de esgoto possuem alto poder preditivo na classificação de probabilidade de mortalidade adulta mais altas nas microrregiões. Isto é, ter entre 15 a 59 anos em microrregiões com alta proporção de negros ou baixa proporção de brancos, baixa cobertura de esgoto nos domicílios e cobertura de vacinação leva a maior mortalidade nesta faixa etária. A desigualdade, desta forma, se expressa nas microrregiões com maior proporção de negros, piores condições de saneamento, de acesso à saúde preventiva (imunização) e com altas taxas de mortalidade por causas externas que acometem em grande proporção jovens e adultos, negros e homens.

Parece-me que todas as variáveis de maior poder preditivo de alguma forma nos levam a segregação social e racial, ou seja, a população negra no Brasil sofre de maior desvantagem social, sua concentração é nas regiões mais pobres do país, seu nível de desenvolvimento é inferior ao da população como um todo, o acesso a saneamento, educação e emprego é significantemente menor (BARATA, 2009). Porém, é de difícil mensuração e de alta complexidade as relações à raça/cor e saúde. Há um conjunto de complexo de determinações nem sempre passíveis de quantificação (BARATA, 2009).

Em relação às desigualdades por sexo quando se trata de olhar as mortalidades de adultos, a sobremortalidade masculina relacionada a exposição a fatores e situações de risco ao longo da vida como situações insalubres de emprego, ou comportamentos nocivos à saúde como maior consumo de álcool, cigarro e drogas entre os homens e exposição frequente a riscos como acidentes e violências.

Estudos usando o método de aprendizado de máquina para determinação de mortalidade e condições de atividade física em alguns países (NGUYEN, 2016; SAEZ;

BALDOMINOS; ISASI, 2016; SINGHA et al., 2016; SAKR et al., 2017; PAN et al.,2017) mostraram que como pode ser difícil saber a priori quais métodos de aprendizado de máquina terão melhor desempenho em uma aplicação específica como o caso deste estudo, investigar uma variedade de modelos com formulações diferentes pode ser uma boa estratégia para entender o problema de forma ampla. Dessa forma, por haver algumas diferenças em alguns dos principais preditores nos modelos, escolheu-se basear as principais conclusões em dois algoritmos, no lugar de basear em somente um. Usando apenas um único método para examinar os resultados poderia nos levar a conclusões sobre a importância de algumas variáveis que entram em conflito com os resultados de outros modelos similarmente apropriados.

CONSIDERAÇÕES FINAIS

O objetivo principal foi entender e investigar as relações entre os fatores socioeconômicos, estruturais, contextuais e de saúde com a probabilidade de morte adulta no nível agregado das microrregiões brasileiras em 2010 aplicando o método de aprendizado de máquina visando entender quais algoritmos possuem melhor performance e compreender essas relações. Timaes; Chackiel e Ruzicka (1996) observaram que as tendências e diferenciais geográficos na mortalidade adulta podem ser diferentes daqueles que influenciam a mortalidade infantil, para a mortalidade adulta, os fatores contextuais podem instigar uma mudança direta nos indicadores de risco próximos como condições macroeconômicas, fatores culturais, sistema de saúde e fatores ecológicos.

Os resultados apontaram que as variáveis mais importantes na determinação da probabilidade de morte adulta nas microrregiões brasileiras foram, principalmente, a taxa de mortalidade por causas externas, taxa de desemprego, proporção de negros, proporção de brancos, cobertura de vacinação e cobertura de esgoto. Usar os métodos propostos pelos algoritmos de aprendizado de máquina confirma que há um melhor entendimento da mortalidade adulta quando investigamos as variáveis de forma conjunta e mesmo com altas correlações entre elas. Ao analisar as variáveis com maior poder preditivo por meio da dependência parcial chegou-se à conclusão de que determinadas variáveis devem ser analisas em conjunto com as outras, mostrando a forte correlação delas e, assim, consegue-se alcançar de forma mais apropriada à realidade de como fatores contextuais expressam diferentes níveis de mortalidade adulta.

Entende-se pelas análises feitas que as desigualdades sociais nas taxas de mortalidade são uma expressão da situação socioeconômica no nível das microrregiões. Sendo assim, os resultados obtidos ajudam a melhorar o planejamento de políticas públicas destinado a promover maiores progressos nas condições de vida e de saúde e diminuir as taxas de mortalidade das populações mais vulneráveis no Brasil. Apropriando-se da definição de raça/cor como construção social e possível determinante importante de diferenciais em saúde, mesmo quando se leva em conta características socioeconômicas diretas como, por exemplo, renda, PIB per capita, taxa de pobreza e taxa de analfabetismo, mostra de alguma maneira as segregações raciais nas quais o Brasil está socialmente estruturado.

O Brasil, em 2018, se encontrou em 13ª lugar em relação a todos os países no que diz respeito às taxas de homicídio (IPEA; FBSP, 2018). Salienta-se a importância que as políticas públicas devem ter em diminuir a mortalidade por causas externas entre a população

economicamente ativa e consideradas mortalidades evitáveis, ou seja, mortes preveníveis por meio de melhoras nas políticas em saúde, os serviços e o acesso dos grupos populacionais mais vulneráveis apresentados.

A desigualdade no perfil da mortalidade no Brasil exige o aprofundamento das análises por diversas variáveis estratégicas – sexo, idade, raça/cor da pele, localização geográfica-além de novas abordagens das causas de morte. A análise das desigualdades por pequenas áreas é fundamental para orientar as prioridades de prevenção e tratamento de doenças que mais matam (BRASIL, 2015b). As políticas sociais destinadas a reduzir a lacuna social entre os grupos populacionais terão influências importantes sobre as condições de saúde da população.

O presente estudo é inovador e pioneiro no Brasil. Foi o primeiro estudo exploratório usando algoritmos de aprendizado de máquina como alternativa às modelagens tradicionais para entender as relações dos fatores contextuais, econômicos e de saúde em nível de pequenas áreas na mortalidade adulta no país. Para estudos futuros é interessante desagregar a probabilidade de morte adulta separadamente por sexo e por grupos etários menores do que o grande grupo de 15 a 59 anos. Pode ser que determinados fatores tenham influências diferentes entre os sexos e entre grupos etários de 15 a 29 anos, de 30 a 44 anos e de 44 a 59 anos separadamente e por níveis dos municípios brasileiros. Além disso, uma extensão metodológica deste trabalho poderia incluir técnicas de modelagem agrupada, combinando os resultados de vários modelos para criar um modelo que leva em consideração os vários elementos dos dados que cada algoritmo identificou como importantes.

Cabe observar que esse estudo é transversal, ou seja, se refere a um corte em um ponto no tempo, constituindo uma limitação do estudo, pois não permite estabelecer determinadas relações de causalidade e tirar algumas conclusões mais aprofundadas sobres essas relações.

Finalmente, os resultados encontrados podem contribuir efetivamente para a melhora das condições da saúde da população brasileira, uma vez que forneceu elementos importantes, auxiliando os planejadores de políticas públicas na tomada de decisões regionais e de acordo com a realidade local.

REFERÊNCIAS

AGOSTINHO, C. S. **Estudo sobre a mortalidade adulta, para Brasil entre 1980 e 2000 e Unidades da Federação em 2000:** uma aplicação dos métodos de distribuição de mortes. 2009. 256f. Tese (Doutorado) – Centro de Desenvolvimento e Planejamento Regional, Universidade Federal de Minas Gerais, Belo Horizonte, MG, 2009.

AGOSTINHO, C. S.; QUEIROZ, B. L. Estimativas da mortalidade adulta para o Brasil no período 1980/2000: uma abordagem metodológica comparativa. In: ENCONTRO NACIONAL DE ESTUDOS POPULACIONAIS, 16., 2008, Caxambu, MG. **Anais...** Belo Horizonte, MG: ABEP, 2008.

AKSAN, A.; CHAKRABORTY, S. Mortality versus morbidity in the demographic transition. **European Economic Review,** Amsterdam, v. 70, p. 470-492, 2014.

ALMEIDA, N. D. A saúde no Brasil, impasses e desafios enfrentados pelo Sistema Único de Saúde – SUS. **Revista Psicologia e Saúde,** Campo Grande, MS, v. 5, n. 1, p. 01-09, 2013.

ANDRADE, M. V. et al. Desigualdade socioeconômica no acesso aos serviços de saúde no Brasil: um estudo comparativo entre as regiões brasileiras em 1998 e 2008. **Economia Aplicada,** Ribeirão Preto, SP, v. 17, n. 4, p. 623-645, 2013.

ARAÚJO, E. M. et al. Spatial distribution of mortality by homicide and social inequalities according to race/skin color in an intra-urban Brazilian space. **Revista Brasileira de Epidemiologia,** São Paulo, SP, v. 13, n. 4, p. 549-560, 2010.

ARAÚJO, E. M. et al. Diferenciais de raça/cor da pele em anos potenciais de vida perdidos por causas externas. **Revista de Saúde Pública,** São Paulo, SP, v. 43, n. 3, p. 405-412, 2009.

ARAÚJO, E. M. **Mortalidade por causas externas e raça/cor da pele:** uma das expressões das desigualdades sociais. 2007. 201f. Tese (Doutorado) – Instituto de Saúde Coletiva, Universidade Federal da Bahia, Salvador, BA, 2007.

ARAÚJO, J. D. Polarização epidemiológica no Brasil. **Epidemiologia e Serviços de Saúde,** Brasília, DF, v. 21, n. 4, p.533-538, 2012.

ARPINO, B.; LE MOGLIE, M.; MENCARINI, L. Machine-learning techniques for family demography: an application of random forests to the analysis of divorce determinants in Germany. In: ANNUAL MEETING PAA, 83., 2018, Denver. **Anais...** New York, NY: PAA, 2018.

ARRETCHE, M. **Trazendo o conceito de cidadania de volta:** a propósito das desigualdades territoriais. In: ARRETCHE, M. (org.). Trajetórias das desigualdades no Brasil: como o Brasil mudou nos últimos 50 anos. São Paulo, SP: Editora da Unesp, 2015. p. 193-223.

ARRIAGA, E. E; DAVIS, K. The pattern of mortality change in Latin America. **Demography,** New York, NY, v. 6, n. 3, p. 223-242, 1969.

ASSUNÇÃO, R. M. et al. Empirical bayes estimation of demographic schedules for small areas. **Demography,** New York, NY, v. 42, n. 3, p. 537-558, 2005.

AWAD, M.; KHANNA, R. **Efficient learning machines:** theories, concept, and applications for engineers and system designers. New York, NY: Apress Open, 2015.

BARATA, R. B. **Como e por que as desigualdades sociais fazem mal à saúde.** Rio de Janeiro, RJ: Editora FIOCRUZ, 2009. (Temas em Saúde Collection).

BARRETO, M. L. et al. Mudanças dos padrões de morbi-mortalidade: uma revisão crítica das abordagens epidemiológicas. **Physis – Revista de Saúde Coletiva,** Rio de Janeiro, RJ, v. 3, n. 1, p.127-146, 1993.

BARROS, R. P. et al. A queda recente da desigualdade de Renda no Brasil. In: BARROS, R. P.; FOGUEL, M.; ULYSSEA, G. (org.) **Desigualdade de renda no Brasil:** uma análise da queda recente. Brasília, DF: IPEA, v. 1, 2007. p. 107-127.

BEER, J. Smoothing and projecting age-specific probabilities of death by TOPALS. **Demographic Research,** Germany, v. 27, n. 20, p. 543-592, 2012.

BELON, A. P.; BARROS, M. B.; MARÍN-LEÓN, L. Mortality among adults: gender and socioeconomic differences in a Brazilian city. **BMC Public Health,** London, v. 12, n. 1, p. 1-10, 2012.

BENNETT, N. G.; HORIUCHI, S. Estimating the completeness of death registration in a closed population. **Population Index,** Princeton, v. 47, n. 2, p. 207-221, 1981.

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.

BORRELL, C. et al. Socioeconomic inequalities in mortality in 16 European cities. **Scandinavian Journal of Public Health,** v. 42, n. 3, p. 245-254, 2014.

BOSWORTH, B. Increasing disparities in mortality by socioeconomic status. **Annual Review of Public Health,** California, v. 39, n. 1, p. 237-251, 2018.

BRASIL. Ministério da Saúde. Programa Nacional de Imunizações: aspectos históricos dos calendários de vacinação e avanços dos indicadores de coberturas vacinais, no período de 1980 a 2013. **Boletim Epidemiológico,** Brasília, DF, v. 46, n. 30, p. 1-13, 2015a.

BRASIL. Ministério da Saúde. **Saúde Brasil 2014:** uma análise da situação de saúde e das causas externas. Brasília, DF, 2015b.

BRASIL. Ministério da Saúde. **Datasus.** Brasília, DF, 2010.

BRASIL. Ministério da Saúde. **Saúde Brasil 2007:** uma análise da situação de saúde. Brasília, DF, 2008.

BRASIL. Ministério da Saúde. Secretaria de Vigilância em Saúde. Análise da morte violenta segundo raça/cor. In: BRASIL. **Saúde Brasil, 2005.** Brasília, DF, 2005.

CAMPOS, M. R. et al. Morbidity and mortality associated with injuries: results of the Global Burden of Disease study in Brazil, 2008. **Cadernos de Saúde Pública,** Rio de Janeiro, RJ, v. 31, n. 1, p. 121-136, 2015.

CAMPOS, N. O. B.; RODRIGUES, R. N. Ritmo de declínio nas taxas de mortalidade dos idosos nos Estados do Sudeste, 1980-2000. **Revista Brasileira de Estudos Populacionais,** Campinas, SP, v. 21, n. 2, p. 323-342, 2004.

CARPENTER, B. et al. Stan: a probabilistic programming language. **Journal of Statistical Software,** Austria, v. 76, n. 1, p. 1-32, 2017.

CARVALHO, F. R. D. **Análise fatorial.** 2013. 61f. Dissertação (Mestrado) – Curso de Matemática, Departamento de Matemática, Universidade de Coimbra, Coimbra, 2013.

CASELLI, G.; MESLE, F.; VALLIN, J. Epidemiologic transition theory exceptions. **Genus**, Roma, v. 58, n. 1, p. 1-34, 2002.

CASTRO, M. C.; SIMÕES, C. C. Spatio-temporal trends of infant mortality in Brazil. In: INTERNATIONAL POPULATION CONFERENCE – IUSSP, 26., Marrakech, Morocco. **Anais…** Paris: IUSSP, 2009.

CESSE, E. A. P. et al. Tendência da mortalidade por doenças do aparelho circulatório no Brasil: 1950 a 2000. **Arquivos Brasileiros de Cardiologia,** São Paulo, SP, v. 93, n. 5, p. 490-497, 2009.

CHAIMOWICZ, F. A saúde dos idosos brasileiros às vésperas do século XXI: problemas, projeções e alternativas. **Revista de Saúde Pública,** São Paulo, SP, v. 31, n. 2, p. 184-200, 1997.

CLEMENS, T.; POPHAM, F.; BOYLE, P. What is the effect of unemployment on all-cause mortality? A cohort study using propensity score matching. **European Journal of Public Health,** Praga, v. 25, n. 1, p. 115-121, 2015.

CLOSS, V. E.; SCHWANKE, C. H. A. A evolução do índice de envelhecimento no Brasil, nas suas regiões e unidades federativas no período de 1970 a 2010. **Revista Brasileira de Geriatria e Gerontologia,** Rio de Janeiro, RJ, v. 15, n. 3, p. 443-458, 2012.

CUNNINGHAM, T. J. et al. Vital signs: racial disparities in age-specific mortality among blacks or African Americans: United States, 1999–2015. **Morbidity and Mortality Weekly Report,** Atlanta, v. 66, n. 17, p. 444-456, 2017.

CUTLER, D. M.; DEATON, A. S.; LLERAS-MUNEY, A. **The determinants of mortality.** Cambridge: National Bureau of Economic Research, 2006. (NBER Working Paper, n. 11963).

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.

DE ROSE, A.; PALLARA, A. Survival trees: an alternative non-parametric multivariate technique for life history analysis. **European Journal of Population,** Amsterdam, v. 13, n. 3, p. 223-241, 1997.

DIVINO, F.; EGIDI, V.; SALVATORE, M. A. Geographical mortality patterns in Italy: a bayesian analysis. **Demographic Research,** Germany, v. 20, n. 18, p. 435-466, 2009.

DOUGHERTY, J.; KOHAVI, R.; SAHAMI, M. Supervised and unsupervised discretization of continuous features. In: TWELFTH INTERNATIONAL CONFERENCE ON MACHINE LEARNING, 1995, Tahoe City. **Proceedings...** California: Morgan Kaufmann, 1995.

DUARTE, E. C. et al. Expectativa de vida ao nascer e mortalidade no Brasil em 1999: análise exploratória dos diferenciais regionais. **Revista Panamericana de Salud Pública,** Washington, DC, v. 12, n. 6, p. 436-444, 2002.

ELO, I. T. Mortality: transitions and measures. In: RITZER, G. (ed.). **The blackwell encyclopedia of sociology.** Malden, MA: Blackwell Pub., 2007.

FERREIRA, F. H. G. et al. Ascensão e queda da desigualdade de renda no Brasil: uma atualização para 2005. In: BARROS, R. P.; FOGUEL, M.; ULYSSEA, G. (org.). **Desigualdade de renda no Brasil:** uma análise da queda recente. Brasília, DF: IPEA, v. 1, 2007. p. 359-378.

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.

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.

FRENK, J.; LOZANO, R.; BOBADILLA, J. L. La transicion epidemiologica en America Latina. **Notas de Población,** Santiago de Chile, Chile, n. 60. p. 79-101, 1994.

FRIES, J. F. Aging, natural death, and the compression of morbidity. **New England Journal of Medicine,** Boston, v. 303, n. 3, p. 130-135, 1980.

GARCY, A. M.; VÅGERÖ, D. The length of unemployment predicts mortality, differently in men and women, and by cause of death: a six year mortality follow-up of the Swedish 1992–1996 recession. **Social Science & Medicine,** Oxford, v. 74, n. 12, p. 1911-1920, 2012.

GAWRYSZEWSKI, V. P. et al. A mortalidade por causas externas no Estado de São Paulo em 2005. **Boletim Epidemiológico Paulista,** São Paulo, SP, v. 3, n. 33, 2006.

GAWRYSZEWSKI, V. P.; HIDALGO, N. T.; VALENICH, D. M. O. A queda nas taxas de homicídios no Estado de São Paulo e apresentação dos dados de mortalidade por causas externas em 2004. **Boletim Epidemiológico Paulista,** São Paulo, SP, v. 2, n. 21, 2005.

GERDTHAM, U. G.; JOHANNESSON, M. A note on the effect of unemployment on mortality. **Journal of Health Economics,** Amsterdam, v. 22, n. 3, p. 505-518, 2003.

GONZAGA, M. R.; QUEIROZ, B. L.; LIMA, E. E. C. Compression of mortality: the evolution in the variability in the age of death in Latin America. **Revista Latinoamericana de Población,** Rio de Janeiro, RJ, v. 12, n. 23, p. 9-35, 2018.

GONZAGA, M. R.; SCHMERTMANN, C. P. Estimating age- and sex-specific mortality rates for small areas with TOPALS regression: an application to Brazil in 2010. **Revista Brasileira de Estudos de População,** Campinas, SP, v. 33, n. 3, p. 629-652, 2016.

HILL, K.; YOU, D.; CHOI, Y. Death distribution methods for estimating adult mortality: sensitivity analysis with simulated data errors. **Demographic Research,** Germany, v. 21, n. 9, p, 235-254, 2009.

HILL, K. Estimating census and death registration completeness. **Asian and Pacific Census Forum,** Honolulu, v. 1, n. 3, p. 8-13; 23-24, 1987.

HORIUCHI, S. Epidemiological transitions in human history. In: CHAMIE, J.; CLIQUET, R. L. (ed.). **Health and mortality:** issues of global concern. New York, NY: United Nations, 1999. p. 54-71. (Proceedings of the Symposium on Health and Mortality).

HUDSON, V. M.; BOER, A. D. A surplus of men, a deficit of peace: security and sex ratios in Asia's largest states. **International Security,** Cambridge, v. 26, n. 4, p.5-38, 2002.

HUMMER, R. A.; ROGERS, R. G.; EBERSTEIN, I. W. Sociodemographic differentials in adult mortality: a review of analytic approaches. **Population and Development Review,** New York, NY, v. 24, n. 3, p. 553-578, 1998.

IBGE – INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. **Censo Demográfico 2010:** resultados gerais da amostra. Rio de Janeiro, RJ, 2010a.

IBGE – INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. **Censo Demográfico 2010:** características gerais da população religião e pessoas com deficiência. Rio de Janeiro, RJ, 2010b.

IBGE – INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. **Observações sobre a evolução da mortalidade no Brasil:** o passado, o presente e perspectivas. Rio de Janeiro, RJ, 2010c.

IPEA – INSTITUTO DE PESQUISA ECONÔMICA APLICADA; FORÚM BRASILEIRO DE SERGURANÇA PÚBLICA – FBSP. **Atlas da violência 2018.** Brasília, DF, 2018.

IPEA – INSTITUTO DE PESQUISA ECONÔMICA APLICADA. **Retrato das desigualdades de gênero e raça.** 4. ed. Brasília, DF, 2011.

ISHITANI, L. H. et al. Desigualdade social e mortalidade precoce por doenças cardiovasculares no Brasil. **Revista de Saúde Pública,** São Paulo, SP, v. 40, n. 4, p. 684-691, 2006.

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

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

JIANG, F. et al. Artificial intelligence in healthcare: past, present and future. **Stroke and Vascular Neurology,** v. 2, n. 4, p. 230-243, 2017.

KAPLAN, G. A. et al. Inequality in income and mortality in the United States: analysis of mortality and potential pathways. **British Medical Journal, London,** v. 312, n. 7037, p. 999-1003, 1996.

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

KUHN, M. Building predictive models in R using caret package. **Journal of Statistical Software,** v. 28, n. 5, p. 1-26, 2008.

KUNITZ, S. J. Explanations and ideologies of mortality patterns. **Population and Development Review,** New York, NY, v. 13, n. 3, p. 379-498, 1987.

KUNST, A. E.; WOLLESWINKEL-VAN DEN BOSCH, J. H.; MACKENBACH, J. P. Medical demography on the Netherlands: recent advances, future challenges. In: VAN WISSEN, L. J. G.; DYKSTRA, P. A. (ed.). **Population issues:** an interdisciplinary focus. New York, NY: Kluwer Academic/Plenum Publishers, 1999.

LERNER, M. **The health transition in social change:** unpublished document. Baltimore: The Johns Hopkins School of Hygiene and Public Health, 1972.

LIMA, M. L. C. et al. Análise espacial dos determinantes socioeconômicos dos homicídios no Estado de Pernambuco. **Revista de Saúde Pública,** São Paulo, SP, v. 39, n. 2, p. 176-182, 2005.

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.

LOVELL, P. A. Development and the persistence of racial inequality in Brazil: 1950-1991. **Journal of Developing Areas,** US, v. 33, n. 3, p. 395-418, 1999.

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

MAATEN, L. V. der; HINTON, G. Visualizing data using t-SNE. **Journal of Machine Learning Research,** v. 9, p. 2579-2605, 2008.

MACINKO, J.; GUANAIS, F. C.; SOUZA, M. F. M. Evaluation of the impact of the family health program on infant mortality in Brazi, 1990-2002. **Journal of Epidemiology and Community Health,** London, v. 60, n. 1, p. 13-19, 2006.

MACKENBACH, J. P.; KUNST, A. E. Socio-economic and cultural determinants of regional mortality patterns in the Netherlands. In: LOPEZ, A. D.; CASELLI, G.; VALKONEN, T. (ed.). A**dult mortality in developed countries:** from description to explanation. Oxford: Claredon Press, 1995. p. 286-303.

MACKENBACH, J. P.; LOOMAN, C. W. Living standards and mortality in the European community. **Journal of Epidemiology & Community Health,** London, v. 48, n. 2, p. 140-145, 1994.

MARTIKAINEN, P. Does unemployment cause mortality? **Bulletin of the International Statistical Institute,** Roma, 1999, p. 59-62.

MARTINE, G.; CARVALHO, J. A. M; ALFONSO, A. R. **Mudanças recentes no padrão demográfico brasileiro e implicações para a agenda social.** Brasília, DF: IPEA, 1994. (Textos para Discussão, n. 345).

MASTERS, R. K. et al. Long-term trends in adult mortality for U.S. blacks and whites: an examination of period- and cohort-based changes. **Demography,** New York, NY, v. 51, n. 6, p. 2047-2073, 2014.

MCKEE, M. For debate - does health care save lives? **Crotia Medical Journal,** v. 40, n. 2, p. 123-128, 1999.

MEDEIROS, C. R. G.; MENEGHEL, S. N.; GERHARDT, T. E. Desigualdades na mortalidade por doenças cardiovasculares em pequenos municípios. **Ciência & Saúde Coletiva,** Rio de Janeiro, RJ, v. 17, n. 11, p. 2953-2962, 2012.

MELLO JORGE, M. H. P.; GAWRYSZEWSKI, V. P.; LATORRE, M. R. D. O. Análise dos dados de mortalidade. **Revista de Saúde Pública,** São Paulo, SP, v. 31, n. 4, p. 5-25, 1997.

MELLO, J. M. P.; SCHNEIDER, A. Mudança demográfica e a dinâmica dos homicídios no Estado de São Paulo. **São Paulo em Perspectiva,** São Paulo, SP, v. 21, n. 1, p. 19-30, 2007.

MENDES, L. V. P. et al. A evolução da carga de causas externas no Brasil: uma comparação entre os anos de 1998 e 2008. **Cadernos de Saúde Pública,** Rio de Janeiro, RJ, v. 31, n. 10, p. 2169-2184, 2015.

MENEZES-FILHO, N.; FERNANDES, R.; PICHETTI, P. Educação e queda recente da desigualdade no Brasil. In: BARROS, R. P.; FOGUEL, M.; ULYSSEA, G. (org.). **Desigualdade de renda no Brasil:** uma análise da queda recente. Brasília, DF: IPEA, v. 1, 2007. p. 285-304.

MOLNAR, C. **Interpretable machine learning:** a guide for making black box models explainable. Canada: LeanPub, 2018.

MONTEZ, J. K. Educational disparities in adult mortality across U.S. States: how do they differ, and have they changed since the mid-1980s? **Demography,** New York, NY v. 56, n. 2, p. 621-644, 2019.

MONTGOMERY, S. et al. Mortality following unemployment during an economic downturn: swedish register-based cohort study. **British Medical Journal Open,** v. 3, n. 7, p. 1-9, 2013.

MORAES, J. C.; RIBEIRO, M. C. S. A. Desigualdades sociais e cobertura vacinal: uso de inquéritos domiciliares. **Revista Brasileira de Epidemiologia,** São Paulo, SP, v. 11, n. 1, p. 113-124, 2008.

MOSLEY, W. H.; CHEN, L. An analytical framework for the study of child survival in developing countries. **Population and Development Review,** New York, NY, v. 10, Supl., p. 25-45, 1984.

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.

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.

MOURA, R.; OLIVEIRA, S.; PÊGO, B. **Escalas da urbanização brasileira.** Brasília, DF: IPEA, 2018. (Texto para Discussão, n. 2372).

MURRAY, C. J. L. et al. What can we conclude from death registration? Improved methods for evaluating completeness. **Plos Medicine,** v. 7, n. 4, p. 1-27, 2010.

NGUYEN, G. **Evaluating statistical and machine learning methods to predict risk of in-hospital child mortality in Uganda.** 2016. 23f. Dissertação (Master) – Public Health, University Of Washington, Washington, DC, 2016.

NORONHA, K. V. M. S; ANDRADE, M. V. O efeito da distribuição de renda sobre o estado de saúde individual no Brasil. **Pesquisa e Planejamento Econômico,** Rio de Janeiro, RJ, v. 37, n. 3, p. 521-555, 2007.

NURU-JETER, A. M.; LAVEIST, T. A. Racial segregation, income inequality, and mortality in US metropolitan areas. **Journal of Urban Health,** New York, NY, v. 88, n. 2, p. 270-282, 2011.

OEPPEN, J.; VAUPEL, J. W. Broken limits to life expectancy. **Science,** Washington DC, v. 296 n. 10, p. 1029-1031, 2002.

OLIVEIRA, B. L. C. A.; THOMAZ, E. B. A. F.; SILVA, R. A. The association between skin color/race and health indicators in elderly Brazilians: a study based on the Brazilian National Household Sample Survey (2008). **Cadernos de Saúde Pública,** Rio de Janeiro, RJ, v. 30, n. 7, p. 1438-1452, 2014.

OLSHANSKY, S. J.; AULT, A. B. The fourth stage of the epidemiologic transition: the age of delayed degenerative diseases. **The Milbank Quarterly,** New York, NY, v. 64, n. 3, p. 355-391, 1986.

OMRAN, A. R. The epidemiologic transition: a theory of the epidemiology of population change. **The Milbank Memorial Fund Quartely,** New York, NY, v. 4, n. 49, p. 509-538, 1971.

PALLONI, A.; PINTO-AGUIRRE, G. Adult mortality in Latin America and the Caribbean. In: ROGERS, R. G.; CRIMMINS, E. M. (ed.). **International handbook of adult mortality.** New York, NY: Springer, 2011. p. 101-132.

PAN, I. et al. Machine learning for social services: a study of prenatal case management in Illinois. **American Journal of Public Health,** New York, NY, v. 107, n. 6, p. 938-944, 2017.

PEREIRA, F. N. A.; QUEIROZ, B. L. Diferenciais de mortalidade jovem no Brasil: a importância dos fatores socioeconômicos dos domicílios e das condições de vida nos

municípios e estados brasileiros. **Cadernos de Saúde Pública,** Rio de Janeiro, RJ, v. 32, n. 9, p. 1-12, 2016.

PEREIRA, F. N. A. **Diferenciais de mortalidade jovem no Brasil:** a importância dos fatores socioeconômicos dos domicílios e das condições de vida nos municípios e UFs. 2014. 122f. Dissertação (Mestrado) – Curso de Demografia, Centro de Desenvolvimento e Planejamento Regional, Universidade Federal de Minas Gerais, Belo Horizonte, MG, 2014.

PERES, M. F. T.; CARDIA, N.; SANTOS, P. C. **Homicídios de crianças e jovens no Brasil:** 1980 a 2002. São Paulo, SP: NEV/USP, 2006. Disponível em: http://nevusp.org/wp-content/uploads/2015/01/down095.pdf.

RENTERÍA PÉREZ, E.; TURRA, C. M. **Desigualdade social na mortalidade no Brasil:** diferenciais por escolaridade entre mulheres adultas. In: ENCONTRO NACIONAL DE ESTUDOS POPULACIONAIS, 16., 2008, Caxambu, MG. **Anais...** Belo Horizonte, MG, 2008.

PETRUCELLI, J. L.; SABOIA, A. L. (or.g). **Características étnico-raciais da população:** classificações e identidades. Rio de Janeiro, RJ: IBGE, 2013. (Estudos e Análises: Informação Demográfica e Socioeconômica, n. 2).

PRATA, P. R. A transição epidemiológica no Brasil. **Cadernos de Saúde Pública**, Rio de Janeiro, RJ, v. 8, n. 2, p. 168-175, 1992.

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.

ROELFS, D. J. et al. Losing life and livelihood: a systematic review and meta-analysis of unemployment and all-cause mortality. **Social Science & Medicine,** Oxford, v. 72, n. 6, p. 840-854, 2011.

ROGGERO, P et al. The health impact of child labor in developing countries: evidence from cross-country data. **American Journal of Public Health,** New York, NY, v. 97, n. 2, p. 271-275, 2007.

SAEZ, Y.; BALDOMINOS, A.; ISASI, P. A comparison study of classifier algorithms for cross-person physical activity recognition. **Sensors,** Switzerland, v. 17, n. 12, p. 66-92, 2016.

SAKR, S. et al. Comparison of machine learning techniques to predict all-cause mortality using fitness data: the Henry ford exercIse testing (FIT) project. **BMC Medical Informatics and Decision Making,** London, v. 17, n. 1, p. 1-15, 2017.

SANTANA, P. et al. Mortality, material deprivation and urbanization: exploring the social patterns of a metropolitan area. **International Journal for Equity in Health,** London, v. 14, n. 1, p. 1-13, 2015.

SANTOS, G. R.; PALES, R. C.; RODRIGUES, S. G. Desigualdades regionais no Brasil – 1991-2010. **InterScience Place,** Campos dos Goytacazes, RJ, v. 1, n. 31, p. 145-173, 2014.

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

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.

SHI, L. et al. Primary care, infant mortality, and low birth weight in the states of the USA. **Journal of Epidemiology & Community Health,** London, v. 58, n. 5, p. 374-380, 2004.

SILVA, A. F.; SILVA, V. P. Nos limites do viver: moradia e segregação socioespacial nas áreas metropolitanas do Nordeste brasileiro. **Scripta Nova – Revista Electrónica de Geografía y Ciencias Sociales.** Barcelona, v. VII, n. 146, 2003.

SILVA, L. E.; FREIRE, F. H. M.; PEREIRA, R. H. M. Diferenciais de mortalidade por escolaridade da população adulta brasileira em 2010. **Caderno de Saúde Pública,** Rio de Janeiro, RJ, v. 32, n. 4, p. 1-12, 2016.

SIMÕES, C. C. S. **Perfis de saúde e de mortalidade no Brasil:** uma análise de seus condicionantes em grupos populacionais específicos. Brasília, DF: OPAS, 2002.

SINGH, G. K.; YU, S. M. Trends and differentials in adolescent and young adult mortality in the United States, 1950 through 1993. **American Journal of Public Health,** New York, NY, v. 86, n. 4, p. 560-564, 1996.

SINGHA, A. K. et al. Application of machine learning in analysis of infant mortality and its factors. **Working Paper,** India, p. 1-5, 2016.

SOARES, R. R. **On the determinants of mortality reductions in the developing world.** Cambridge: National Bureau of Economic Research, 2007. (NBER Working Paper, n. 12837).

SOUSA, M. F.; HAMANN, E. M. Programa saúde da família no Brasil: uma agenda incompleta? **Ciência & Saúde Coletiva,** Rio de Janeiro, RJ, v. 14, n. 1, p. 1325-1335, 2009.

SPIJKER, J. J. A. **Socioeconomic determinants of regional mortality differences in Europe.** 2004. 325f. Thesis (Doctor of Philosophy) – University of Groningen, 2004.

STEPHENS, C. et al. **Environment and health in developing countries:** an analysis of intra-urban differentials using existing data. São Paulo, SP: Fundação SEADE; London: London School of Hygiene & Tropical Medicine, 1994.

THATCHER, R.; KANNISTO, V.; ANDREEV, K. The survivor ratio method for estimating numbers at high ages. **Demographic Research,** Germany, v. 6, n. 1, p. 1-18, 2002.

TIMAEUS, I. M. Measurement of adult mortality in less developed countries: a comparative review. **Population Index,** Princeton, NJ, v. 57, n. 4, p. 552-568, 1991.

TIMAEUS, I. M.; CHACKIEL, J.; RUZICKA, L. **Adult mortality in Latin America.** New York, NY: Clarendon Press, 1996.

TINEU, R.; BORGES, C. M. M. Desigualdade e segregação socioespacial da população negra na cidade de São Paulo. **Revista Belas Artes,** São Paulo, SP, n. 22, 2016.

TRAVASSOS, C.; MARTINS, M. Uma revisão sobre os conceitos de acesso e utilização de serviços de saúde. **Cadernos de Saúde Pública,** Rio de Janeiro, RJ, v. 20, n. 2, p. 190-198, 2004.

VALLIN, J.; MESLÉ, F. Convergences and divergences in mortality. A new approach to health transition. **Demographic Research,** Germany, Special Collection 2, Article 2, p. 11-44, 2004.

VAN POPPEL, F. W. A. Roman catholicism and regional mortality differences in the Netherlands, 1850-1933. In: EUROPEAN POPULATION CONFERENCE, 1991, Paris. **Anais…** 1991.

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. **Anais…** 2012.

VASCONCELOS, A. M. N.; GOMES, M. M. F. Transição demográfica: a experiência brasileira. **Epidemiologia e Serviços de Saúde,** Brasília, DF, v. 21, n. 4, p. 539-548, 2012.

VAUPEL, J. W. Biodemography of human ageing. **Nature**, London, v. 464, n. 7288, p. 536-542, 2010.

VICTORA, C. G. et al. Explaining trends in inequities: evidence from Brazilian child health studies. **The Lancet,** London, v. 356, n. 9235, p. 1093-1098, 2000.

WAISELFISZ, J. J. **Mapa da violência 2014:** homicídios e juventude no Brasil. Brasília, DF: Secretaria Nacional de Juventude, 2014.

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.

WIILEKENS, F. J. **Demographic transitions in Europe and the world.** Germany: MPIDR, 2014. **(**Working Paper).

WILLEKENS, F. J. Demographic forecasting: state of the art and research needs. In: FRINKING, G. A. B.; HAZEAU, C. A. (ed.). **Emerging issues in demographic research.** Amsterdam: Elsevier, 1990.

WILLIAMS, C. et al. Effects of economic crises on population health outcomes in Latin America, 1981-2010: an ecological study. **The BMJ,** London, v. 6, n. 1, 2016.

WILLIAMS, D. R.; PRIEST, N.; ANDERSON, N. B. Understanding associations among race, socioeconomic status, and health: patterns and prospects. **Health Psychology,** US, v. 35, n. 4, p. 407-411, 2016.

WILLIAMS, D. R. Miles to go before we sleep: racial inequities in health. **Journal of Health and Social Behavior,** US, v. 53, n. 3, p. 279-295, 2012.

WINGARD, D. L. The sex differential in morbidity, mortality, and lifestyle. **Annual Review of Public Health,** California, v. 5, n. 1, p. 433-458, 1984.

WHO – WORD HEALTH ORGANIZATION. Urbanization and health. **Bulletin World Health Organization,** Geneva, v. 88, n. 4, p. 241-320, 2010.

WOLPERT, D. H. The lack of a prior distinctions between learning algorithms and the existence of a priori distinctions between learning algorithms. **Neural Computation,** Cambridge, v. 8, p. 1341-1390, 1996.

WOOD, C. H.; CARVALHO, J. A. M. **A demografia da desigualdade no Brasil.** Brasília, DF: IPEA, 1994.

ZIJDEMAN, R.; SILVA, F. R. Life expectancy at birth (Total). **IISH Dataverse,** 2015. Disponível em: http://hdl.handle.net/10622/LKYT53.

ANEXO A – Códigos de implementação dos métodos

```
## Rotina com as funções utilizadas na Seção 3.1.1: - ACP
#                                                  - t-SNE
#                                                  - K-Means

#### Pacotes utilizados ####

library(caret); library(tidyverse); library(data.table); library(pROC); library(ROCR);
library(klaR); library(corrplot); library(e1071); library(readxl); library(hydroGOF);
library(lattice); library(psych); library(reshape); library(rpart.plot); library(rpart);
library(MLmetrics); library(ggRandomForests); library(dplyr); library(mclust);
library(gbm); library(ipred); library(randomForest); library(Rtsne); library(FactoMineR)
library(pdp); library(ALEPlot)

################################### ACP ##########################
# Análise de Componentes Principais
### função ACP ###
res.pca <- PCA(banco[3:23], graph = FALSE, scale.unit = TRUE) #gerando ACP
eig.val <- get_eigenvalue(res.pca) ### eigenvalues (autovalores)
data_eig<-as.matrix(res.pca$eig)
var <- get_pca_var(res.pca) #### Principal Component Analysis Results for variables

## Gráfico 7 – Porcentagem da variância total explicada por cada componente
fviz_eig(res.pca, addlabels = TRUE, ylim = c(0, 50))

## Gráfico 8 – Contribuição das variáveis para cada componente principal
# Contributions of variables to PC1
dim1<-fviz_contrib(res.pca, choice = "var", axes = 1, top = 15)+ggtitle("A) Componen
te Principal 1")+
                        ylab("Contribuições (%)")+theme(axis.text=element_text(size=9))
# Contributions of variables to PC2
dim2<-fviz_contrib(res.pca, choice = "var", axes = 2, top = 15)+ggtitle("B) Componen
te Principal 2")+ ylab("Contribuições (%)")+theme(axis.text=element_text(size=9))

# Contributions of variables to PC3
dim3<-fviz_contrib(res.pca, choice = "var", axes = 3, top = 15)+ggtitle("C) Componen
te Principal 3")+ylab("Contribuições (%)")+theme(axis.text=element_text(size=9))

# Contributions of variables to PC4
dim4<-fviz_contrib(res.pca, choice = "var", axes = 4, top = 15)+ggtitle("D) Componen
te Principal 4")+ylab("Contribuições (%)")+theme(axis.text=element_text(size=9))

################################### t-SNE ############################

data <- read.table ("(Corrigido)Banco_ML.csv", header = T ,sep=';', dec=',')
data_filter <- dplyr::select(banco,-cod, -qx) #Retirar variavel de codigo para o t-SNE

#Padronizar(calcular o z escore para todas as variaveis)
quantis_scale<- as.data.frame(scale(data_filter))
```

```
qx<-banco$qx
cod<-banco$cod
data <- cbind(cod, qx, quantis_scale)

tsne<- Rtsne(data[3:23], dims=2, theta=0.5, perplexity=100, iterations=100000, learning=
10000) ## algoritmo t-SNE


plot(tsne$Y, t='n', main="tSNE", xlab="Dimensão 1", ylab="Dimensão 2", "cex.main"=2,
 "cex.lab"=1.5)
points(tsne$Y) ## coordenadas criadas pelo t-SNE

#### função para K-means usando as coordenadas criadas pelo algoritmo t-SNE
d_tsne_1 = as.data.frame(tsne$Y)
d_tsne_1_original=d_tsne_1

fit_cluster_kmeans=kmeans(scale(d_tsne_1), 2)
d_tsne_1_original$cl_kmeans = factor(fit_cluster_kmeans$cluster)

plot_cluster=function(data, var_cluster, palette)
{
  ggplot(data, aes_string(x="V1", y="V2", color=var_cluster)) +
    geom_point(size=2, show.legend = TRUE) +
    guides(colour=guide_legend(override.aes=list(size=6))) +
    xlab("") + ylab("") +
    ggtitle("") +
    theme_light(base_size=20) +
    theme(axis.text.x=element_blank(),
          axis.text.y=element_blank(),
          legend.direction = "horizontal",
          legend.position = "bottom",
          legend.box = "horizontal") +
    scale_colour_brewer(palette = palette)
}

## Gráfico 9 – Divisão dos dados em dois grupos, K= 2 usando K-Means
plot_k=plot_cluster(d_tsne_1_original, "cl_kmeans", "Accent")

######## Rotina com as funções usadas na Seção 3.2: - Random Forest
#                                                   - Extreme Boosted Trees
#                                                   - Naive Bayes
#                                                   - Support Vector Machine


########################## PRÉ-PROCESSAMENTO ######################
#Treinamento de modelos preditivos
##Tecnica de reamostragem (para evitar sobreajuste)
#to obtain different performance measures (accuracy, Kappa, area under ROC curve, s
ensitivity, specificity)
fiveStats <- function(...)c(twoClassSummary(...),defaultSummary(...),prSummary(...),m
```

```
nLogLoss(...))

#função controle usada para os quatro algoritmos
set.seed(123)
ctrl<- trainControl(method = "cv", #validação cruzada 10 subamostras
                    number = 10, #número de subamostras
                    savePredictions = TRUE, # salva as informações da reamostragem
                    classProbs = TRUE, #calcula a prob de cada classe
                    summaryFunction = fiveStats, #sumariza as informações
                    verboseIter = TRUE) #A logical for printing a training log

############################ 1. RANDOM FOREST ########################

#objeto (rfqx) com a função gerada para o método Random Forest (method="rf")
rfqx <- train(qx ~ ., data=banco,
              method="rf",
              trControl=ctrl,
              ntree=1000,
              preProc=c("center", "scale"),
              metric="ROC",
              importance=T, proximity = TRUE)

rfqxPred<-as.data.frame(rfqx$pred[,1:2]) #cria como data frame os valores preditos e
observados usando k-fold cross validation

### calculca matriz de confusão e principais medidas de desempenho
confMatRF <- confusionMatrix(rfqxPred$pred, rfqxPred$obs, mode="prec_recall")

## Gráfico de Importância para o modelo Random Forest
varimp_RF<-varImp(rfqx, scale=TRUE)
plot(varimp_RF, top=20,scales=list(y=list(cex=.95)))

################ 2. EXTREME BOOSTED GRADIENT TREES ###############

#objeto (xgbqx) com a função gerada para o método XBG (method="xgbTree")
set.seed(134)
xgbqx <- train(qx ~ ., data=banco,
               method="xgbTree",
               trControl=ctrl,
               preProcess=c("center","scale"),
               metric="ROC", plot=TRUE,importance=TRUE)
plot(xgbqx)
boosted<-xgbqx$finalModel$params
xgb.plot.tree(model = xgbqx$finalModel, trees = 0:10)
xgbqxPred<-as.data.frame(xgbqx$pred[,1:2]) #create as data frame the observed and p
redicted values

### calculca matriz de confusão e principais medidas de desempenho
```

```
confMatXGB <- confusionMatrix(xgbqxPred$pred, xgbqxPred$obs,mode = "prec_recall")

## Gráfico de Importância para o modelo XGB
varimp_XGB<-varImp(xgbqx, scale=FALSE)
plot(varimp_XGB, top=20)

############################## 3. NAIVE BAYES ##############################
#objeto (NBqx) com a função gerada para o método NB (method="naive_bayes")
set.seed(134)
NBqx <- train(qx ~ ., data=banco,
                    method="naive_bayes",
                    trControl=ctrl,
                    preProcess=c("center","scale"),
                    metric="ROC",
                     importance=TRUE)

plot.train(NBqx)

NBqxPred<-as.data.frame(NBqx$pred[,1:2]) #cria como data frame os valores preditos
 e observados usando k-fold cross validation

### calculca matriz de confusão e principais medidas de desempenho
confMatNB <- confusionMatrix(NBqxPred$pred, NBqxPred$obs,mode = "prec_recall")

## Gráfico de Importância para o modelo NB

varimp_NB<-varImp(NBqx, scale=TRUE)
plot(varimp_NB, top=21)




###################### 4. SUPPORT VECTOR MACHINE ###############
library(kernlab)
set.seed(202)

sigmaRangeReduced <- sigest(as.matrix(banco[2:22]))
svmRGridReduced <- expand.grid(.sigma = sigmaRangeReduced[1],
                                        .C = 2^(seq(-4, 4)))
set.seed(476)
svmRModel <- train(qx ~., banco,
                           method = "svmRadial",
                           metric = "ROC",
                           preProc = c("center", "scale"),
                           tune = svmRGridReduced,
                           fit = FALSE,
                           trControl = ctrl)
```

```
svmPred<-as.data.frame(svmRModel$pred[,1:2]) #cria como data frame os valores preditos e observados usando k-fold cross validation

### calculca matriz de confusão e principais medidas de desempenho
confMatSVM <- confusionMatrix(svmPred$pred, svmPred$obs, mode = "prec_recall")

## Gráfico de Importância para o modelo SVM

varimp_SVM<-varImp(svmRModel, scale=TRUE)
plot(varimp_SVM, top=20)


########### GRÁFICOS COMPARAÇÃO - REAMOSTRAGEM ##############
resamp <- resamples(list(RF = rfqx,
                         SVM = svmRModel,
                         XGB = xgbqx,
                         NB  = NBqx))
summary(resamp)
        theme1 <- trellis.par.get()
        theme1$plot.symbol$col = rgb(.2, .2, .2, .4)
        theme1$plot.symbol$pch = 16
        theme1$plot.line$col = rgb(1, 0, 0, .7)
        theme1$plot.line$lwd <- 2
        trellis.par.set(theme1)
        bwplot(resamp, layout = c(5, 1))
        splom(resamp)
```

ANEXO B – Glossário

Anos potenciais de vida perdidos – Indicador de mortalidade que determina o efeito das mortes ocorridas precocemente em relação à duração da vida esperada (ARAÚJO, 2007).

Causas externas – Conjunto de agravos à saúde que provocam algum tipo de lesão, seja física, mental ou psicológica, podendo ou não levar ao óbito (DATASUS). Incluem acidentes (Acidentes de Transporte), lesões autoprovacadas intencionalmente (suicídios), agressões, eventos cuja intenção é indeterminada, sequelas de causas externas de morbidade e mortalidade, Fatores suplementares relacionados com as causas de morbidade e de mortalidade classificados em outra parte (evidência de alcoolismo, por exemplo).

Coeficiente de Gini (em desigualdade) – Indicador que visa medir o grau de concentração de renda em determinado grupo.

Coeficiente de Gini (em aprendizado de máquina) – Aplica-se à classificação binária e requer um classificador que possa de alguma forma classificar os exemplos de acordo com a probabilidade de estar em uma classe positiva.

Coorte – Um grupo de pessoas compartilhando uma experiência demográfica temporal comum que é observada ao longo do tempo. Por exemplo, a coorte de nascimentos de 1900 são as pessoas nascidas naquele ano. Há também coortes de casamento, coortes de classes escolares e assim por diante (*Population Reference Bureau's*).

Doenças do aparelho circulatório – Incluem febre reumática aguda, doenças reumáticas crônicas do coração, doenças hipertensivas, doenças isquêmicas do coração, doenças cardíaca pulmonar e da circulação pulmonar, outras formas de doenças do coração, doenças cerebrovasculares, doenças das artérias, das arteríolas e dos capilares, doenças das veias, Doenças das veias, dos vasos linfáticos e dos gânglios linfáticos, não classificadas em outra parte e Outros transtornos, e os não especificados do aparelho circulatório (I95 – I99) (DATASUS).

Equidade (princípio, SUS) – Possui relação direta com os conceitos de igualdade e de justiça. Esse princípio reconhece as diferenças tanto nas condições de vida quanto de saúde de cada pessoa considerando que o direito à saúde deve considerar as diferenças sociais e atender a diversidade (FIOCRUZ).

Incidência – O número de pessoas que contraem uma doença por 1.000 habitantes em risco, por um determinado período de tempo (*Population Reference Bureau's*).

Integralidade (principio SUS) – "Compreensão de que as pessoas têm o direito de serem atendidas no conjunto de suas necessidades, e que os serviços de saúde devem ser organizados de modo a oferecer todas as ações requeridas por essa atenção integral" (ALMEIDA, 2013).

Morbidade – A frequência de doenças, doenças, lesões e deficiências em uma população (*Population Reference Bureau's*).

Morbimortalidade – Refere-se ao número de pessoas mortas em decorrência de uma doença específica dentro de determinado grupo populacional.

Mortalidade – Mortes como um componente da mudança populacional (*Population Reference Bureau's*).

Universalidade (principio, SUS) – determina que todos os cidadãos brasileiros, sem qualquer tipo de discriminação, têm direito ao acesso às ações e serviços de saúde (FIOCRUZ).

Vulnerabilidade – "Conjunto de aspectos individuais e coletivos relacionados ao grau de modo de exposição a uma dada situação e, de modo indissociável, ao maior ou menor acesso a recursos adequados para se proteger das consequências indesejáveis daquela situação" (ARAÚJO, 2007).