https://eventos.utfpr.edu.br//sicite/sicite2018

Modelagem espaço-temporal de tuberculose na região sul do país utilizando DAG e modelos aditivados generalizados

Spatio temporal modeling of tuberculosis in the southern region of DAG and generalized additive models

Tamires Cristina Cassiano
tamirescassiano8@gmail.com
Universidade Tecnológica Federal do Paraná, Cornélio Procópio, Paraná, Brasil

Elisangela Aparecida da Silva Lizzi
elisangelizzi@utfpr.edu.br
Universidade Tecnológica Federal do Paraná, Cornélio Procópio, Paraná, Brasil

RESUMO

Introdução: O uso combinado das áreas de matemática/estatística, epidemiologia, ciência da computação e big data possibilitam estudar informações em saúde de forma inovadora e levam à ações focadas e diferenciadas. Neste trabalho utilizou-se dados de tuberculose, esta doença existe há milhares de anos e configura-se como um problema de saúde pública global. O objetivo é trabalhar com refinamento de informação e relacionar indicadores sociais com as taxas de incidência de tuberculose na região sul do país, ponderando estas informações pelo posicionamento geográfico e o tempo. Métodos: Estudo epidemiológico ecológico com dados referentes aos municípios de residência da região sul do Brasil para os anos de 2011 a 2017. Este trabalho foi desenvolvido em 4 etapas: bigdata, redes bayesianas, modelos aditivados generalizados para espaço e tempo e modelos aditivados generalizados para espaço e tempo com covariáveis. Utilizou-se redes bayesianas para auxiliar no processo de refinamento de informações para estudos de indicadores sociais, pois estes são altamente correlacionados impossibilitando técnicas simples e usuais. Depois utilizou-se modelos generalizados aditivados para entender os relacionamentos considerando espaço e tempo, em seguida fez-se inserção das covariáveis de interesse neste modelo, sendo estas escolhidas pelos resultados das redes. Conclusão: Os indicadores mostraram que ao longo do tempo tem comportamentos distintos, sendo que o IDHM- educação, renda e longevidade aumentaram ao longo dos anos e mortalidade infantil teve queda significativa, os mapas temáticos das taxas de tuberculose mostram que a doença não está estável no território estadual, podendo indicar municípios para ações prioritárias em saúde.

PALAVRAS-CHAVE: Estatística. Big data. Tuberculose. Modelos aditivados generalizados.

ABSTRACT

Introduction: The combined use of math/statistics, epidemiology, computer science, and big data makes it possible to study health information in an innovative way and lead to focused and differentiated actions. In this work we used data from tuberculosis, this disease has existed for thousands of years and is a global public health problem. The goal is to work with information refinement and to relate social indicators to the incidence rates of tuberculosis in the southern region of the country, pondering this information by geographic positioning and time. Methods: An ecological epidemiological study with data referring to the municipalities of residence of the southern region of Brazil for the years 2011 to 2017. This work was develop in 4 stages: big data, Bayesian networks, generalized additive models for space and time and generalized additive models for space and time with covariates. Bayesian networks were used to aid in the process of information refinement for studies of social indicators, since these are highly correlated, making simple and usual techniques impossible. Then generalized models were add to understand the relationships considering space and time, then the covariates of interest in this model were inserted, being chosen by the results of the networks. Conclusion: The indicators showed that over time have different behaviors, being that the HDI-education, income and longevity increased over the years and infant mortality had a significant fall, the thematic maps of tuberculosis rates show that the disease is not stable in the state territory, and may indicate municipalities for priority health actions.

KEYWORDS: Statistic**.** Big Data.Tuberculosis. Generalized additive models

Recebido: 31 ago. 2018.
Aprovado: 04 out. 2018.

INTRODUÇÃO

Informações em saúde, especificamente informações epidemiológicas sobre morbidades no território nacional, podem ser estudadas considerando informações disponíveis nos sistemas de saúde. Este trabalho possibilitou o estudo combinado de áreas como: epidemiologia, matemática aplicada e estatística, *Big data* e ciências da computação.

A técnica de Redes Bayesianas surgiu com o objetivo de verificar qual a influência probabilística não direta de uma variável para as demais (Neapolitan, 2004). Cada nó representa uma variável aleatória; as ligações entre os nós são as arestas e representam as dependências probabilísticas entre as variáveis. Assim consegue demonstrar as incertezas na forma de grafos acíclicos e direcionados (*DAG*) como suas dependências probabilísticas entre os nós (Ben-Gal, 2007).

Os Modelos Aditivados Generalizados (GAM) (HASTIE e TIBSHIRANI, 2008), é um modelo linear generalizado com um preditor linear envolvendo a soma das funções suavizadoras (não lineares) das covariáveis, sendo uma abordagem flexível para identificar e descrever relações não-lineares entre preditores e a variável resposta por funções suavizadoras. Podendo ser descrito matematicamente, por:

$$g(\mu) = \beta \cdot X + f_1(X_1) + f_2(X_2) + f_3(X_3) + \ldots + fn(Xn) \quad (1)$$

Neste modelo é possível incorporar a informação espacial, considerando no modelo os pares geo-referenciados de latitude e longitude e a uma função do tempo. Logo para estudar estes métodos combinados o estudo de caso deste trabalho refere-se à dados tuberculose, esta é uma doença milenar.

METODOLOGIA

Os dados de tuberculose para esta pesquisa foram obtidos do SINAN especificamente por município de residência para os estados do Paraná, Rio Grande do Sul e Santa Catarina, nos anos de 2011 à 2017.

As técnicas utilizadas para analisar estes dados foram aplicadas em etapas, onde cada etapa será responsável por uma determinada ação e implicará no resultado da próxima etapa. Desta forma a sequência de manipulações será a seguinte:

a) Etapa 1 - Manipulação das bases de dados provindas de diferentes sistemas como: IBGE com relação a população de cada município nos três estados da região Sul (Paraná, Santa Catarina e Rio Grande do Sul) e arquivo *shapefile* vetorizado para a manipulação do mapa da região Sul do Brasil; Atlas dos indicadores sociais do último censo do ano de 2010; e SINAN com as informações de tuberculose que foram organizados como banco de dados para *input* no programa que realiza as análises;

b) Etapa 2 – Uso das Redes Bayesiana para mostrar a estrutura causal subjacente sobre a Tuberculose e como as variáveis estão relacionadas entre si, utilizando *DAG's*. Foram utilizados quatro algoritmos distintos para estudar a estrutura de dependência condicional, porém foi escolhido o algoritmo Incremental *Association Markov Blanket (IAMB),* por ter refinado melhor as informações com relação aos outros e obteve-se os melhores resultados do ponto de vista epidemiológico.

c) Etapa 3 - Modelos generalizados aditivados com inserção do espaço e tempo, via inserção dessas variáveis, sendo elas: anos, latitude e longitude, respectivamente, no processo de modelagem;

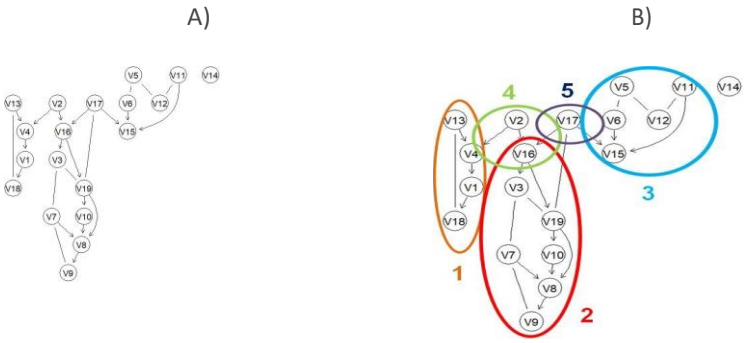
d) Etapa 4 - Modelos Aditivados Generalizados com covariáveis para estudar o padrão espaço temporal da tuberculose ponderados pelos indicadores (covariaveis inseridas no modelo), identificados como primordiais pelas redes bayesianas. Estas etapas foram consolidadas com o apoio computacional do software R.

RESULTADOS E DISCUSSÕES

Abaixo estão descritas as variáveis utilizadas nas redes bayesianas que representam a etapa "b" deste trabalho:

Esperança de vida (V1), Mortalidade Infantil (V2), Razão de dependência (V3), Probabilidade de sobrevivência até os 60 anos (V4), Taxa de envelhecimento (V5), Expectativa de anos de estudo (V7), Índice de Gini (V7), Proporção de pobres (V8), Percentual da renda apropriada pelos 10% mais ricos (V9), Renda per capita (V10), Percentual dos ocupados no setor agropecuário (V11), Percentual de ocupados de 18 anos ou mais que são empregados com carteira (V12), Percentual de ocupados de 18 anos ou mais que são empregados sem carteira (V13), População economicamente ativa de 18 anos ou mais de idade (V14), Subíndice de escolaridade – IDHM Educação (V15), Índice de desenvolvimento humano municipal (IDHM) (V16), IDHM Educação (V17), IDHM Longevidade (V18), IDHM Renda (V19). Na figura 1 é apresentada o *DAG* da rede gerada pelo algoritmo IAMB.

Figura 1- *DAG* da interação entre os indicadores sociais e compostos estudados (Algoritmo IAMB) e seus respectivos agrupamentos interpretados e gerados.



Fonte: Autoria própria (2018).

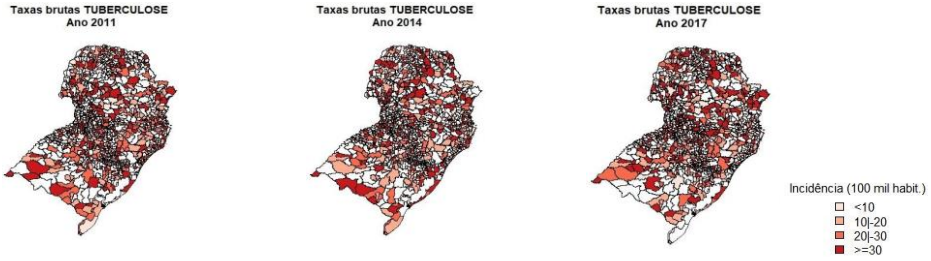
O *DAG* representado no painel A) da figura 1 é relativo ao resultado do *algoritmo IAMB*, onde é possível visualizar as interações entre os indicadores. Pode ser observado no painel B) da figura 1 os subgrupos identificados na rede Bayesiana. Com os subgrupos identificados, pode-se observar que as variáveis V18- IDHM Longevidade, V19 - IDHM Renda, V15 - IDHM Educação, influenciam diretamente nas outras variáveis do seu respectivo agrupamento, por isso só essas variáveis foram selecionadas para ponderação dos modelos. A variável V2 - Mortalidade Infantil está correlacionada com o agrupamento 2 e 1, logo decidiu- se pela inserção dela no modelo, pois ela está correlacionada com longevidade e renda e pode gerar uma nova interpretação dessas informações combinadas.

Interpretando os relacionamentos, IDHM Longevidade (V18) influência na probabilidade com relação a sua esperança de vida. O IDHM Renda (V19) por sua vez, influencia o agrupamento, para a influência probabilística do Índice de Desenvolvimento Humano (IDHM). Já o IDHM Educação (V15) influência na expectativa de anos de estudo do agrupamento.

Na figura 2 é possível visualizar o padrão espaço temporal da distribuição de casos de tuberculose no território da Região Sul, com o auxílio de mapas temáticos, ou seja, as cores mais escuras remetem as maiores taxas de incidência de TB, enquanto que as cores mais claras indicam menores taxas, é uma boa representação, pois permite fácil visualização por meio do gradiente e paleta de cores. Os mapas foram gerados dos anos de 2011 a 2017, porém os números de figuras são limitados neste resumo e então mostra-se os mapas, do ano inicial, do ano

do meio e do último ano do estudo.

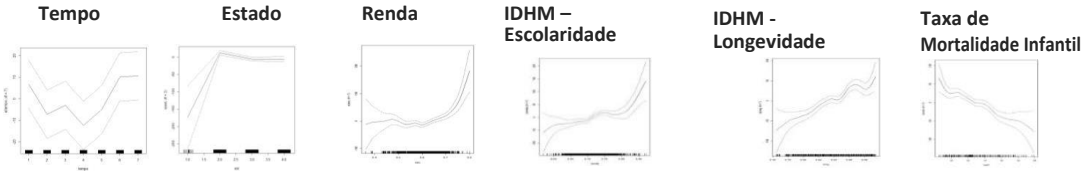
Figura 2 - Mapas temáticos das taxas brutas de TB (casos/100 mil hab.) para a região sul do Brasil nos anos de 2011, 2014 e 2017.



Fonte: Autoria própria (2018).

Na figura 3 estão relacionados os gráficos de tendência para os indicadores estudados, por meio do GAM, da qual foram gerados 5 modelos. A figura 3 mostra o padrão dos indicadores, a evolução ao longo do tempo o comportamento da tuberculose, não está estável sofrendo aumento e declínio ao longo dos anos. Com relação ao comportamento das taxas por estado, o Paraná é o estado com maior volatilidade, já Santa Catarina e Rio Grande do Sul se mantém com valores próximos. Com relação aos indicadores IDHM- Educação, renda e longevidade todos aumentam ao longo dos anos e a taxa de mortalidade teve uma diminuição, com queda significativa.

Figura 3 - Gráficos de tendência para os indicadores estudados, segundo estimativa dos modelos GAM.



Fonte: Autoria própria (2018).

CONCLUSÃO

Informações em saúde, especificamente informações epidemiológicas sobre morbidades no território nacional, podem ser estudadas considerando informações disponíveis nos sistemas de saúde. Este trabalho possibilitou o estudo combinado de áreas como: epidemiologia, matemática aplicada e estatística, *Big data* e ciências da computação. A interação destas áreas levou ao uso inteligente dos resultados e propiciou mostrar o cenário da região sul do país em relação à tuberculose e aos indicadores sociais, que mostram sua situação de saúde atual neste quesito. O uso combinado de análise, possibilitou transformar este grande volume de informação em resultados interpretáveis, possibilitando propor implementação de políticas públicas diversificadas por municípios prioritários na região sul do país.

REFERÊNCIAS

Ben-Gal I. **Bayesian Networks**. In: Ruggeri F, Faltin F, Kenett R. Encyclopedia of Statistics in Quality & Reliability, Wiley & Sons; 2007.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. H. **The elements of statistical learning: data mining, inference, and predicition**. 2º edição. New York, NY: Springer, 2008.

Neapolitan, R. E. **Learning Bayesian Networks**. Upper Saddle River: Pearson, 2004.

AGRADECIMENTOS

Agradeço a Deus, a UTFPR-CP, a minha orientadora Elisangela e a minha família.