Relatório técnico

William Nilson de Amorim

24 de março de 2019

Título da tese: Ciência de dados, poluição do ar e saúde

Instituição de ensino: Instituto de Matemática e Estatística da Universidade de São Paulo (IME-USP)

Departamento: Departamento de Estatística

Orientador: Prof. Dr. Antonio Carlos Pedroso de Lima

Introdução

A Estatística é uma ferramenta imprescindível para a aplicação do método científico, estando presente em todos os campos de pesquisa. As metodologias estatísticas usuais estão bem estabelecidas entre os pesquisadores das mais diversas áreas, sendo que a análise de dados em muitos trabalhos costuma ser feita pelos próprios autores. Nos últimos anos, a área conhecida como Ciência de Dados (*Data Science*) vem exigindo de estatísticos e não-estatísticos habilidades que vão muito além de modelagem, começando na obtenção e estruturação das bases de dados e terminando na divulgação dos resultados. Dentro dela, uma abordagem chamada de aprendizado automático (*machine learning*) reuniu diversas técnicas e estratégias para modelagem preditiva, que, com alguns cuidados, podem ser aplicadas também para inferência. Essas novas visões da Estatística foram pouco absorvidas pela comunidade científica até então, principalmente pela ausência de estatísticos em grande parte dos estudos. Embora pesquisa de base em Probabilidade e Estatística seja importante para o desenvolvimento de novas metodologias, a criação de pontes entre essas disciplinas e suas áreas de aplicação é essencial para o avanço da ciência. O objetivo desta tese foi aproximar a ciência de dados, discutindo metodologias novas e usuais, da área de pesquisa em poluição do ar, que, segundo a Organização Mundial da Saúde, é o maior risco ambiental à saúde humana.

Atividades desenvolvidas

Neste tese, apresentamos diversas estratégias de análise e as aplicaremos em dados reais de poluição do ar. Entre as metodolodias estatísticas usuais, discutimos

- modelos de regressão linear, suas suposições e algumas extensões;
- modelos lineares generalizados, no contexto de dados assimétricos e dados de contagem;
- modelos aditivos generalizados e a utilização de funções não lineares para modelar relações complexas entre variáveis;
- breve discussão sobre modelos de séries temporais e modelos não-supervisionados.

Dentro da abordagem conhecida como aprendizado automático, discutimos:

- modelos de árvores, como árvores de decisão e florestas aleatórias;
- o modelo XGBoost;
- técnicas de validação cruzada e seleção de variáveis;
- regularização;
- técnicas gráficas para a interpretação de modelos "caixa-preta".

Para explificar a aplicação dessas técnicas, consideramos um estudo cujo objetivo era associar a proporção de carros rodando a gasolina com a concentração de ozônio na cidade de São Paulo, e uma extensão desse trabalho, na qual analisamos o efeito do uso de gasolina/etanol na mortalidade de idosos e crianças. As seguintes análises foram realizadas:

- associação das médias diárias de ozônio com a proporção de carros a gasolina;
- associação da mortalidade diária em crianças e idosos com a proporção de carros a gasolina na cidade de São Paulo;
- associação do um índice NO_x/O_3 com a proporção de carros a gasolina na cidade de São Paulo;
- associação da mortalidade diária em crianças e idosos com a concentração de ozônio de óxidos de nitrogênio.
- associação da mortalidade diária por doenças cardio pulmonares em crianças e idosos com a concentração de ozônio de óxidos de nitrogênio.

Em todos os modelos, controlamos o efeito por variáveis de clima, trânsito, tendência e calendário.

Também discutimos estratégias de análise exploratória de dados e apresentamos os principais conceitos envolvidos na raspagem de dados web (*web scraping*).

Resultados obtidos

As análises realizadas indicaram que a associação entre a concentração de ozônio e a proporção estimada de carros a gasolina não é linear. Para valores muito baixos ou muito altos da proporção estimada de carros a gasolina, a relação com a concentração de ozônio não é muito clara. A não-linearidade sugerida pelos modelos reflete bem a complexidade do fenômeno sob estudo.

Os modelos também apontaram que a direção dessa associação é, em geral, negativa, mas a forma variou consideravelmente a depender do modelo escolhido. Também mostramos que os resultados dependem da estação de monitoramento, indicando a importância da química atmosférica local na formação do ozônio ou a impossibilidade da estimativa da proporção de carros a gasolina representar toda a cidade.

Na maioria dos modelos, observamos um maior erro para valores altos da concentração de ozônio. Isso indica que, possivelmente, estamos deixando fora da análise alguma variável que explique os picos desse poluente.

Os modelos ajustados para investigar o efeito da proporção estimada de carros a gasolina e da concentração de ozônio na mortalidade geral em crianças e idosos não indicaram uma relação clara entre as variáveis. Ao considerarmos a concentração de ozônio e de óxidos de nitrogênio como variáveis explicativas, em vez da proporção de carros a gasolina, encontramos associações positivas entre esses preditores e a mortalidade, principalmente quando defasamos as concentrações em um dia. Essa relação também foi observado quando filtramos a mortalidade apenas para doenças cardio pulmonares.

Os resultados nas duas análises mostraram, sobretudo, a dificuldade de se analisar dados de poluição do ar, as diversas estratégias existentes para abordar cada problema e o quanto devemos ser cuidadosos ao interpretar os resultados. Em problemas tão complexos, dificilmente vamos conseguir respostas simples, sendo preferível, em alguns casos, levantar mais hipóteses do que tentar chegar em respostas definitivas.

Considerações finais

Durante a última década, novas tecnologias vêm mudando bastante a forma como estatísticos encaram a análise de dados. A utilização de computadores, instrumentos de medição cada vez mais sofisticados e a internet geram um volume de informação cada vez maior, abrindo novas possibilidades em áreas que historicamente sofriam com a escassez de dados. As preocupações que antes giravam em torno de encontrar o modelo mais adequado para os dados e derivar resultados assintóticos para a construção de intervalos de

confiança e testes de hipóteses, foram em certo grau substituídas pelo desenvolvimento de algoritmos que melhor capturam os padrões escondidos nos dados e que possam ser generalizados para além da amostra. Além disso, a necessidade de extrair dados da internet, manipular bases gigantescas e comunicar resultados de forma atraente passaram a frequentar o radar do estatístico.

Essa nova *forma* de se fazer Estatística chamada Ciência de Dados incorporou diversas ideias interessantes à análise de dados, mas o seu principal requisito é um que sempre fez parte da formação dos estatísticos: a análise crítica. Entender o fenômeno de interesse, o objetivo do estudo, as restrições de cada técnica e ser capaz de avaliar criticamente os resultados é essencial independentemente da abordagem de análise que estamos seguindo.

O aspecto multidisciplinar faz os estudos de poluição do ar um exemplo claro disso. Os principais desafios nessa área envolvem a modelagem de fenômenos naturalmente complexos, e a utilização de modelos muito simples para representá-los pode levar a conclusões superficiais e, por consequência, a criação políticas públicas insuficientes.

Esta tese buscou ser uma ponte entre a Estatística e as outras disciplinas que compõem o estudo da poluição do ar. O compartilhamento de conhecimento de forma acessível é essencial para que haja colaboração entre pesquisadores e para que os trabalhos caminhem na direção certa. Este esforço continuará, pois há muitos tópicos que não foram abordados aqui, como estudos de previsão, estratégias de imputação de dados ou o uso de modelos de redes neurais. Além disso, novas técnicas e modelos vêm surgindo todos os dias, demandando que a construção dessa ponte seja um processo contínuo, e não parado no tempo.

Ao passo que a ciência cria novas tecnologias, a tecnologia também muda o jeito como fazemos ciência. A Estatística não é diferente.