`weeee`: uma *wee little* análise do dia que "virou noite"

André Machado André Silva Ianca Leite Mariana Costa Vítor Ramos

Dados

Origem. Se tratam de séries temporais de variáveis meteorológicas coletadas pelo INMET (Instituto Nacional de Meteorologia), em um grande número de estações no Brasil. Para este projeto, os dados foram obtidos via *query* do BDMEP (Banco de Dados Meteorológicos para Ensino e Pesquisa) através da biblioteca `inmetr`, em R.
Dataset. O *dataset* principal possui cerca de 14 milhões de linhas, mais de 15 variáveis numéricas observadas – as mais antigas datando dos anos 70, além de metadados georeferenciados.
Desafio. Um dos desafios encontrados foi manuseio dos dados (problema de *big data*) para o fim de realizar as operações necessárias para análise (problema também de computação numérica de alta performance).

Análise

Contexto. Para nortear a análise, foi utilizada informação externa de que este ano houve atividade atípica de queimadas no Brasil. Pontuamos o evento anômalo em que o dia "virou noite" em São Paulo no dia 19/08/2019 por conta de fumaça proveniente de queimadas.
Objetivo. Com este norte, realizamos uma investigação exploratória dos dados com um viés de visualização geográfica. A investigação possuiu o objetivo de aferir esta anomalia sob a ótica das variáveis meteorológicas observadas nos dados. No total, foram realizadas três análises (listadas abaixo), correspondendo aos 3 slides de conteúdo apresentado possível.

1. Temperatura do ar. De início, realizamos uma visualização da temperatura do ar em agosto de 2019, em relação ao *baseline* histórico para cada dia e para cada estação. O *baseline* foi obtido calculando a média (regressão polinomial de ordem zero, sem regularização), para cada variável, dia, e estação. A maneira simples de obter o *baseline* se deu pela grande (1) irregularidade dos dados e (2) volume de dados. O resultado é um mapa de contorno cuja terceira dimensão é a cor – um escalar que representa a diferença da temperatura observada para o *baseline* (dado em ° C). Notamos que em São Paulo houve um período frio no mês de agosto.
2. PCA. O passo incremental mais intuitivo é olhar para todas as variáveis meteorológicas do dia 19/08/2019. Notamos que haviam 15 dimensões numéricas. Logo, lançamos mão da análise de componentes principais (PCA) para observar a variação da primeira componente principal (PC). Obtivemos a base para o PCA considerando todos os dados históricos dos dias 19 de agosto. Notamos através de um *scree plot* que a variância explicada pela primeira PC foi de 42%, indicando alta covariância entre as variáveis. Transformamos as observações de 19/08/2019 para esta base. Intuímos que uma grande variação da primeira PC indica grande atividade climática. Esta análise foi visualizada geograficamente. Vimos que houve grande variação na região sudeste.
3. *Forecast* da temperatura máxima. Por fim, realizamos um *forecast* utilizando um modelo ARIMA com ajuste automático da *trends* por dia do ano e da semana de uma estação do INMET em São Paulo. A ferramenta que usamos, específica para grande volume de dados, também estima o ajuste de *trend*. Incluímos também limites do intervalo de confiança de 80% do *forecast*. Todos estes parâmetros estão presentes na apresentação. Concluímos que a estação analisada teve temperaturas distantes do modelo obtido, demonstrando que houve, de fato, um evento anômalo no período analisado.

Ferramentas

A análise somente foi possível através da utilização de poderosas ferramentas de computação numérica. Utilizamos R para extrair os dados e Python para realizar as demais computações. Listamos abaixo as principais ferramentas que foram utilizadas. O projeto está disponível no GitHub em `vitorsr/ccd`.

Google Colab. Ferramenta do Google *Research* para pesquisa e educação em aprendizado de máquina. Se trata de um servidor de Jupyter Notebook em uma máquina virtual (VM) Linux. Foi utilizado para escrever e executar os *notebooks* necessários para o trabalho, e fazer uso de ferramentas Linux em uma plataforma de alto poder computacional.
`inmetr`. Pacote de R para realizar *query* do BDMEP. Tivemos que minerar os dados de 2019 para prosseguir com nossas análises.
`pandas`. Biblioteca para manipulação de dados tabulados.
`geoplot`. Biblioteca de visualização geoespacial. Utilizada para desenhar o *outline* das UFs do Brasil.
`scikit-learn`. Ferramentas de mineração de dados e análise. Usada para pré-processamento e decomposição PCA.
`fbprophet`. Ferramenta do Facebook *Core Data Science* para *forecasting*. Utilizado para realizar o *forecast* de dados de temperatura.