Saúde humana e adaptação à mudanças climáticas no Brasil: Uma abordagem utilizando Ciência de Dados (Tratamento de dados)

Lucas Hideki Ueda 156368
Prof. Paula Dornhofer Paro Costa

Novembro 2018

1 Introdução

As mudanças climáticas no planeta estão cada vez mais frequentes e essas alterações podem interferir na saúde, no bem-estar e no desfecho de doenças em seres humanos.

Utilizando a abordagem da ciência de dados é possível analisar a influência e o impacto dessas mudanças na saúde da população.

O projeto visa estudar base de dados de clima e de informações médicas da população de Campinas-SP afim de analisar esses efeitos. A primeira abordagem será o tratamento dessas bases e limpeza dessas bases.

2 Ciência de dados

A ciência de dados é uma área multi-disciplinar que utiliza de conhecimentos matemáticos, estatísticos e de computação combinados ao método científico e algoritmos para extrair insights e conhecimentos a partir de dados.

É uma área crescente tanto no meio acadêmico quanto no meio empresarial e será utilizada para estudar as bases de dados disponibilizadas.

3 Base de dados

As bases de dados utilizadas e seus respectivos resumos de dados constam abaixo.

3.1 Dados prefeitura

A base da prefeitura possui dados da população de Campinas e algumas informações clínicas. Serão utilizadas para analisar alterações das informações clínicas em eventuais mudanças climáticas.

Resumo da base:

Tabela 1: Resumo da base da prefeitura

	Quantidade
Observações	1.407.452
Variáveis	14
Anos coletados	2008-2016

Volumetria da base:

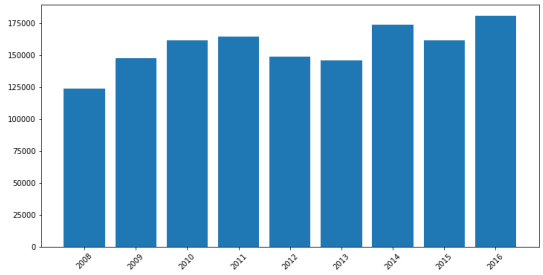

Figura 1: Volumetria ao longo dos anos observados na base da prefeitura

3.2 Dados CEPAGRI

O CEPAGRI é o Centro de Pesquisas Meteorológicas e Climáticas Aplicadas à Agricultura, localizado na Unicamp, o mesmo disponibilizou informações coletadas ao longo dos anos referente ao clima de Campinas. A base disponibilizada consta com informações de: Velocidade do vento, Temperatura média, Umidade relativa do Ar, entre outros dados climatológicos. Será utilizada para enriquecer a base da prefeitura com indícios de mudanças climáticas.

Ao todo 3 bases foram disponibilizadas, diferenciando-se pela periodicidade em que os sensores colhiam as informações. São elas:

- Base 111: Dados coletados a cada 10 hora-minutos
- Base 222: Dados coletados no final do dia (hora-minuto 2400)
- Base 333: Dados coletados em 5 hora-minutos diferentes: 700, 900, 1400, 1500, 2100

Inicialmente todas as análises foram realizadas na base de dados mais granular, ou seja, a Base 111.

Resumo da base:

Tabela 2: Resumo da base do CEPAGRI

	Quantidade
Observações	1.050.445
Variáveis	18
Anos coletados	1997-2017

Volumetria da base:

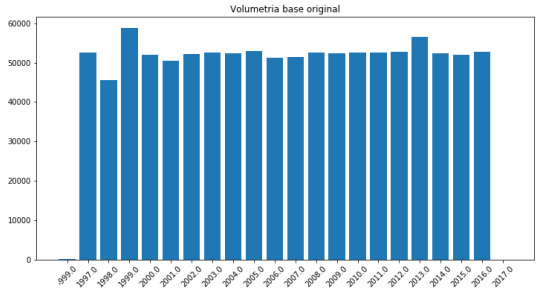

Figura 2: Volumetria ao longo dos anos observados na base do CEPAGRI

É possível notar que a volumetria do ano de 2017 é bem pequena, no entanto ainda sim será considerada visto que são observações mais recentes.

4 Limpeza de dados

A limpeza de dados é uma etapa fundamental na análise dos dados, pois dados com erros de medida, valores nulos ou mesmo valores errados podem inviesar as análises e consequentemente levar a conclusões erradas. O tratamento das bases foi realizado em 3 etapas de limpeza, são elas:

- Limpeza de missings (é considerado missing todo tipo de dado sujo quanto a natureza da variável)
- Limpeza de dados inconsistentes (é considerado dado inconsistente todo dado que não condiz com suas próprias limitações, ex: Dia negativo)
- Limpeza de outliers (em andamento)

4.1 Dados missings

Na base existiam 185.689 observações com missing em pelo menos uma variável, e portanto foram removidas da base. A distribuição de volumetria da base sem missings se encontra na Figura 3.

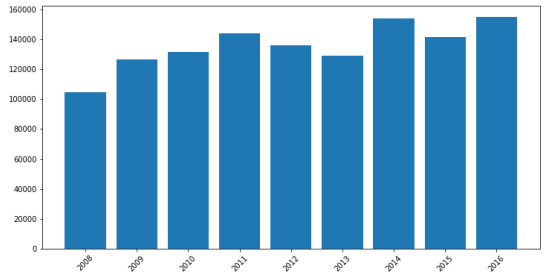

Figura 3: Volumetria ao longo dos anos observados na base da prefeitura sem missings

Já na base do CEPAGRI 2.418 observações eram missings. A distribuição da volumetria da base sem missings se encontra na Figura 4.

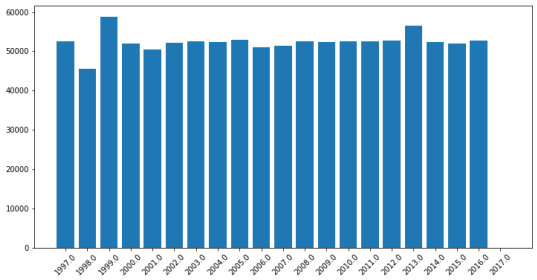

Figura 4: Volumetria ao longo dos anos observados na base do CEPAGRI sem missings

De modo geral, podemos ver pelas Figuras 1 e 3 e Figuras 2 e 4, que o impacto em volumetria não é grande, assim sendo perdemos pouca informação nessa etapa.

4.2 Dados inconsistentes

Alguns valores inconsistentes foram encontrados em ambas as bases.

Na base da prefeitura foi encontrado 3544 idades iguais a 175 anos, dado que a pessoa mais velha do mundo tem em torno de 120 anos, as observações com tal idade foram retiradas da base como dados inconsistentes.

Na base do CEPAGRI diversos valores inconsistentes foram encontrados. Observações absurdas e fora de seus intervalos normais, como temperaturas de 6999 graus, Dias com valores negativos, Ano no valor de 2139, entre outros. É possível notar através das figuras 5 e 6 que outras variáveis também possuem valores inconsistentes.

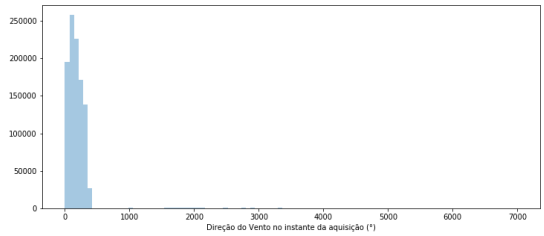

Figura 5: Distribuição da variável "Direção do vento"da base do CEPAGRI. A distribuição concentrada indica que existem valores muito fora do esperado, distorcendo a distribuição da variável.

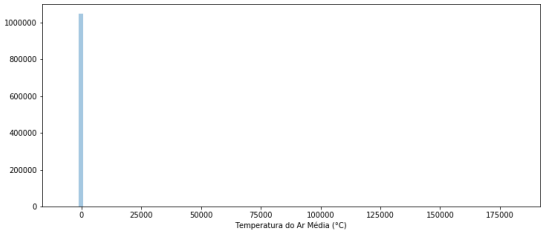

Figura 6: Distribuição da variável "Temperatura do Ar"da base do CEPAGRI. A distribuição concentrada indica que existem valores muito fora do esperado, distorcendo a distribuição da variável.

A limpeza da base nessa etapa se fez a partir de conversas com uma especialista da base que orientou alguns filtros de consistência nas principais variáveis, são elas:

- Direção do vento: Valor entre 0º e 360º
- Umidade relativa do ar: Valor entre 0% e 100%
- Temperatura média: Valor entre 0ºC e 45ºC
- Pressão: Valor entre 822Pa e 1030Pa

A base resultante possui 1.013.879 observações, e sua distribuição ao longo do anos é mostrada na Figura 7.

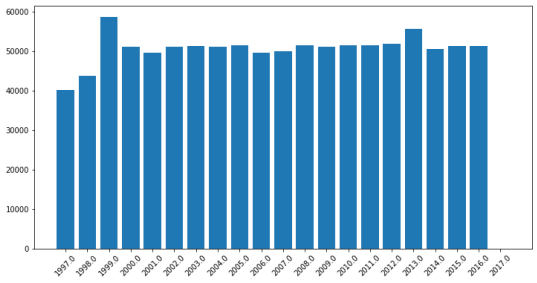

Figura 7: Volumetria da base do CEPAGRI após tratamento por intervalos de valores consistentes

4.3 Outliers

A base da CEPAGRI apresenta valores duvidosos em quase todas as variáveis, como por exemplo nas Figuras 5 e 6, sendo assim afim de analisar as variáveis sem tratá-las manualmente optou-se por uma análise de outliers.

Outliers são observações que fogem muito do usual de uma determinada distribuição, essa mesma pode ser fruto de um erro de medida ou de um evento extremo aleatório.

Uma análise inicial foi feita admitindo-se normalidade na distribuição dos dados. Dessa forma, dois métodos foram aplicados:

- Método do Z-Score
- Método de Tukey

Método do Z-Score

O método consiste em calcular a média (μ) e o desvio padrão (σ) de uma variável. A partir desse valores definimos um intervalo confiança a partir da quantidade de σ que consideramos.

No tratamento em questão foi utilizado 3σ, assim o intervalo de confiança abrange 99.7% dos valores analisados em torno da média μ. O intervalo de valores consistentes é definido como:

$$\mathrm{I} = [\mu - 3\sigma, \mu + 3\sigma]$$

Qualquer valor fora desse intervalo é considerado outlier.

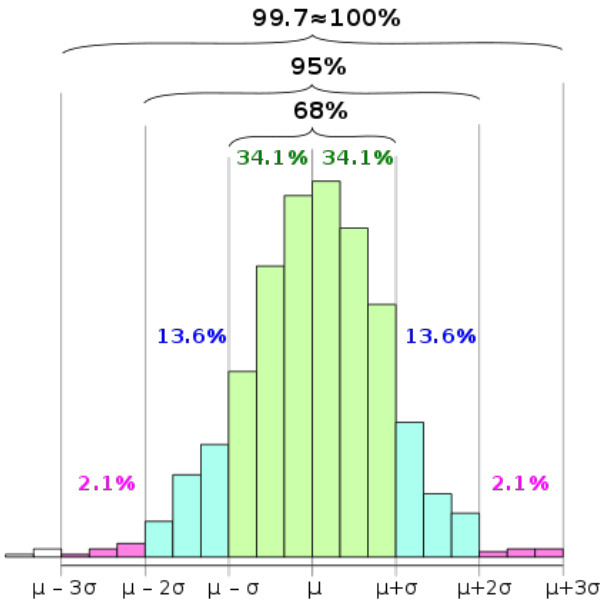

Figura 8: Representação do intervalo utilizado no método Z-Score

O método foi aplicado em todas as variáveis da base para cada ano observado de forma a obter uma assertividade maior à outliers. Toda observação que não estivesse no intervalo foi retirada da base, resultando assim em uma nova base com a seguinte volumetria:

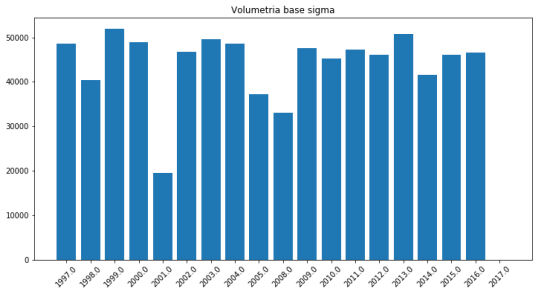

Figura 9: Volumetria ano a ano da base tratada pelo método Z-Score. A base resultando tem 796.144 observações

A base resultante perdeu cerca de 10% de observações.

Em relação as variáveis (como exemplo será tomado as mesmas variáveis observadas nas Figuras 5 e 6: Direção do vento e Temperatura do ar) obtemos os seguintes resultados:

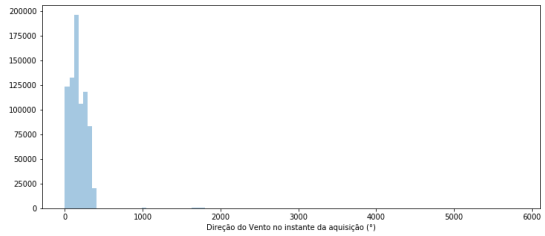

Figura 10: Distribuição da variável "Direção do vento"após tratamento pelo método Z-Score.

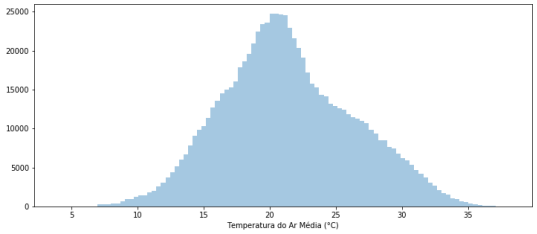

Figura 11: Distribuição da variável "Temperatura do Ar"após tratamento pelo método Z-Score.

Nota-se que o método não é tão assertivo para certas variáveis, como a Direção do vento onde a distribuição permaneceu distorcida e seus valores ficaram fora do intervalo de consistencia dado na seção anterior.

Método de Tukey

O método consiste em calcular os quartis Q1 e Q3 da base e apartir disso calcular o Interquantile Range (IQR), definido como:

$$IQR = Q3 - Q1$$

A partir do IQR definimos o intervalo de valores que adotamos como consistente para o modelo. Esse intervalo é definido como sendo:

$$\text{I} = [\text{Q1} - 1.5 \times \text{IQR} \ , \ \text{Q3} + 1.5 \times \text{IQR}]$$

Assim, qualquer valor fora do intervalo é considerado outlier.

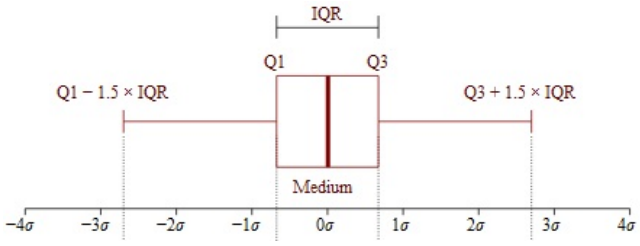

Figura 12: Representação do método de Tukey

De forma análoga ao método anterior, o método de Tukey foi aplicado em todas as variáveis da base para cada ano observado de forma a obter uma assertividade maior à outliers. Toda observação que não estivesse no intervalo foi retirada da base, resultando assim em uma nova base com a seguinte volumetria:

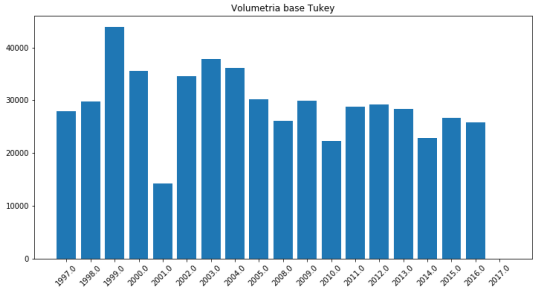

Figura 13: Volumetria ano a ano da base tratada pelo método de Tukey. a base resultante tem 530.323 observações.

A base resultante perdeu cerca de 40% de observações.
Em relação as variáveis obtemos os seguintes resultados:

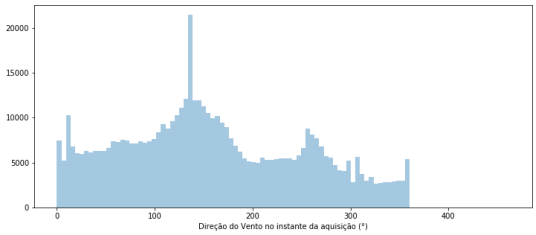

Figura 14: Distribuição da variável "Direção do vento"após tratamento pelo método de Tukey.

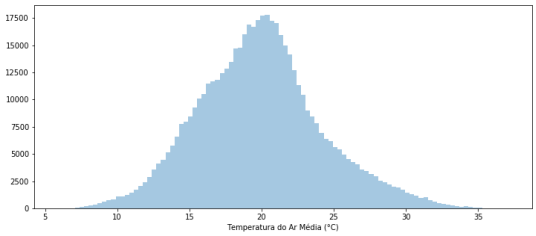

Figura 15: Distribuição da variável 'Tmeperatura do Ar"após tratamento pelo método de Tukey.

Podemos notar que o método é mais assertivo quanto a limpeza de outliers, no entanto o método é custoso quanto à perda de observações.

A premissa desses métodos é a de que a distribuição das variáveis sejam próximas à distribuição normal, podemos presumir um comportamento não normal dado os resultados obtidos.

Resultados e análises completas podem ser encontradas em [1].

5 Conclusão e trabalho futuro

A limpeza de dados é uma etapa extremamente importante no processo de análise de dados, além de ser um trabalho complicado e delicado. Exige-se conhecimento técnico e teórico tanto dos métodos de análise quanto da base de dados em si para limpá-la da forma mais assertiva possível.

Através dos estudos aqui apresentados foi possível entender melhor a base para estruturar a base de forma mais assertiva.

Como trabalho futuro será realizado métodos não paramétricos (como o uso do método DBSCAN) para identificar outliers e análise exploratória dos dados, afim de identificar padrões e comportamentos entre variáveis.

Referências

[1] Códigos do projeto,
`https://github.com/lucashueda/MS777`

[2] Método de Tukey,
`sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704`$_{s}ummarizingdata$
$/bs704_{s}ummarizingdata7.html$

[3] Outliers,
`https://www.ctspedia.org/do/view/CTSpedia/OutLier`