Trabalho Final INF-0611 Análise de Dados

Daniel Noriaki Kurosawa
Eric Uyemura Suda
Fernando Shigeru Wakabayashi

Objetivo

Este trabalho se propõe analisar o conjunto de dados climatológicos CEPAGRI (disponível https://www.ic.unicamp.br/~zanoni/cepagri/cepagri.csv), para o trabalho final da matéria INF-0611 Análise de Dados do curso Mineração de Dados Complexos da Universidade Estadual de Campinas.

Introdução

O conjunto de dados do estudo é disponibilizado em um arquivo CSV sem cabeçalho e é composto por medições espaçadas a cada de 10 min de temperatura (ºC), velocidade do vento (km/h), umidade (%) e sensação térmica (ºC) da Cidade de Campinas, SP.

Foi possível notar falhas nas medições, que foram previamente tratadas de acordo com a metodologia descrita no texto.

Metodologia

Pré-tratamento dos dados

a) Inserção do nome das colunas do arquivo

Devido a ausência de identificadores para as colunas do arquivo, se fez necessária a inserção manual dos mesmos para o tratamento dos dados.

b) Tratamento de linhas sem todos as colunas preenchidas

Devido a um erro de preenchimento dos dados, alguns registros apareciam apenas com a coluna de horário e temperatura preenchidos e uma quebra de linha logo em seguida. O interpretador de csv teve de ser configurado para permitir tal erro e preencher as colunas vazias com NA.

c) Delimitação do intervalo a ser considerado

Para este trabalho, consideramos o intervalo de 01/01/2015 a 31/12/2020, removendo entradas anteriores e posteriores a este. Durante a conversão de horas

para o formato POSIXct, percebeu-se que o R automaticamente contabiliza os horários de de verão, logo deve-se tomar cuidado caso não exista o horário em vigor.

d) Remoção de colunas marcadas como [ERRO] na coluna de temperaturas

Neste conjunto de dados, algumas temperaturas foram registradas com a string [ERRO], o que impossibilita a análise desta coluna sem o devido tratamento. Felizmente, a simples conversão da coluna para o tipo numérico transforma tais valores em valores nulos, que possibilitam a remoção de tais colunas.

Nesta etapa foram removidas 22.081 entradas, o que corresponde a 7,08% dos dados originais no intervalo considerado. O tamanho do conjunto após a remoção é de 289.699 entradas.

e) Análise e remoção de outliers

Após a correção da coluna temperatura, verificamos os valores máximos e mínimos de cada coluna, e quando o valor foi considerado inadequado, foram investigadas entradas anteriores e posteriores à entrada contendo o dado a ser investigado. Tal investigação se deu necessária nas colunas sensação térmica e vento. A Figura 01 apresenta as distribuições e o boxplot das variáveis para análises.

No caso da sensação térmica, foram constatados valores demasiadamente elevados (99.9ºC) e demasiadamente baixos (abaixo de zero graus celsius). Para as sensações elevadas, a simples remoção das linhas com este valor máximo mostrou-se suficiente, sendo o próximo valor máximo encontrado condizente com os seus arredores. Já no caso da sensação térmica mínima, os valores encontrados explicam-se pela alta velocidade do vento e alta umidade, e por isso foram considerados adequados.

No caso dos ventos de elevada intensidade (143,6 km/h), verificamos que os valores dos registros anteriores e posteriores são condizentes e por isso os dados foram considerados adequados. Além disso, uma rápida pesquisa por notícias nesta data ajudaram a corroborar tal análise.
(http://g1.globo.com/sp/campinas-regiao/noticia/2015/12/temporal-tem-ventos-de-ate-143-kmh-na-regiao-de-campinas-diz-cepagri.html)

Para a variável de umidade, verificamos que existiam medições com valores marcados com 0% de umidade, porém após uma pesquisa verificamos que estes valores extremos existem apenas em lugares muito áridos como desertos. Logo consideramos estas medições como outliers e as removemos da análise.
(https://super.abril.com.br/blog/oraculo/como-e-medida-a-umidade-do-ar-ela-alcanca-0-ou-100/)

Além disso, também encontramos alguns clusters com valores de umidade 100% somente nos anos de 2015 , 2016 e 2017. A Tabela 1 ilustra um exemplo de análise de entorno realizada para determinar se realmente os valores de umidade são iguais a 0%.

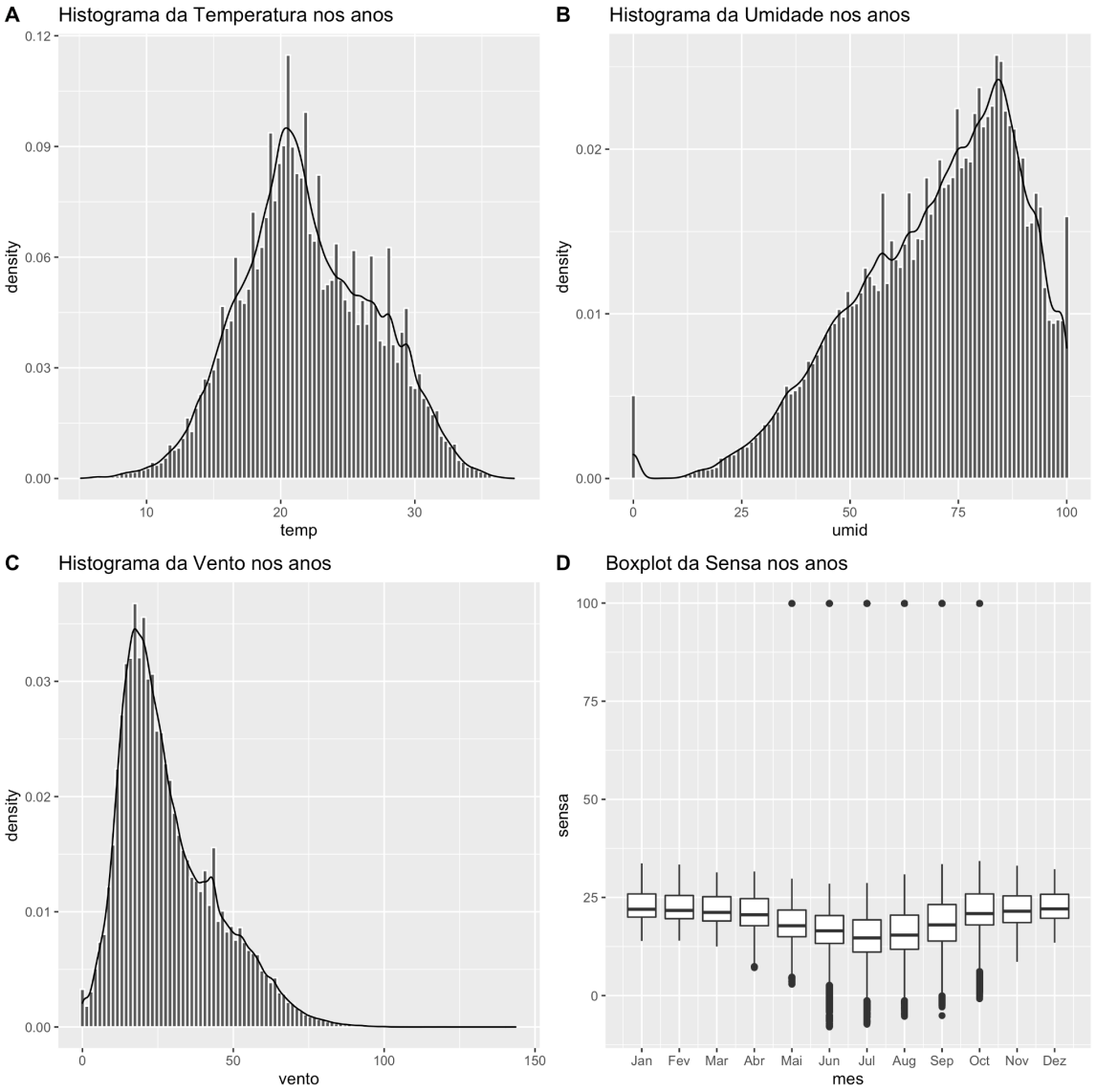

Figura 01 - Visualização das distribuições univariadas para análise da qualidade da informação.

Horário	Temperatura	Vento	Umidade	Sensação
08-02-2020 6:10	19,2	20,4	88,7	17,9
08-02-2020 6:20	19,4	17,2	88,7	18,2
08-02-2020 6:30	19,8	16,8	88,7	18,6
08-02-2020 6:40	20,5	15,6	88,7	19,3
08-02-2020 6:50	21,3	15,9	88,7	20,1
08-02-2020 7:00	21,6	15,2	0	20,5
08-02-2020 7:10	21,5	15,8	0	20,4
08-02-2020 7:20	22,5	18	83,6	21,4

08-02-2020 7:30	22,7	16,8	82,9	21,5
08-02-2020 7:40	22,9	18,1	81,8	21,8
08-02-2020 7:50	23.4	22,4	79,3	22,3
08-02-2020 8:00	23.9	27,3	77,9	22,8

Tabela 01 - Análise de entorno de valores de umidade igual a 0%

f) Remoção de medições com valores constantes durante diversas horas

Primeiramente, verificou-se qual o tamanho do intervalo de tempo em que umas das variáveis medidas foi constante seria adequado à remoção dos dados com alto número de repetições, testando-se valores entre 08 horas e 05 dias. Após as simulações com diferentes intervalos, foi escolhido o período de 1 dia que é o equivalente a 144 medições consecutivas.

Nesta etapa foram removidas 9.141 entradas, o que corresponde a 11,64% dos dados originais no intervalo considerado. O tamanho do conjunto após a remoção é de 275.491 registros, restando o que foi considerado adequado e robusto o suficiente para seguir com as análises.

Etapa de Processamento	**#Registros no Dataframe**	**#Registros Excluídos**	**%Registros excluídos (referente ao passo anterior)**	**%Acumulada de registros excluídos**
Dataframe Original	311.780	0	0	0
Remoção - [ERRO] Sensação Térmica	289.699	22.081	7,08%	7,08%
Remoção - Outliers Umidade	284.765	3.458	1,58%	8,66%
Remoção - Outliers Sensação Térmica = 99.9	284.632	133	0,043%	8,71%
Remoção - Valores repetidos por mais de um dia	275.491	9.141	2,93%	11,64%

Tabela 02 - Etapas de pré processamento dos dados e número de registros removidos.

Análise de Dados

a) Análise da sazonalidade mensal das temperaturas

Observando o heatmap podemos observar no subplot A a distinção entre os meses de inverno, metade do ano, e no verão, começo e fim de ano. Ao analisar a umidade e a sensação térmica podemos observar que possuem um padrão semelhante à temperatura de inverno e verão. Porém o vento é a única variável que não segue um padrão claro de sazonalidade apresentando cores misturadas de vermelho e azul.

Logo supomos que existe uma possível correlação entre as séries de temperatura, umidade e sensação térmica.

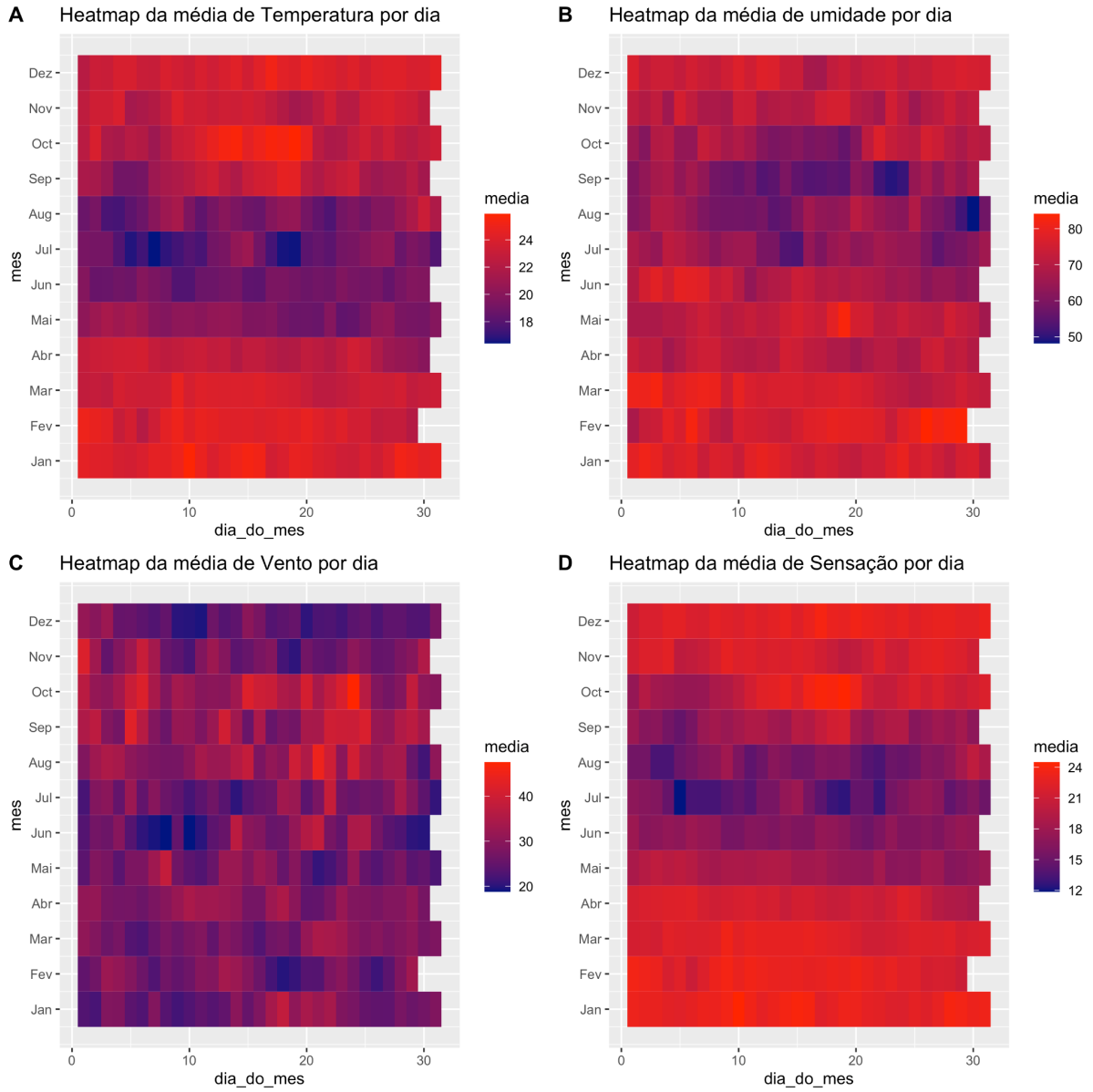

Figura 02 - Heatmap da média mensal da temperatura entre 2015 e 2020

b) Correlação entre as variáveis

Devido às suspeitas de correlação entre as variáveis surgidas durante o item anterior, decidiu-se investigar as mesmas através de uma matriz de correlações. Calculados os valores de correlação, foi feito o cálculo dos valores *p* (Tabela 03) para confirmar os valores obtidos. No entanto, devido ao grande volume de dados, não foi possível tirar conclusões usando o *valor p*, que se demonstrou próximo de zero (Tabela 04) para todos os valores, tal fato explica-se pelo elevado volume de dados utilizado para o cálculo das mesmas.

Para contornar este problema tiramos a média para cada mês de cada ano e refizemos os cálculos de correlação e de *valor p* (Tabelas 05 e 06). Com isso, foi possível constatar que existem evidências de correlação entre vento/umidade, umidade/sensação e temperatura/sensação. É possível notar também, que para o nosso dataset, a variável vento não possui indícios de correlação com as outras variáveis, o que era esperado pela observação dos heatmaps.

Correlação	**Temperatura**	**Vento**	**Umidade**	**Sensação**
Temperatura	1	-0,1723693	-0,6259322	0,9309181
Vento	-0,1723693	1	0.07757017	-0,23026199
Umidade	-0,6259322	0.07757017	1	-0.,1822822
Sensação	0,9309181	-0,23026199	-0.51822822	1

Tabela 03 - Valores de Correlação entre as variáveis usando todos os dados

p-Value	**Temperatura**	**Vento**	**Umidade**	**Sensação**
Temperatura	0	0	0	0
Vento	0	0	0	0
Umidade	0	0	0	0
Sensação	0	0	0	0

Tabela 04 - Valores de P-Value das correlações entre as variáveis usando todos os dados

Correlação	**Temperatura**	**Vento**	**Umidade**	**Sensação**
Temperatura	1	-0,077483227	0,292259327	0,8724265
Vento	-0,07748323	1	-0,009939052	-0,1328859
Umidade	0,29225933	-0,009939052	1	0,299132
Sensação	0,87242653	-0,132885917	0,299132080	1

Tabela 05 - Valores de Correlação entre as variáveis com a média mensal a cada ano

p-Value	**Temperatura**	**Vento**	**Umidade**	**Sensação**
Temperatura	0	0,5299761	0,01558858	3,367146e-22
Vento	0,5299761	0	0,93588575	0,2800135
Umidade	1.558858e-02	0,9358857	0	0,01321217
Sensação	3,367146e-22	0,2800135	0,01321217	0

Tabela 06 - Valores de P-Value das correlações entre as variáveis com a média mensal a cada ano

c) Análise histórica das médias mensais de temperatura

Observando os gráficos de temperatura por ano, podemos observar uma linha de tendência decrescente no intervalo considerado.Apesar disso, não podemos afirmar que temos uma tendência de queda das temperaturas uma vez que grande parte dos meses quentes de 2020 continham erros. Além disso, o intervalo considerado é pequeno para qualquer afirmação sobre mudanças climáticas, uma vez que estas se manifestam em uma escala de décadas conforme podemos observar na Figura 04.

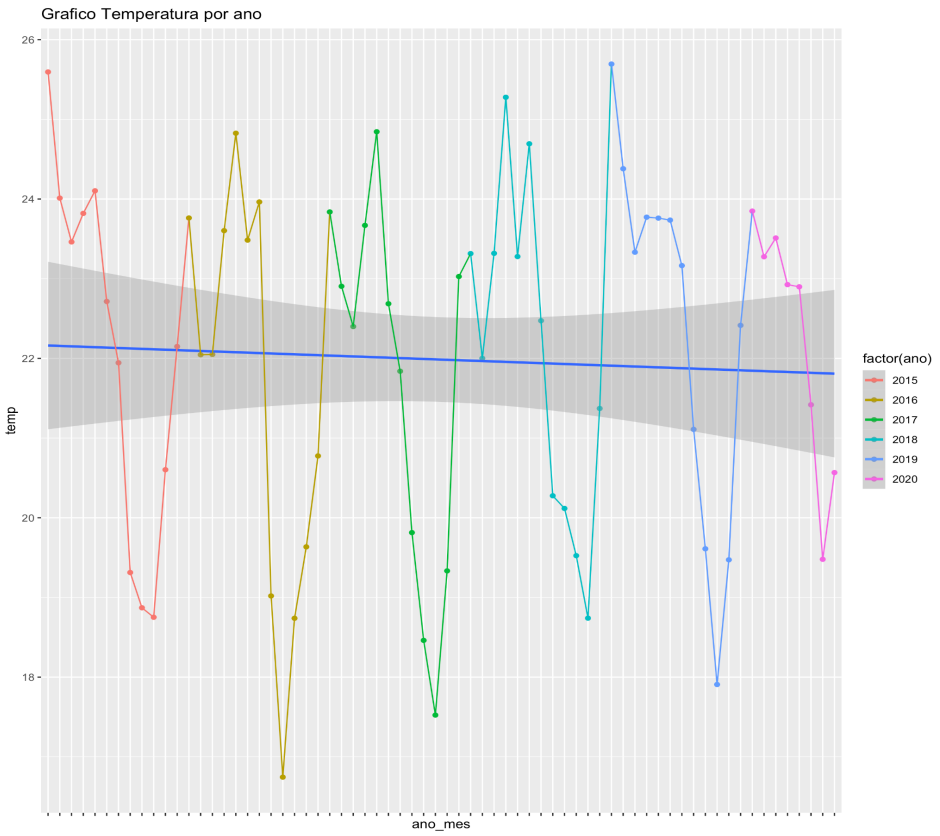

Figura 03 - Série histórica da temperatura de Campinas entre 2015 e 2020 com linha de tendência.

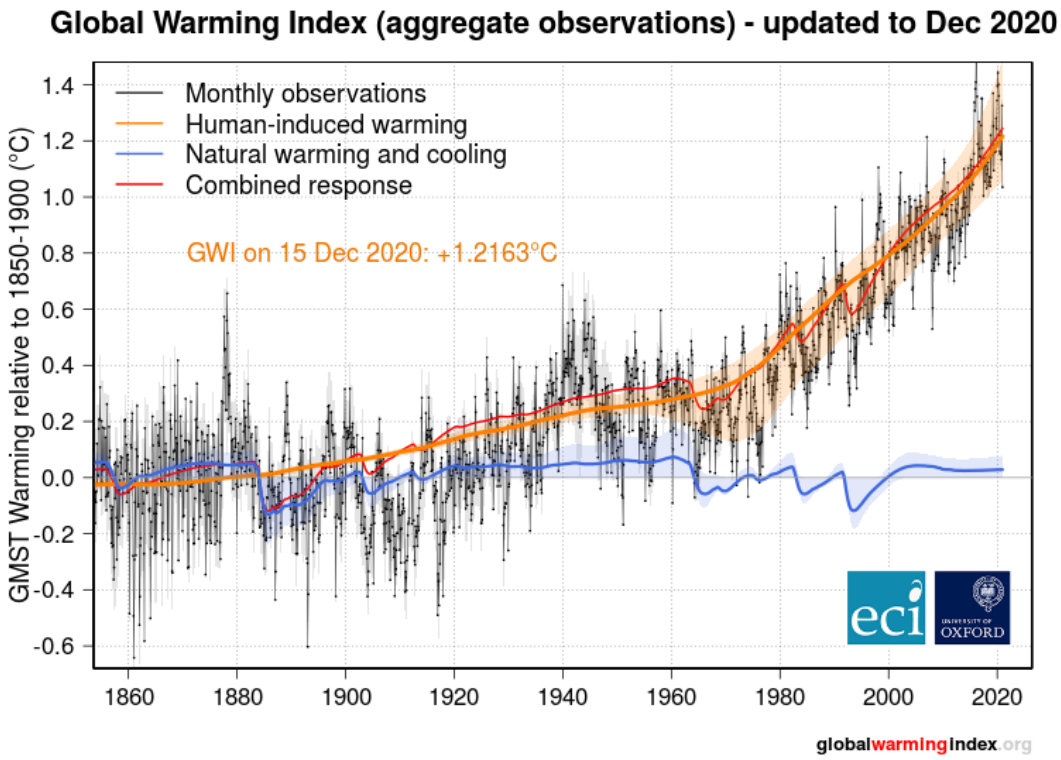

Figura 04 - Série histórica do indicador global de aquecimento.
(https://www.globalwarmingindex.org/)

d) Análise de boxplot da umidade

Observando tanto o gráfico da distribuição dos pontos de Umidade por hora quanto o gráfico de boxplot por ano , podemos observar que os anos de 2015 e 2016 tem apresentado uma umidade relativa bem maior que os anos seguintes de 2017 a 2018, um ponto observado durante a análise é que nesses anos existem uma concentração muito grande de valores de umidade maiores que 90%.Nessa ocasião pode estar acontecendo um erro devido a calibração do aparelho de medição.

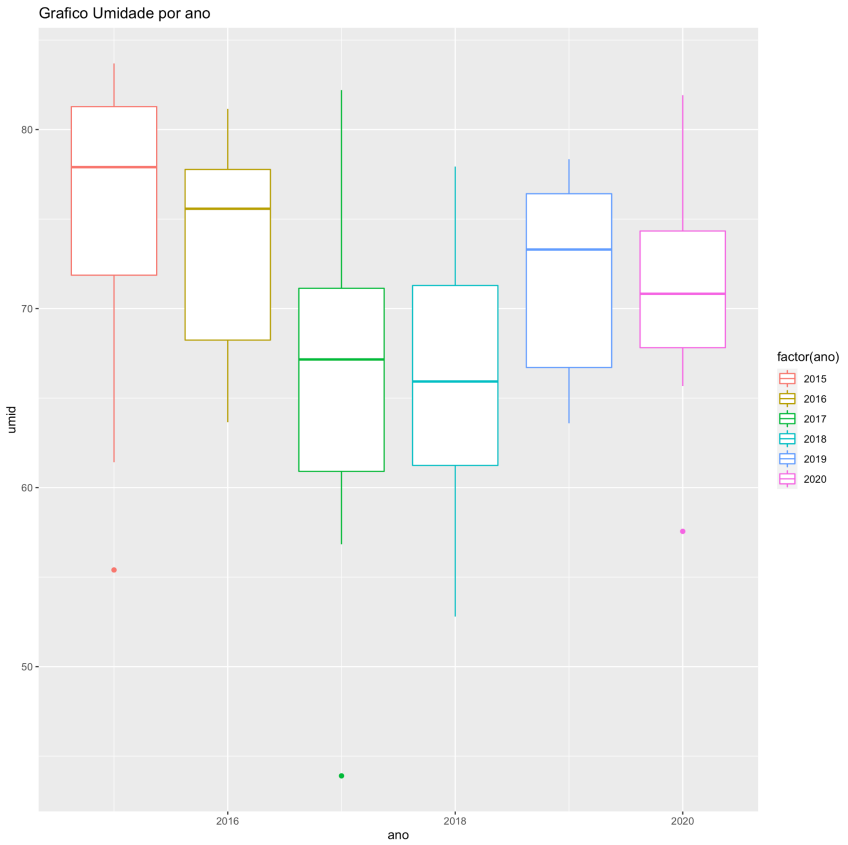

Figura 05 - Boxplot da umidade relativa por ano dentro da janela de 2015 a 2020

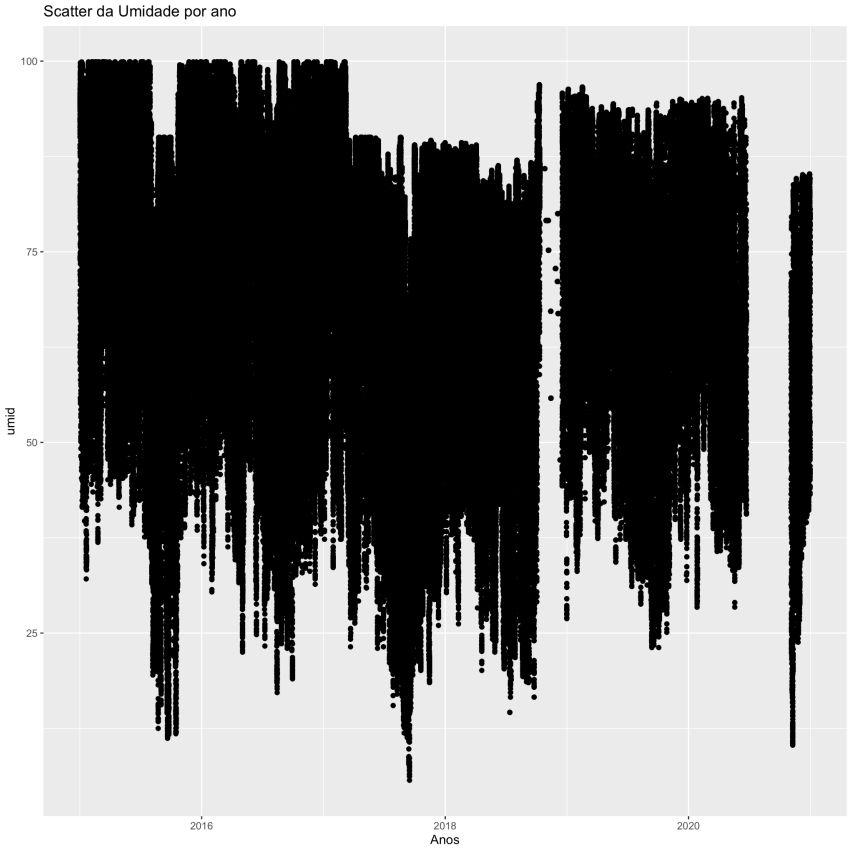

Figura 06 - Scatter plot da umidade relativa por hora dentro da janela de 2015 a 2020.