UNIVERSIDADE ESTADUAL DE CAMPINAS

ESCOLA DE EXTENSÃO DA UNICAMP

INSTITUTO DE COMPUTAÇÃO

CURSO DE APERFEIÇOAMENTO EM MINERAÇÃO DE DADOS COMPLEXOS

ANÁLISE DE DADOS

TRABALHO FINAL

Guilherme Ramos Gouveia
Paola São Thiago da Cunha
Marina Abichabki Pivato

CAMPINAS
2020

INTRODUÇÃO

Como trabalho final da disciplina de Análise de dados do curso de aperfeiçoamento em Mineração de Dados Complexos pela Escola de Extensão da Unicamp (EXTECAMP) foi proposta uma análise dos dados climatológicos da Cidade de Campinas no intervalo de 01/01/2015 a 31/12/2019, utilizando a linguagem R e o R Studio como ferramentas de análise. Os dados foram obtidos através do Centro de Pesquisas Meteorológicas e Climáticas Aplicadas à Agricultura da Unicamp (CEPAGRI/UNICAMP), disponibilizados no endereço **https://www.ic.unicamp.br/~zanoni/cepagri/cepagri**.

Ao longo deste relatório são apresentados os métodos utilizados para o processamento e tratamento dos dados, as dificuldades encontradas com a amostra em questão, assim como a exibição dos resultados da análise exploratória dos dados de forma gráfica.

Todos os trechos de código apresentados neste trabalho fazem referência aos códigos do arquivo **inf0612_trabalho_final.R** do repositório https://github.com/grgouveia/inf0612-trabalho-final no GitHub.

1. TRATANDO OS DADOS

Coletamos os dados abrindo uma conexão com o endereço onde os dados estão armazenados, de acordo com o trecho de código 1.

Devido a existência de anomalias nos dados capturados, foi necessário realizar o tratamento dos dados para que a análise não apresentasse resultados distorcidos e portanto, não prejudicasse a compreensão final.

1.1 COERÇÃO IMPLÍCITA

Observamos que a coluna de temperatura foi definida como *factor* durante a leitura dos dados, pois o parâmetro *StringsAsFactors* tem como valor default **TRUE**. Portanto, efetuamos uma conversão de *fator* para *string* e por conseguinte, para *numeric*, como demonstrado no trecho de código 1.1.

1.2 FORMATAÇÃO DA DATA

Os dados também podem estar com um formato não compatível com a sintaxe da linguagem R, portanto efetuamos uma conversão dessas informações para o formato POSIXct, de acordo com o trecho de código 1.2. Além disso, para simplificar algumas consultas foram criadas colunas em separação de ano, mês e dia.

1.3 OBSERVAÇÕES AUSENTES

Observações ausentes podem surgir desde a perda de informação bem como a falta de resposta durante a coleta. Nos dados coletados, existem diversas informações ausentes, portanto marcamos essas informações com uma constante lógica indicadora de valor ausente (**NA)** durante o tratamento dos dados, para depois removê-los de acordo com o trecho de código 1.3.

1.4 REMOVENDO OUTLIERS

Outro erro comum é a existência de *outliers***,** que são valores que fogem do padrão ou que não fazem sentido para o tipo de dado analisado devido a erros de input, como por exemplo, a sensação térmica máxima de 99.9°C encontrada durante um sumário da coluna de sensação térmica que exibe valores mínimos, máximos, mediana, média, primeiro quartil e terceiro quartil dos dados. Efetuando a mesma análise para a umidade foram encontrados valores de umidade iguais a 0, que não correspondem a valores válidos. Também foi utilizado gráficos simples de boxplot e histograma com os dados ainda contendo outliers. O tratamento dessas anomalias envolve a remoção ou substituição por valores padrões e foram realizados no trecho de código 1.4. Para esses casos, como as outras medidas continham dados validos, colocamos como *NAs*. Assim mesmo após o tratamento acima para remoção dos *NAs*, alguns existirão dos dados de forma proposital para sinalizar esse ajuste.

1.5 DADOS REDUNDANTES

Encontramos em nossas análises, valores repetidos devido a uma interrupção na coleta de dados pelos sensores, entre outros motivos geralmente associados a falhas de sensores. Estes dados foram tratados e as linhas com repetições removidas da base de dados, conforme o trecho de código 1.5.

2. ANÁLISE EXPLORATÓRIA DOS DADOS

A análise exploratória baseia-se em todas as variáveis fornecidas pela base de dados da CEPAGRI. A fim de encontrar relacionamentos entre as variáveis disponíveis foram propostas diversas formas de visualização gráfica dos dados, que serão apresentadas nos tópicos seguintes.

2.1 MEDIDAS DE POSIÇÃO COM A BASE TRATADA

Após serem realizados os tratamentos necessários na base de dados, foi feita a sumarização das variáveis, conforme o trecho de código 2.1.

Variáveis	Mínimo	1º Quartil	Mediana	Média	3º Quartil	Máximo	NA's
temp	5.10	18.60	21.50	21.95	25.50	37.40	0
vento	0.00	17.20	24.90	28.77	38.00	143.60	0
umid	5.70	57.20	73.40	70.31	85.10	100.00	675
sensa	-8.00	16.50	20.00	19.82	24.00	34.30	142

Tabela 1 – Medidas de Posição dos Dados Tratados

Na Tabela 1 acima, observamos que a média e a mediana estão próximas, em todas as variáveis, indicando que a distribuição dos dados está relativamente simétrica e os outliers foram removidos da base de dados.

Para as variáveis que representam a umidade e sensação térmica, foi detectada uma quantidade considerável de valores com algum problema. Estas linhas foram removidas da base de dados, significando que existem períodos que estão com a medição do clima prejudicada.

Existe uma grande variação entre os mínimos e os máximos de todas as variáveis, porém isto é esperado, pois a base registra dados de clima em todas as estações do ano. Existe uma atenção para o mínimo da sensação térmica (- 8.0ºC), indicando que em alguns dias do ano este indicativo do clima ficou fora do esperado para um clima tropical, mas ainda assim, possível de ocorrer. Por este motivo, não foi considerado como um outlier. A mesma interpretação foi dada para os valores máximos e mínimos das variáveis vento e umidade.

2.2 MEDIDAS DE DISPERSÃO COM A BASE TRATADA

No trecho de código 2.2, foi calculado as medidas de dispersão das variáveis. Estas medidas, observam se os valores dos dados estão compactados ou espalhados.

Variáveis	Média	Desvio Padrão	Coeficiente de variação
temp	22	4.96	22.55 %
vento	29	15.84	54.62 %
umid	71	18.77	26.44 %
sensa	20	6.09	30.45 %

Tabela 2 - Medidas de Dispersão dos Dados Tratados

O desvio-padrão na Tabela 2, se mostrou mais disperso na variável umidade. O coeficiente de variação, representa o desvio-padrão expresso como porcentagem da média. Apesar de o vento ter um desvio-padrão menor do que a variável umidade, o coeficiente de variação do vento é

maior em relação à média, indicando uma dispersão maior nos dados da variável vento, do que na umidade do ar. Podemos comprovar, observando os mínimos e os máximos das duas variáveis.

2.3 BOXPLOT E HISTOGRAMA DAS VARIÁVEIS

Os gráficos do boxplot e histograma, trecho de código 2.3, possuem o intuito de fornecer informações sobre a variabilidade dos dados. O conjunto de medidas avaliadas nesses gráficos fornece evidências acerca da posição, dispersão, assimetria e valores atípicos.

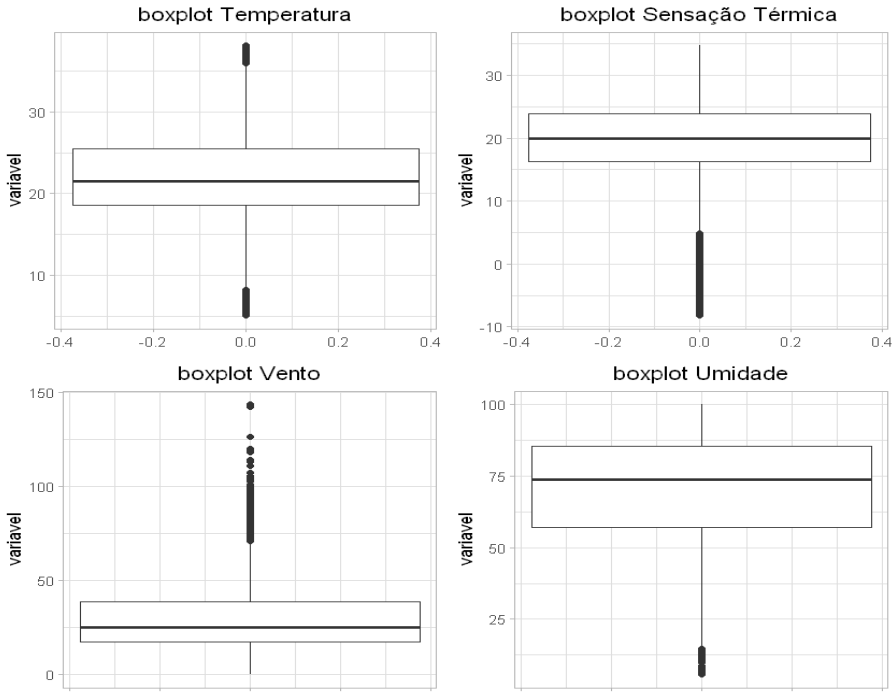


Gráfico 1 - Boxplot para cada variável quantitativa

A temperatura possui sua mediana na posição aproximada 21Cº do gráfico, seus dados estão distribuídos relativamente de maneira simétrica, como é possível observar no histograma de dispersão e nos quartis do boxplot. As outras variáveis sensação térmica e vento, apresentaram uma quantidade considerável de dados fora do limite superior ou inferior, indicando que existe uma assimetria na frequência das observações e uma maior dispersão, suas medianas estão em 20ºC e 25 km/h respectivamente. A umidade também apresentou dados fora do limite inferior, a mediana ficou próxima dos 75 % de e os limites inferior e superior estão entre 73 % e 87 %.

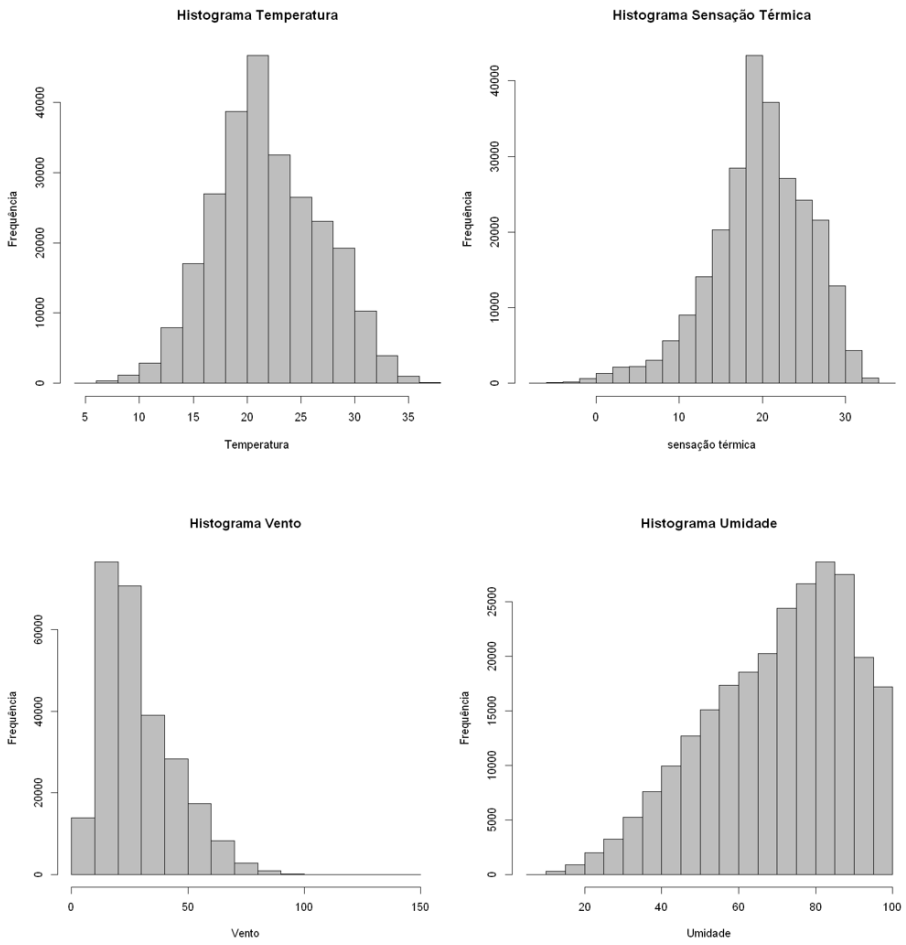


Gráfico 2 - Histograma para cada variável quantitativa

O Histograma apresenta a frequência na qual um evento acontece dentro de um intervalo dado. Os histogramas foram criados com intervalos diferentes, de acordo com a unidade de medida de cada variável. A variável que mais se mostrou assimétrica em sua distribuição foi o vento, tendo o seu pico de frequência na distribuição entre 5 km/h e 10 km/h. A umidade ficou entre 80% e 90%, a sensação térmica entre 18ºC e 20ºC e a temperatura entre 20ºC e 22ºC.

3. ANALISANDO DADOS

3.1 COMPARAÇÃO ENTRE AS MEDIDAS COLETADAS

Para verificar se existe alguma relação simples entre todas as medidas, temperatura, sensação térmica, umidade e velocidade do vento, foi extraído a média por mês das temperaturas, considerando todos as medidas do mesmo mês independente do ano. Os valores Nas das medidas sensação térmica e umidade foram removidos, pois os registros não iriam influenciar na análise. O trecho de código 3.1 corresponde aos gráficos a seguir.

Como não seria medido a frequência dos dados, qualquer valor não encontrado foi removido. Foram encontrados os valores médios conforme a tabela abaixo.

mes	Temperatura	Umidade	Sensacao	Vento	Diferenca_temp_sensa
1	24.33	76.35	23.05	26.29	1.28
2	24.01	76.52	22.74	25.35	1.28
3	23.37	77.09	22.12	26.01	1.26
4	22.64	71.08	21.33	28.51	1.31
5	19.91	73.67	18.27	26.61	1.64
6	18.68	70.54	16.16	26.39	2.52
7	18.39	63.99	14.82	26.90	3.57
8	19.56	61.37	15.84	32.23	3.72
9	21.94	59.76	18.03	33.11	3.91
10	23.67	66.66	21.31	34.57	2.36
11	23.17	73.15	21.88	31.39	1.29
12	24.08	74.10	22.81	27.71	1.27

Tabela 3 - Valores médios por mês das medidas coletadas de todos os anos

É possível ver pelos dados apresentados que a diferença da sensação térmica costuma ser maior nos meses do inverno. Além da tabela, foi utilizado o gráfico das medidas, com valores não normalizados para verificar o comportamento durante os meses dos valores. O fato de não estar normalizado para essa análise não impactará nos resultados pois o foco de comparação é temperatura e sensação térmica, que apresentam a mesma escala de valores. As outras medias estão junto para verificar se nos meses que a diferença térmica foi maior, existe alguma queda ou aumento nos valores das outras medidas.

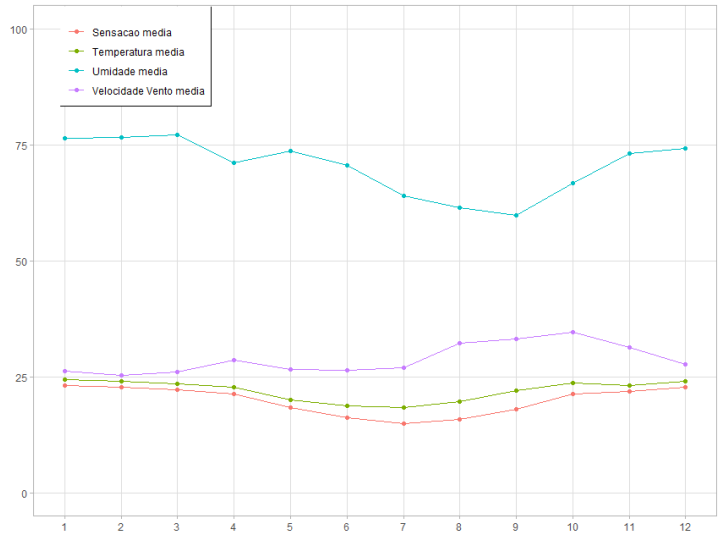


Gráfico 3 - Visualização de todas as variáveis numéricas sem normalização

No gráfico é possível observar que a distância entre as medidas sensação térmica e temperatura é maior nos meses de inverno do que no verão, início de outono e final da primavera. Além disso, quando a umidade começa a diminuir, essa distância começa a aumentar, indicando uma possível relação.

Para verificar a variação das medidas, o gráfico com os valores normalizados foi realizado. No gráfico com os valores normalizados, é possível ver que a temperatura possui uma variação muito mais acentuada que a sensação térmica, apesar das duas seguirem curvas semelhantes nos meses avaliados. A variação da umidade segue uma curva parecida com a temperatura, porém com um deslocamento de meses, indicando alguma relação entre a queda de temperatura e a queda de umidade posterior.

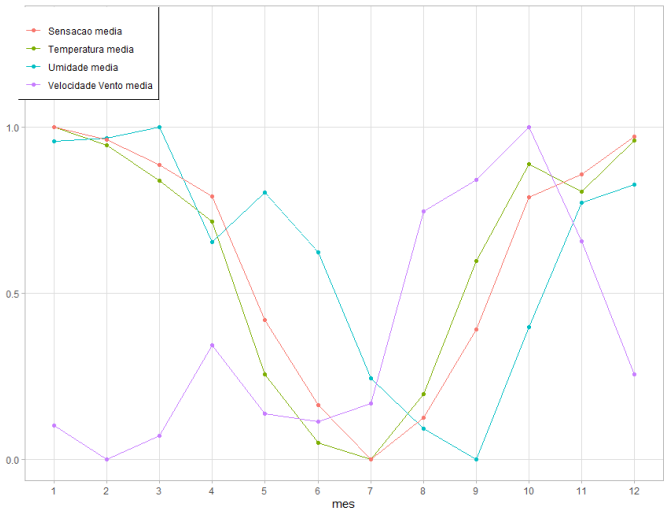


Gráfico 4 - Comparativo dos valores médios normalizados ao mês de todo o período

Porém esse análise é superficial, conta com poucos dados, não foi feito uma avaliação em detalhe do período do dia pois a temperatura e umidade mudam durante o dia e noite e foram usadas apenas médias mensais, sendo uma avaliação que não pode ser utilizada para conclusões referente a verdadeira relação entre temperatura e umidade.

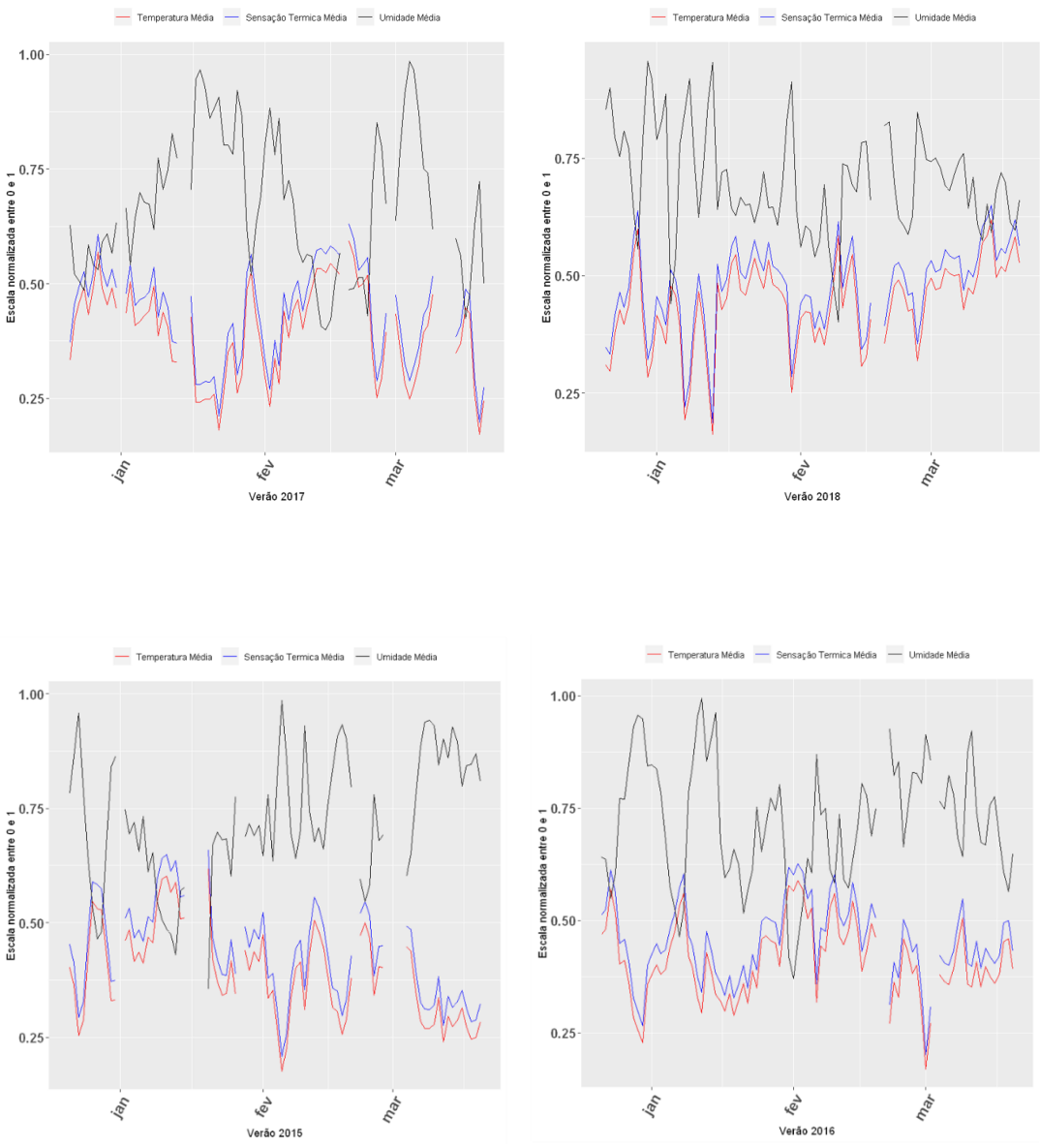
3.2 ANÁLISE DAS ESTAÇÕES VERÃO E INVERNO

O trecho de código 3.2 observa a relação entre as variáveis temperatura, sensação térmica e umidade durante as estações do inverno e do verão de cada ano, foi realizada a média por dia de cada variável. Como a unidade de medida das variáveis não são as mesmas, foi realizada a normalização de cada uma, colocando os valores entre zero e um.

O Gráfico 5 abaixo, mostra uma série temporal nos períodos compreendidos entre verão (21 dezembro – 20 março) e inverno (21 junho - 20 setembro) em cada ano da base de dados.

É possível observar que a umidade influencia de maneira inversa na temperatura e sensação térmica, pois geralmente quando acontece um pico em um dos extremos de umidade a temperatura e sensação térmica vão para o extremo oposto. Por outro lado, a movimentação das variáveis temperatura e sensação térmica em todos os gráficos é muito similar, mas com a temperatura em todo o período ficando ligeiramente mais alta do que a sensação térmica. Uma possível interpretação dos valores altos da umidade do ar durante os verões é que o tempo em Campinas é muito abafado durante esse período, ficando na faixa dos 75% de umidade do ar.

Nos gráficos de cada ano é visível que há valores de medição faltantes na base de dados, identificados com a interrupção da série temporal em alguns períodos.



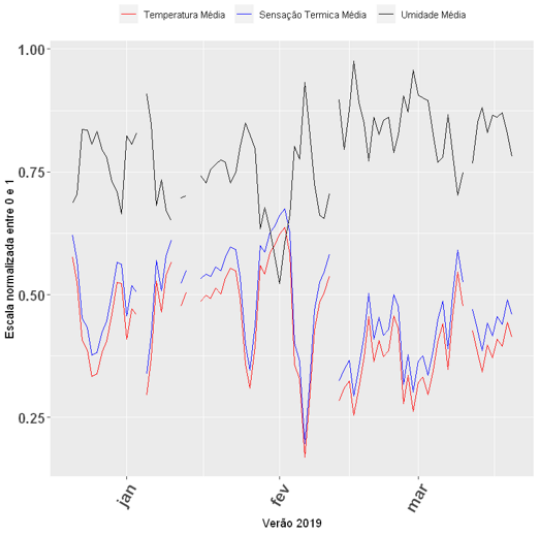


Gráfico 5 – Médias da Temperatura, Sensação Térmica e umidade no verão por ano

Para o Gráfico 6, referente ao inverno, observamos uma variação maior entre as três variáveis. A umidade no período do inverno fica em torno de 50 % e ainda que em todos os anos a umidade assuma valores maiores do que as outras variáveis, nos anos de 2016 e 2018 a temperatura ficou mais próxima da umidade do que a sensação térmica. Uma possibilidade para isso ocorrer é que a variável sensação térmica pode estar sofrendo ação dos ventos no inverno. É importante mencionar, que na base de dados analisada, existem alguns valores extremo que não foram considerados como outliers, mas que não ocorrem com frequência nas estações. Esses picos que ficam fora da média podem estar afetando o comportamento das variáveis ao longo dos anos.

A sensação térmica no período do inverno acompanhou a temperatura, com exceção dos anos de 2016 e 2018. Para que fosso possível investigar de maneira mais profunda os motivos pelos quais ocorreu essa discrepância, seriam necessárias mais informações sobre o clima que não abrangem esta pesquisa.

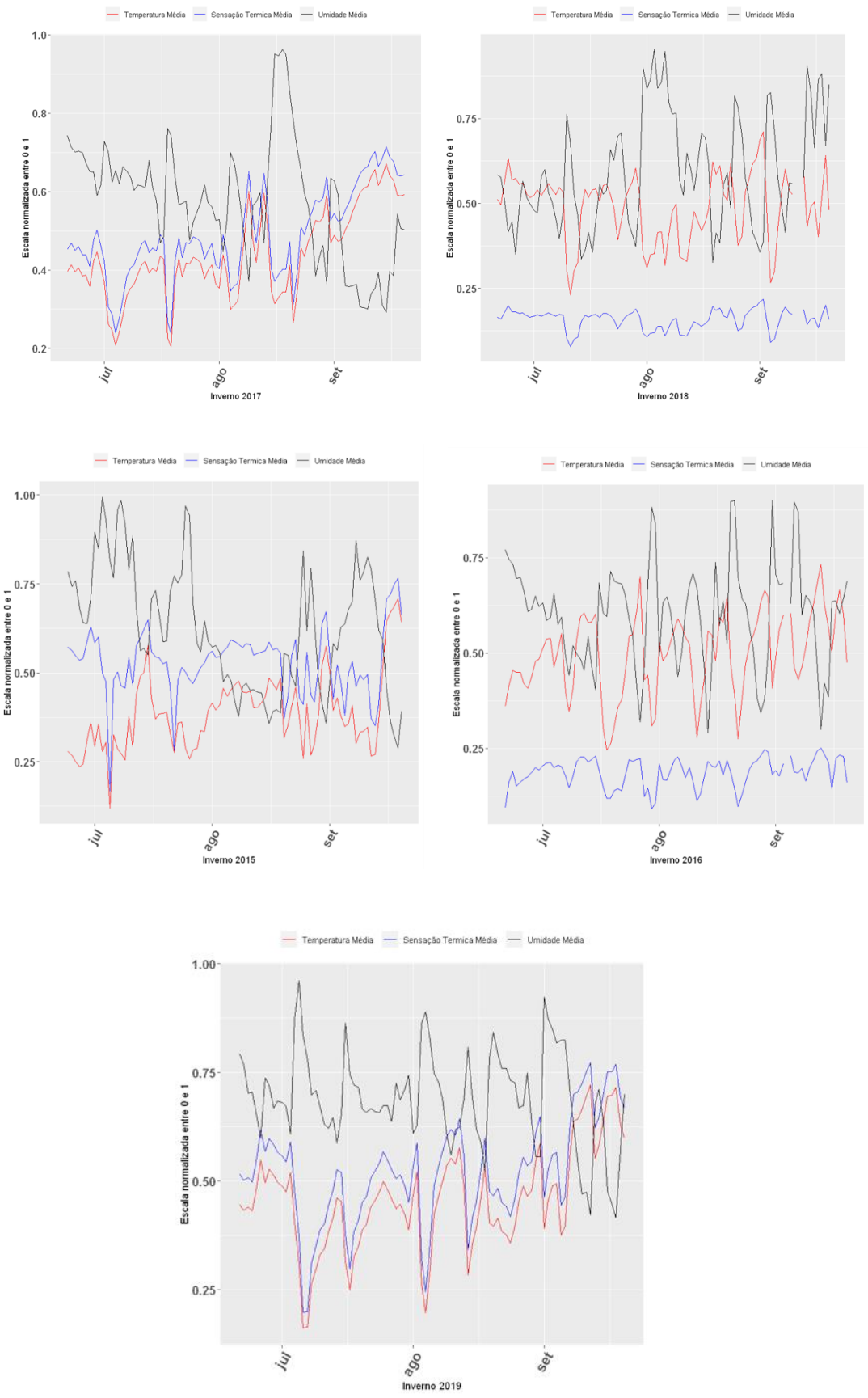


Gráfico 6 - Médias da Temperatura, Sensação Térmica e umidade no inverno por ano

3.3 COMPARAÇÃO ENTRE AS MÉDIAS DA VELOCIDADE E DO VENTO POR PERÍODO DO DIA AO LONGO DOS ANOS

Com a intenção de observar a evolução da velocidade do vento ao longo do dia durante todo o período analisado de 2014 a 2020, o dia foi subdividido entre os períodos manhã, tarde, noite e madrugada.

Após o agrupamento dos dados de acordo com o período do dia, foi calculada a média das velocidades do vento em cada período do dia considerando todos os dias do ano para cada ano subsequente, conforme exibido na Tabela 4 abaixo.

Note que os anos nos dois extremos do conjunto de dados, 2014 e 2020, foram desconsiderados na análise por conter dados desbalanceados, devido ao início e fim da coleta ocorrerem durante o ano corrente.

Ano	Manhã	Tarde	Noite	Madrugada
2015	24.9	27.7	30.9	26.8
2016	26.8	28.8	33.4	29.5
2017	26.0	27.0	31.9	28.7
2018	24.0	25.3	29.7	26.7
2019	27.4	30.7	35.3	30.2

Tabela 4 - Média das velocidades do vento em km/h para cada período do dia e cada ano subsequente

A partir dos dados da Tabela 4, é fácil observar o comportamento crescente da velocidade do vento ao longo do dia em cada período e uma queda na velocidade durante a madrugada para completar o ciclo levando a uma manhã mais calma.

O trecho de código 3.3 mostra o Gráfico 5 de **barras agrupadas** exibido abaixo e facilita a visualização deste comportamento cíclico, através de uma comparação de tamanho das barras do conjunto de dados analisado , já que a análise possui uma variável numérica (velocidade do vento [km/h]), e duas variáveis categóricas (ano e período do dia).

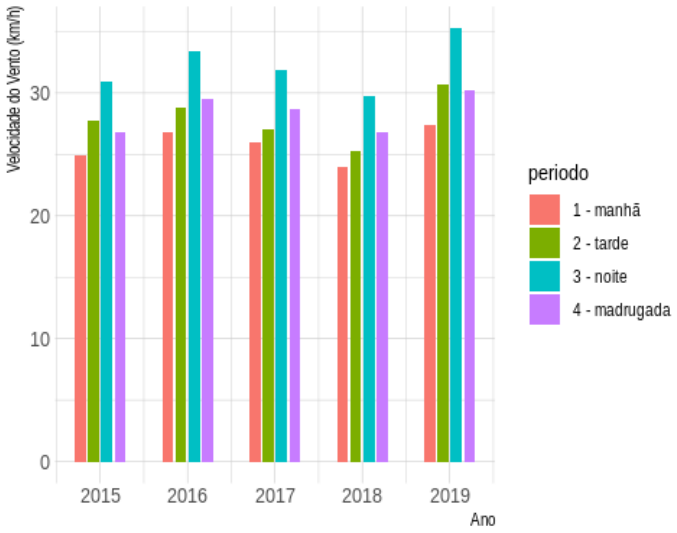


Gráfico 7 - Velocidade do vento por período do dia

Notamos um comportamento bastante padronizado do vento ao longo do dia e dos anos, pois o gráfico mostra de forma explícita e clara como a velocidade do vento cresce ao longo dia e diminui durante a madrugada.

Além disso, é interessante ressaltar que em 2016, Campinas foi acometida de um fenômeno climático anormal com ventos chegando a velocidades altíssimas, o que pode explicar a alta velocidade média do vento neste ano para todos os períodos

3.5 COMPARAÇÃO VENTO X TEMPERATURA DURANTE AS ESTAÇÕES DO VERÃO E DO INVERNO

Para essa análise, no trecho de código 3.5, foram utilizadas as medidas de vento e temperatura. A temperatura foi classifica em frio, normal, calor e muito calor. Frio foi considerado as temperaturas abaixo de 19 graus, normal entre 19 e 27 graus, calor entre 27 e 31 graus e muito calor acima de 31 graus. O gráfico escolhido foi o de barras, que não dá muita informação e apenas é possível comparar a quantidade dos registros se frequente ou não. Os valores Nas das medidas sensação térmica e umidade foram removidos, pois os registros não iriam influenciar na análise.

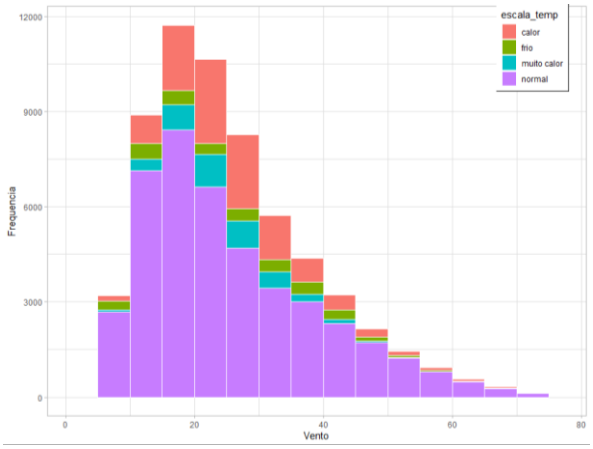


Gráfico 8 - Vento e Temperatura durante o verão

Na estação verão, foi possível verificar que como aumento da velocidade do vento, a classificação de temperatura calor e muito calor apareceram em menor frequência que as demais.

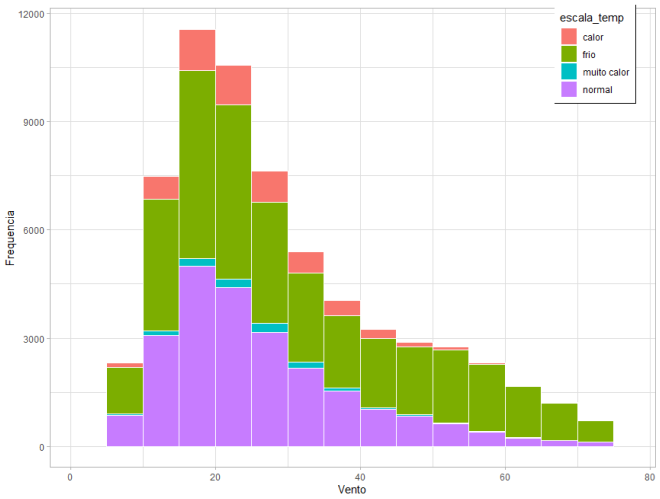


Gráfico 9 - Vento e Temperatura durante o Inverno

No inverno, já é esperado uma frequência baixa de temperaturas calor e muito calor, porém também sofrem influência nos dias de muito vento apresentou um valor muito pequeno ou nenhuma ocorrência de muito calor e calor. Contudo, novamente o número de dados analisados é pequeno e de uma região apenas, não sendo possível concluir que existe essa relação direta da classificação da temperatura com a quantidade de ventos.

Estacao	Temperatura	Vento	Frequencia
Inverno	calor	forte	874
Inverno	calor	fraco	1295
Inverno	calor	medio	3344
Inverno	frio	forte	12986
Inverno	frio	fraco	7278
Inverno	frio	medio	14423
Inverno	muito calor	forte	235
Inverno	muito calor	fraco	240
Inverno	muito calor	medio	817
Inverno	normal	forte	4750
Inverno	normal	fraco	6258
Inverno	normal	medio	13094
Verao	calor	forte	1564
Verao	calor	fraco	1854
Verao	calor	medio	7957
Verao	frio	forte	923
Verao	frio	fraco	1004
Verao	frio	medio	1443
Verao	muito calor	forte	361
Verao	muito calor	fraco	723
Verao	muito calor	medio	2964
Verao	normal	forte	9358
Verao	normal	fraco	14304
Verao	normal	medio	20332

Tabela 5 - Frequência das classificações calor e ventos nas estações inverno e verão

Outra possível visão dessas informações é verificar pela estação a frequência de ocorrência de ventos e muito calor. Abaixo a Tabela 5 mostra que em medições de muito calor, a velocidade do vento é baixa se comparada às outras medições.

3. CONCLUSÃO

Após o tratamento de dado, foi possível realizar algumas análises simples dos dados, mostrando alguma relação entre a sensação térmica, temperatura, umidade em determinadas estações do ano. Também foi possível verificar o comportamento referente a classificação da temperatura em relação ao vento e por fim o comportamento do vento em diferentes períodos do dia. Porém como os dados são restritos a 5 anos, não é possível generalizar os resultados. Além disso, as análises efetuadas são insuficientes, devido às limitações das informações fornecidas pelo banco de dados da CEPAGRI.

As variáveis são influenciadas por fatores que possuem relações de causalidade com outras variáveis que fogem ao nosso escopo de análise, como a evaporação das águas oceânicas, movimentação das massas de ar e a cobertura vegetal da região. Além disso, a presença de muita inconsistência nos dados, assim como períodos de ausência de informação ou repetitividade de dados devido a falhas nos sensores tornam a análise, inconclusiva.