Universidade Estadual de Campinas
Faculdade de Engenharia Elétrica e de Computação

Projeto final de curso
2018/2019

Projeto clima e saúde:

Análise de dados climáticos

Aluno: Lucas Hideki Ueda
RA: 156368
lucashueda@gmail.com
Graduando em Matemática Aplicada e Computacional
Instituto de Matemática, Estatística e Computação (IMECC)

Orientadora: Profa. Dra. Paula Dornhofer Paro Costa
paula@fee.unicamp.br
Departamento de Engenharia de Computação e Automação Industrial (DCA)
Faculdade de Engenharia Elétrica e de Computação (FEEC)

Campinas, Junho 2019

1 Apresentação

O presente relatório visa atender os requisitos da disciplina "MS877 - Projeto Supervisionado" e apresentar um resumo das principais atividades e resultados obtidos ao longo do segundo semestre de 2018 e primeiro semestre de 2019.

As atividades conduzidas durante este período fazem parte do projeto FAPESP 17/20013-0: "Saúde Humana e Mudanças Climáticas: uma abordagem envolvendo Ciência dos Dados".

As seções seguintes contemplam as etapas de descrição do projeto, a origem dos dados analizados, o preprocessamento dos dados, e a análise dos dados, incluindo uma breve discussão sobre os mesmos.

Todo código escrito no contexto deste projeto está compartilhado publicamente na plataforma *Github*, acessível pelo endereço eletrônico: https://github.com/lucashueda/climate_health_project

2 Introdução

Muito se fala de mudanças climáticas nos dias de hoje. Aquecimento global, gases do efeito estufa e derretimento de geleiras polares são termos frequentemente associados ao tema. Durante o século XX a temperatura média da superfície da Terra aumentou em torno de 0.6ºC e aproximadamente dois terços desse aumento ocorreu após o ano de 1975, de acordo com [10]. A atividade humana está diretamente ligada a alterações climáticas, aumentando, por exemplo, a concentração de dióxido de carbono na atmosfera devido à queima de combustiveis fósseis e de florestas, amplificando o efeito estufa.

Por outro lado, as mudanças climáticas afetam a saúde humana de forma direta, como em períodos de estresse térmico (longos períodos com temperaturas muito baixas ou altas) ou acidentes naturais como enchentes e tempestades, e também de forma indireta, afetando a qualidade do ar, da água ou interferindo nos vetores de transmissão de doenças [4]. Além disso diversos setores dependem de condições climáticas favoráveis, como a agricultura, a pecuária e as usinas hidrelétricas, fazendo com que estudos nessa área sejam cada vez mais relevantes e frequentes como mostrado em [1].

Com o intuito de estudar os impactos das mudanças climáticas e, particularmente, dos eventos climáticos extremos, na saúde humana, a Faculdade de Ciências Médicas (FCM), o Centro de Pesquisas Meteorológicas e Climáticas Aplicadas à Agricultura (Cepagri) e a Faculdade de Engenharia Elétrica e Computação (FEEC) da Unicamp, deram origem ao grupo de pesquisas em "Clima e Saúde", alavancado pelo projeto seminal "Human Health and Adaptation to Climate Change in Brazil: A Data Science Approach", financiado pela Fundação de Amparo a Pesquisa de São Paulo (FAPESP). O projeto tem como propósito estudar os tipos de correlações existentes entre dados climáticos regionais (cidade de Campinas) e bancos de dados de variáveis relevantes na área da saúde, na expectativa de desenvolver modelos preditivos que poderão ser potencialmente aplicados na mitigação dos impactos dos eventos climáticos extremos na saúde da população.

Partindo deste contexto, o presente relatório foca na análise de dados climáticos coletados pela estação meteorológica mantida pelo Cepagri, dentro do campus de Campinas da UNICAMP. O principal objetivo do presente trabalho foi a caracterização do microclima desta região da cidade de Campinas (São Paulo), tendo-se como base dados coletados ao longo das últimas duas décadas. Em particular, o presente trabalho buscou evidências de tendências nas variáveis estudadas ao longo dos anos e aprofundou-se na análise de eventos climáticos extremos de ondas de calor, ondas de frio e *humidex* (índice que combina valores de temperatura e umidade e oferece uma métrica de desconforto térmico).

Em particular, são contribuições deste deste trabalho:

- Preprocessamento da base de dados coletados pela estação meteorológica do Cepagri, Unicamp;
- Limpeza dos dados, análise de outliers e remoção de dados medidos erroneamente;
- Definição de intervalos de valores válidos para cada variável estudada;
- Cálculo e análise de métricas de *humidex* e eventos climáticos extremos de ondas de calor e ondas de frio.

3 Metodologia

A metodologia adotada neste trabalho é ilustrada na Figura 1. A figura foi adaptada do trabalho de 5 e ilustra um processo de extração de conhecimento a partir de uma base dados, também referenciado como processo *KDD* (do inglês, Knowledge-discovery in databases).

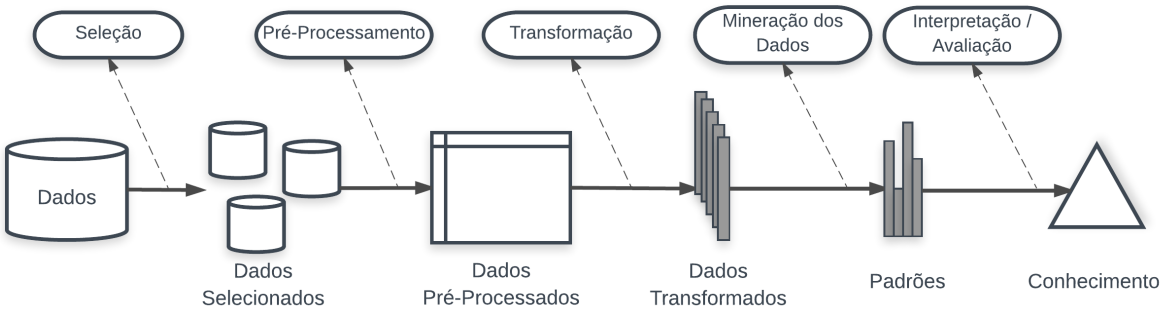


Figure 1: Processo de extração de conhecimento *KDD* (adaptado de [5]).

O processo consiste em passos iterativos que buscam extrair informação de base de dados:

- **Passo 1 (Entendimento do Problema):** Nessa etapa o objetivo é entender primordialmente o motivo que justifica o processo de *KDD*, ou as perguntas de pesquisa que guiarão o processo, do ponto de vista de quem ou o quê utilizará esse conhecimento. No contexto deste trabalho, esta etapa foi cumprida por meio de várias reuniões com o grupo de especialistas vinculados ao projeto.
- **Passo 2 (Criação da base alvo):** Selecionar as bases de dados de interesse, que possuam as variáveis a serem estudadas. Partindo-se da disponibilidade inicial apenas dos dados da estação meteorológica do Cepagri, definiu-se que seria de interesse analisar todo o período disponível de dados entre 1997 e 2018. Uma descrição desta base é apresentada na Seção 3.1.
- **Passo 3 (Limpeza e preprocessamento de dados):** Retirar dados com ruídos ou inconsistentes, tratar valores faltantes e processar a base para uma forma utilizável. No contexto deste trabalho, esta etapa foi realizada de maneira interativa. Estatísticas descritivas inicialmente extraídas dos dados, associadas a gráficos de distribuição de cada variável, foram discutidos e analisados em reuniões periódicas com o grupo de especialistas vinculados ao projeto. A partir destas reuniões, foram definidos intervalos e valores de referência que serviram de base para a limpeza dos dados. Esse processo é descrito na Seção 3.2.
- **Passo 4 (Transformação dos dados):** Encontrar ou transformar, se necessário, informações na sua base que melhor a representa, dependendo do seu objetivo. Para o presente projeto foram selecionadas um subconjunto de variáveis para estudos, estas variáveis são descritas na Seção 3.1. Também foi construído variáveis que melhor caracterizavam eventos climáticos extremos, como descrito na Seção 3.3.
- **Passo 5 (Checagem de objetivo):** Checar se os tratamentos realizados até então estão de acordo com os objetivos do **Passo 1**. Podendo, em caso de negativa, repetir passos anteriores. Para o projeto está etapa consistiu em conversas com especialistas vinculados ao projeto, apresentando resultados parciais e retornando às análises iterativamente.
- **Passo 6 (Análise exploratória):** Definir metodologias de análises e de seleção para aplicar na base, visando o objetivo principal. No contexto desse projeto esta etapa consistiu em analisar hipóteses levantadas pela equipe de especialistas de forma a validá-las quantitativamente atráves de análises gráficas e com estatísticas descritivas, todas descritas na Seção 4.1.
- **Passo 7 (Mineração de dados):** Buscar padrões nos seus dados que possam ser utilizados em um futuro modelo. Buscando estudar padrões e características climáticas nas bases disponibilizadas foram realizadas análises graficas e de correlação nas principais variáveis e nos índices utilizados. Esse processo é descrito na Seção 4.

- **Passo 8 (Entendimento dos padrões):** Entender os padrões obtidos no **Passo 7**, podendo ser necessário realizar passos anteriores novamente. Esse processo ocorreu iterativamente a partir das reuniões periódicas com os especialistas, levantando-se hipoteses a partir de análises descritivas e gráficas e assim gerando resultados apresentados na Seção 4.
- **Passo 9 (Utilização do conhecimento):** A partir dos conhecimentos obtidos nos passos anteriores, utilizar deles diretamente seja implementando um modelo ou elaborando um relatório com os resultados. Todos os resultados estão descritos no presente relatório, assim como no ambiente github, permitindo assim a utilização dos resultados e códigos desenvolvidos no restante do projeto principal.

3.1 Descrição da Base de Dados de Variáveis Climáticas Cepagri

Para o estudo do presente trabalho foram utilizados dados da estação meteorológica do Cepagri, Unicamp. O Cepagri é um centro de estudos climatológicos, criado em 1983, inicialmente criado com a finalidade de desenvolver atividades relacionadas à agricultura. Atualmente, as áreas de pesquisa do Cepagri incluem agrometeorologia, sensoriamento remoto aplicado à agricultura e ecofisiologia, dentre outras.

A estação meteorológica mantida pelo Cepagri foi inicialmente instalada nas dependências da Faculdade de Engenharia Agrícola da Unicamp (FEAGRI), nas coordenadas geográficas latitude sul 22°49'10.14", longitude 47°03'39.14", altitude 634m. A estação meteorológica permaneceu na FEAGRI entre os anos de 1997 e 2003.

Em 14 de maio de 2003, a estação foi transferida para as dependências do Museu Exploratório de Ciências da Unicamp, coordenadas geográficas latitude sul 22°48'54.26", longitude 47°3"27.86", altitude 660m. Sua mudança de localização teve duplo propósito: (1) situá-la num local mais isolado e, portanto, mais independente de influências ambientais próximas que interferissem nas leituras dos sensores da estação; (2) e torná-la um item de visitação do Museu Exploratório de Ciências, contribuindo para a formação de alunos de escolas que visitam o Museu.

Destaca-se que, apesar das medições terem sido obtidas em locais diferentes, os locais de medição estão distantes apenas 586 m entre si. Dessa maneira, foram consideradas desprezíveis as diferenças nas medições devido às diferentes localizações. Assim sendo, foram considerados na análise deste trabalho todos os dados coletados entre os anos de 1997 e 2018, sem distinção.

A estação meteorológica do Cepagri monitora, de maneira automática, as seguintes variáveis:

- Temperatura do ar (°C);
- Temperatura do solo a 3 e 6 cm do solo, e sensor ultravioleta (°C);
- Direção do vento (°);
- Umidade relativa do ar (%);
- Pressão atmosférica (hPA);
- Chuva total (mm);
- Velocidade do vento a 2 e 5 m do solo (m/s);
- Radiação solar ($\frac{kW}{m^2}$);
- Fluxo de calor (positivo e negativo) ($\frac{W}{m^2}$);

As medições dos diferentes sensores da estação são coletadas e armazenadas periodicamente e automaticamente por um *data-logger*, de forma que existem 3 tipos de consolidação dos dados:

- Medições de 10 em 10 minutos ao longo do dia;
- Medições consolidadas ao fim do dia (meia-noite);

- Medições consolidadas às 7:00.

Os dados armazenados pelo *data-logger* são coletados por profissionais do Cepagri e disponibilizados em arquivos ".dat" (formato simplificado geralmente utilizado para dados do tipo texto ou binário).

Cada linha do arquivo ".dat" contém os seguintes identificadores primários:

- Tipo de medição: os valores possíveis são 111 (medição realizada de 10 em 10 minutos), 265 (medições consolidadas realizadas às 7:00h da manhã), 222 (consolidação das medidas diárias concluída à meia-noite).
- Ano da medição;
- Data (dia juliano de 1 a 366);
- Hora no formato 24 horas.

As medições então são agrupadas por ano em um arquivo .dat (formato simplificado geralmente utilizado para dados do tipo texto ou binário), totalizando assim 22 anos de informação.

Todas as bases de dados possuiam informações semelhantes, no entanto optou-se pelo uso da base com medições de 10 em 10 minutos devido a maior granularidade de informação permitindo assim uma consolidação em informações diárias mais flexível. Mesmo assim os tratamentos foram realizados em todas as bases disponíveis.

3.2 Tratamento e preprocessamento dos dados

A fim de consolidar as bases de uma maneira mais apropriada para as análises realizadas na linguagem Python todas elas foram consolidadas pelo tipo de medição, obteve-se assim 3 bases com as informações dos 22 anos em cada:

- Base 111: Base com medições de 10 em 10 minutos ao longo do dia;
- Base 222: Base com informações consolidadas à meia noite;
- Base 265: Base com informações consolidadas às 7:00 da manhã;

O formato utilizado para armazenar as bases foi o *Comma Separated Vector* (CSV) pela sua facilidade de comunicação com a linguagem Python.

3.2.1 Variáveis de controle

Considera-se aqui variáveis de controle aquelas que caracterizam a observação e não são relacionadas a clima, são elas:

- Código da base;
- Ano da observação;
- Dia da observação;
- Horário da observação;

Ao verificar essas variáveis notou-se que existiam informações inconsistentes facilmente detectáveis, como:

- Observações repetidas

- Informação de ano acima de 2019
- Informação de dia acima de 366 e abaixo de 1
- Informação de hora acima de 2400 (esta que indica a última hora do dia)
- Código da base inconsistente

Além dessas inconsistências cerca de 1% da base continha informações faltantes, e assim todos os dados inconsistentes ou faltantes foram removidos. É importante notar que para cada ano o esperado são 144 medições referentes à medidas de 10 em 10 minutos para cada um dos 365 dias do ano totalizando um total de 52560 observações esperadas por ano, no entanto nota-se pela Figura 2 que alguns anos não possuíam o total de observações esperadas.

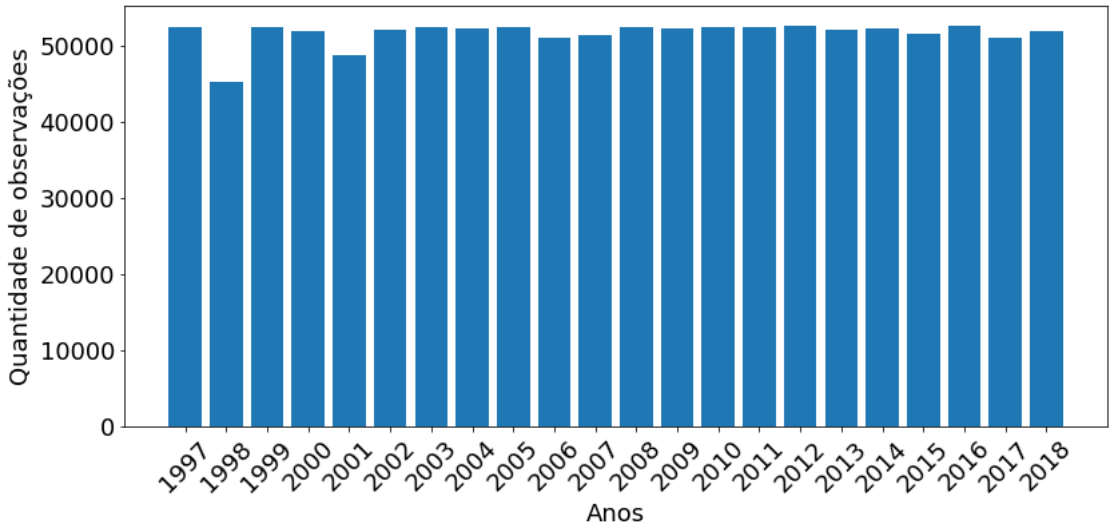


Figure 2: Volumetria da base ano a ano após tratamento de inconsistências nas variáveis de controle..

Isso ocorre devido a falta de medições em determinados dias, a Tabela 1 mostra, por anos, a quantidade de dias com informação faltante.

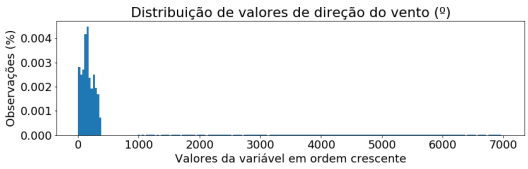
Ano	Quantidade de dias sem observação intervalo
1998	49
2000	2
2001	24
2002	1
2006	9
2007	7
2013	1
2015	4
2017	7
2018	2

Table 1: Valores duvidosos nas variáveis de clima.

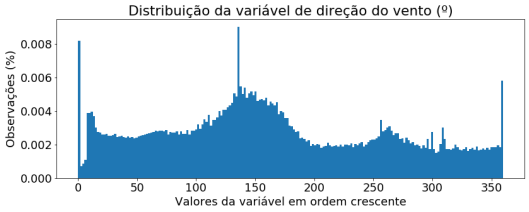
3.2.2 Variáveis de clima

Ao analisar as variáveis de clima foram constatadas distribuições de dados inconsistentes, como mostram as Figuras 3 e 4, onde nota-se que as distribuições de direção do vento e temperatura estão distorcidas por valores *"outliers"* que não representam a realidade física. Dessa forma, com auxilio de especialistas do

Cepagri foi estipulado o intervalo de valores considerados realistas para as principais variáveis da base, como mostra a Tabela 2.

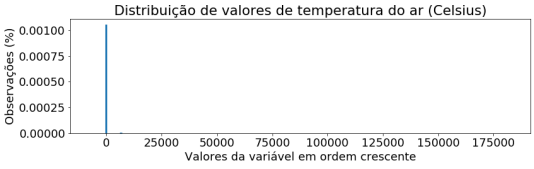


(a) Antes do novo intervalo

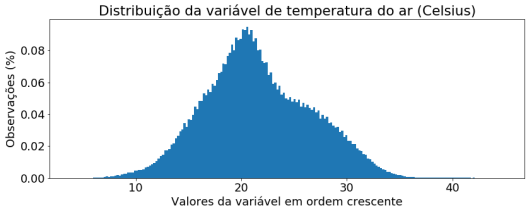


(b) Depois do novo intervalo

Figure 3: Gráfico de distribuição da variável “direção do vento” antes (a) e depois (b) dos novos intervalos.



(a) Antes do novo intervalo.



(b) Depois do novo intervalo

Figure 4: Gráfico de distribuição da variável “temperatura do ar” antes (a) e depois (b) dos novos intervalos.

Variável	Intervalo de valores observado	Novo intervalo
Temperatura do ar	[-6999, 182266]	[-2, 45]
Umidade relativa do ar	[-6371, 6836]	[0, 100]
Pressão atmosférica	[-6999, 2802]	[822, 1030]
Direção do vento	[0, 6999]	[0,360]

Table 2: Valores duvidosos nas variáveis de clima.

Todos os valores fora do novo intervalo foram retirados da base, garantindo-se que as informações da base correspondem a medições coerentes fisicamente. Em torno de 4% da base foi removida após esses tratamentos, sendo o ano de 1997 o mais impactado, como mostra a Figura 5.

3.2.3 Radiação solar

Ao longo do primeiro ciclo de análise exploratória dos dados, observou-se que algumas das variáveis disponibilizadas pela base apresentavam valores que se distanciavam drasticamente da média, de maneira aleatória, muitas vezes saturando num valor máximo, num comportamento típico de ruído de medição.

Tais variáveis não foram consideradas no processo de limpeza e preprocessamento dos dados e optou-se, num primeiro momento, por mantê-las com os seus valores medidos originalmente. Esta estratégia foi adotada pois o grupo de especialistas entendeu, primeiramente, que tais variáveis não eram prioritárias nas análises iniciais desejadas (veja discussão na Seção 3.4) e, adicionalmente, que seria necessário se obter um conhecimento mais aprofundado das grandezas físicas envolvidas e dos sensores responsáveis pelas medições destas grandezas.

Uma das variáveis ignoradas no processo de limpeza e preprocessamento dos dados foi a radiação solar, cujos valores da base vão de 0 a 6999, como mostra o gráfico da Figura 6. No gráfico, a maior concentração de medições no intervalo entre 0 e 1500 kW/m^2 indica a grande tendência de valores de radiação solar ao longo do ano. Nota-se que as medidas são condizentes com o comportamento físico esperado. No meio do ano, período que corresponde ao inverno no hemisfério sul, as medidas de

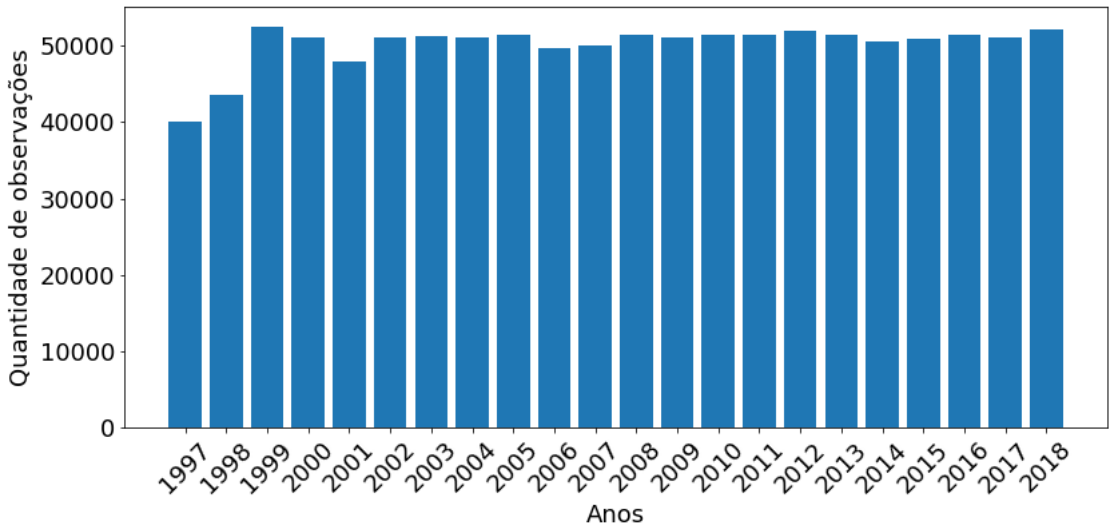


Figure 5: Volumetria da base ano a ano após o novo intervalo de valores.

radiação solar tendem a ser menores que no início ou final de ano, período de verão. No entanto, o gráfico também mostra que cerca de 5% das medições são superiores a 1500 kW/m^2, mas que elas se distribuem de maneira aleatória, sem uma tendência clara, compatível com um comportamento ruidoso oriundo da sensibilidade do sensor.

Adicionalmente, o gráfico da Figura 7 mostra que 92% das medidas ruidosas ocorrem entre 6:00 e 18:00h, ou seja, período diurno, quando a energia solar de fato incide no sensor, é notável ainda que nos períodos de transição entre o período diurno e noturno (as 6:00 e as 18:00), são a região onde se concentra a maior parte dos dados ruidosos. Estas evidências fortalecem a hipótese que os valores medidos fora da faixa 0-1500 kW/m^2, não correspondem a erros sistemáticos do sensor mas são oriundos da natureza e sensibilidade da medida.

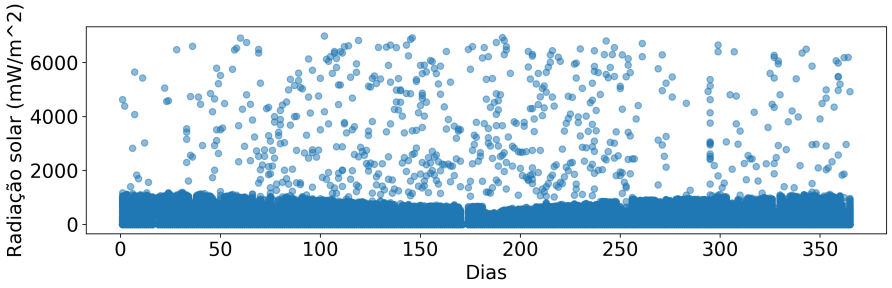


Figure 6: Radiação solar ao longo do ano de 2015. A maior concentração de medições no intervalo de 0 a 2000 indica a grande tendência de valores de radiação solar ao longo do ano. Existe porém uma grande quantidade de medidas fora desse intervalo que são distribuídas de maneira aleatória, indicando um comportamento ruidoso.

É notável pela Figura 7 que os valores de radiação condizem com o esperado e apresenta maiores valores em torno das 12:00h.

3.3 Transformação dos dados e documentação

Uma vez concluído o primeiro ciclo de análise exploratória dos dados, foi possível concluir que, num primeiro momento, seria interessante identificar as variáveis climáticas prioritárias ou mais relevantes a serem analisadas, tendo como referência as perguntas de pesquisa a serem respondidas pelo grupo "Clima e Saúde".

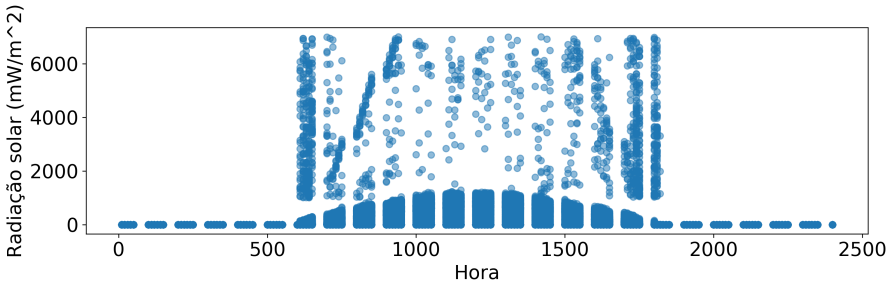


Figure 7: Radiação solar por hora do dia para o ano de 2012.

As reuniões com especialistas explicitaram que as análises iniciais prioritárias estavam relacionadas à identificação de eventos climáticos extremos de temperatura, umidade e pressão atmosférica.

Após o tratamento dos dados descrito na Seção 3.2, uma visão diária da série de dados 111 foi processada de tal forma a se adquirir novas variáveis de interesse. São elas:

- Temperatura do ar mínima e máxima computadas ao longo de 24 horas (°C);
- Pressão atmosférica média, mínima e máxima (hPa) computadas ao longo de 24 horas;
- Umidade relativa do ar mínima, máxima e média (%), onde a umidade média do dia é definida como a média aritmética entre a umidade máxima e a umidade mínima do dia;
- Amplitude térmica (°C), onde definimos a amplitude térmica como sendo a diferença entre a temperatura do ar máxima e mínima para cada dia observado;
- Amplitude da umidade relativa do ar (%), onde definimos similarmente à amplitude térmica como sendo a diferença entre a umidade relativa máxima e mínima para cada dia observado;

Para cada uma das demais variáveis da série 111 também foram geradas novas variáveis de máximo e mínimo valor observado em cada dia, facilitando assim trabalhos futuros, no entanto não foram utilizadas no presente projeto.

Além das novas variáveis, foram calculados índices climáticos utilizados na identificação de eventos extremos, são eles:

- Onda de calor, descrito na Seção 3.4.1;
- Onda de frio, descrito na Seção 3.4.2;
- *Humidex*, descrito na Seção 3.4.3;
- Condição crítica de umidade relativa mínima do ar, descrito na Seção 3.4.4;

Dessa forma, a nova base já preprocessada e com as informações de interesse geradas foi utilizada no presente trabalho para as análises realizadas.

Todo o processo descrito está documentado no ambiente github no link: https://github.com/lucashueda/climate_health_project.

Como proposta para a documentação do metadados das bases utilizadas optou-se pelo uso de arquivos XML. Do inglês *Extensible Markup Language*, XML é uma linguagem de marcação recomendada pela W3C para a criação de documentos com dados organizados hierarquicamente, tais como textos, banco de dados ou desenhos vetoriais. A linguagem XML é classificada como extensível porque permite definir os elementos de marcação. Por esse motivo foi utilizado a linguagem para definir o metadados das bases, os arquivos podem ser encontrados em https://github.com/lucashueda/climate_health_project/tree/master/XML.

3.4 Análises realizadas

As análises dos dados disponibilizados foram realizados visando 2 objetivos: avançar no estudo do clima da cidade de Campinas e estudar fenômenos climáticos extremos na cidade.

Na primeira etapa foram realizadas análises de relação entre variáveis, tendências e variações ao longo dos anos de acordo com o que especialistas descreveram qualitativamente como características climáticas típicas de Campinas. As análises inicialmente realizadas levaram em consideração as seguintes variáveis climáticas:

- Umidade relativa do ar (%);
- Temperaturas mínimas (°C);
- Pressão atmosférica (hPa);
- Amplitude térmica (°C);

A partir dessas variáveis as seguintes relações foram estudadas:

1. Relação entre amplitude térmica e amplitude da umidade relativa do ar;
2. Relação entre amplitudes térmicas e umidade relativa do ar média;
3. Relação entre pressão atmosférica máxima e temperatura do ar mínima;
4. Pressão atmosférica média ao longo dos anos.

A relação entre as variáveis foi medida utilizando-se a estatística de Pearson para amostras, definida como:

$$\rho(x,y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}.\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad (1)$$

Onde x e y são as séries de dados das variáveis sob análise de correlação, $\bar{x}$ e $\bar{y}$ são as médias aritméticas das variáveis x e y, respectivamente, e n é o tamanho da série de dados.

Para a segunda etapa, foram realizados estudos de índices descritos na literatura para detecção de eventos associados diretamente aos dados climáticos.

Para a avaliação de fenômenos extremos de calor foi utilizado o conceito de ondas de calor. Na literatura não existe um consenso na definição de ondas de calor, podendo ser definidas como temperaturas que excedem 35ºC citado em [13] ou, de forma mais flexível, como um período de dias consecutivos com temperaturas acima do percentil 95 [8] [1]. Para o estudo em questão utilizaremos como base o índice CTXP90 (valor de corte a partir do P_{90} de temperaturas máximas em uma janela de 15 dias) e o *"Heat Wave Magnitude Index"* (HWMI) utilizados em [7] e [12], respectivamente, onde uma onda de calor é descrita como sendo um período de 2 ou mais dias consecutivos com a temperatura máxima do dia maior que o percentil 90 (P_{90}) das temperaturas máximas em uma janela de tempo centrada no dia de avaliação, mais detalhes serão dados na Seção 3.4.1.

De maneira análoga, utizaremos o conceito do índice *"Cold Wave Magnitude Index"* (CWMI) apresentado em [6] e [2] onde define-se uma onda de frio como um período de 2 ou mais dias consecutivos com a temperatura mínima do dia menor que o percentil 10 (P_{10}) das temperaturas máximas em uma janela de tempo centrada no dia de avaliação, mais detalhes serão dados na Seção 3.4.2.

Similarmente, foi analisado condições críticas de umidade relativa mínima a partir de uma janela temporal e um percentil de interesse. O processo é descrito na Seção 3.4.4.

Finalmente, analisou-se o *humidex* [11], que relaciona a temperatura do ambiente com a umidade relativa do ar como uma métrica de desconforto térmico que uma pessoa sente. Mais detalhes serão apresentados na Seção 3.4.3.

[1] O percentil k, ou P_k, corresponde ao valor da série de dados onde $k\%$ dos dados tem valores menores que o valor de P_k

3.4.1 Ondas de calor

Com base nos índices CTXP90 e HDWI apresentados, respectivamente, em [7] e [12] , adotou-se como definição de onda de calor um período de 2 ou mais dias consecutivos com temperaturas máximas acima do percentil 90 de temperaturas em uma janela de 30 dias centrada no dia de avaliação calculada no histórico de 22 anos disponíveis. Para a marcação de dias em onda de calor seguiu-se o seguinte algoritmo:

1. Para cada dia d calcula-se o percentil 90 ($P_{90}(d)$) das temperaturas máximas situadas no intervalo de dias $[d-15, d+15]$ para todos os anos da base.
2. Percorre-se todas as observações da base até que se chegue a condição:

   Se a temperatura máxima dos dias d, $d+1$ e $d+2$ forem maiores que $P_{90}(d)$, $P_{90}(d+1)$ e $P_{90}(d+2)$, respectivamente, então em d inicia-se uma onda de calor.
3. Encontrado um início de onda de calor, define-se $i = 0$:

   Enquanto a temperatura máxima em $d+2+i$ for maior que $P_{90}(d+2+i)$ marca-se $d+2+i$ como uma onda de calor e incrementa-se i ($i = i+1$).
4. Os passos 2 e 3 são repetidos até que todas as observações da base estejam marcadas como ondas de calor ou não.

3.4.2 Ondas de frio

De forma análoga às ondas de calor, a definição de ondas de frio têm como base o índice CWMI apresentado em [6]. Adotou-se como definição de onda de frio um período de 2 ou mais dias consecutivos com temperaturas mínimas abaixo do percentil 10 de temperaturas em uma janela de 30 dias centrada no dia de avaliação calculada na janela de 22 anos disponíveis. Para a marcação de dias em onda de frio seguiu-se o seguinte algoritmo:

1. Para cada dia d calcula-se o percentil 10 ($P_{10}(d)$) das temperaturas mínimas situadas no intervalo de dias $[d-15, d+15]$ para todos os anos da base.
2. Percorre-se todas as observações da base até que se chegue a condição:

   Se a temperatura mínima dos dias d, $d+1$ e $d+2$ forem menores que $P_{10}(d)$, $P_{10}(d+1)$ e $P_{10}(d+2)$, respectivamente, então em d inicia-se uma onda de frio.
3. Encontrado um início de onda de frio, define-se $i = 0$:

   Enquanto a temperatura mínima em $d+2+i$ for menor que $P_{10}(d+2+i)$ marca-se $d+2+i$ como uma onda de frio e incrementa-se i ($i = i+1$).
4. Os passos 2 e 3 são repetidos até que todas as observações da base estejam marcadas como ondas de frio ou não.

3.4.3 Humidex

O *humidex* foi proposto em 2016 e é descrito por [11] como uma métrica do desconforto térmico de uma pessoa em uma determinada condição de temperatura e umidade relativa do ar. No mesmo trabalho é descrito uma tabela comparativa entre *humidex* e o desconforto humano, como mostra a Tabela 2.

O *humidex* é definido pela fórmula:

$$H = T + (0.555.[E - 10]), \tag{2}$$

onde T é a temperatura em Graus Celsius e E é a pressão de vapor de água em milibars.

A pressão de vapor é descrita pela fórmula:

$$E = 6.11.e^{5417.7530.(\frac{1}{273.16} - \frac{1}{td+273.16})} \tag{3}$$

Valor de *humidex*	Sensação térmica
Menor que 29	Sem desconforto
De 30 a 39	Desconforto ameno
De 40 a 45	Desconforto, evitar esforço
Acima de 45	Perigo
Acima de 54	Insolação iminente

Table 3: Tabela de *humidex* e desconforto humano.

Na equação (3) o termo td é o ponto de orvalho, que consiste na temperatura de saturação do vapor de água para que haja condensação em unidades de graus Celsius.

Utilizamos uma aproximação para o ponto de orvalho (td) que leva em consideração a temperatura e a umidade relativa, descrita em [9], conforme apresentado na equação (4).

$$td = T - \frac{100 - RH}{5}, \tag{4}$$

onde T é a temperatura e RH é a umidade relativa.

A partir das medidas de temperatura e umidade relativa do ar fornecidas de 10 em 10 minutos pela base 111, as equações (2), (3) e (4) permitiram calcular o *humidex* correspondente. Visando avaliar os impactos dos valores extremos das variáveis climáticas na saúde humana, foi escolhido como valor representativo diário o valor máximo de *humidex* diário.

3.4.4 Umidades mínimas críticas

Segundo a *World Health Organization* (WHO), dias com umidade abaixo de 60% já exigem atenção quanto a saúde humana, e de acordo com [3] umidades abaixo de 30% se tornam críticas, dessa forma entende-se que a metodologia avalia corretamente casos críticos de umidade relativa.

O cálculo desses parâmetros seguiu algoritmo semelhante ao descrito anteriormente:

1. Para cada dia d calcula-se o percentil 10 ($P_{10}(d)$) das umidades mínimas e o percentil 90 ($P_{90}(d)$) das amplitudes térmicas situadas no intervalo de dias $[d-15, d+15]$ para todos os anos da base.
2. Percorre-se todas as observações da base até que se chegue a condição:

   Se a umidade mínima/amplitude térmica dos dias d, $d+1$ e $d+2$ forem menores que $P_{10}(d)$ e $P_{10}(d+1)$ e $P_{10}(d+2)$ (ou $P_{90}(d)$, $P_{90}(d+1)$ e $P_{90}(d+2)$ para amplitudes), respectivamente, então em d inicia-se um período de umidade crítica (ou amplitude crítica).
3. Encontrado um início de período crítico, define-se $i = 0$:

   Enquanto a umidades mínima/amplitude térmica em $d+2+i$ for menor que $P_{10}(d+2+i)/P_{90}(d+2+i)$ marca-se $d+2+i$ como um período crítico e incrementa-se i ($i = i+1$).
4. Os passos 2 e 3 são repetidos até que todas as observações da base estejam marcadas como período crítico ou não.

4 Resultados

4.1 Relações de variáveis

4.1.1 Amplitude térmica x amplitude de umidade relativa

Foi calculado a amplitude térmica para cada uma das duas variáveis e a estatística de correlação de Pearson (ρ), assim permitindo a verificação da correlação entre as variáveis.

Com $\rho = 0.8$, a correlação entre as variáveis se mostra alta, sendo essa relação clara pelas Figuras 10 e 11. Evidencia-se então uma relação direta entre dias com amplitudes térmicas elevadas e amplitudes de umidade relativa também . O resultado tem potencial a servir como um indicador de risco a saúde humana visto que umidades muito baixas são de grande risco para a população. É notável também que na série de dados consta alguns pontos isolados em valores altos de umidade relativa, indicando possíveis "outliers" a serem investigados. (Figura 11).

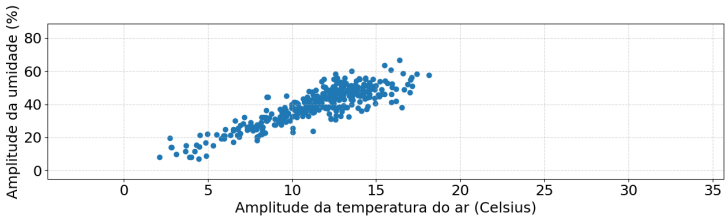


Figure 8: Amplitude térmica e amplitude da umidade relativa do ar para o ano de 2018.

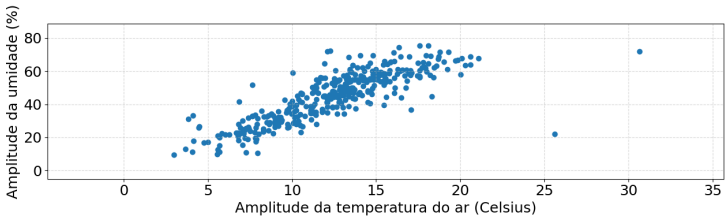


Figure 9: Amplitude térmica e amplitude da umidade relativa do ar para o ano de 2003.

4.1.2 Amplitude térmica x umidade relativa média

A análise anterior evidencia a relação entre amplitudes térmicas altas e amplitudes de umidade relativa também altas, da onde existe um risco em existir nesses dias valores de umidade relativa muito baixos. Analisando as variáveis de amplitude térmica com a umidade relativa média diária consegue-se uma correlação $\rho = -0.65$, indicando uma correlação negativa alta, evidenciada nas Figuras 10 e 11. Uma correlação alta e negativa indica uma relação inversa entre as variáveis, dessa forma destacando-se que dias com amplitudes térmicas maiores tendem a ser dias mais secos, consequentemente sendo dias com mais riscos para a saúde humana.

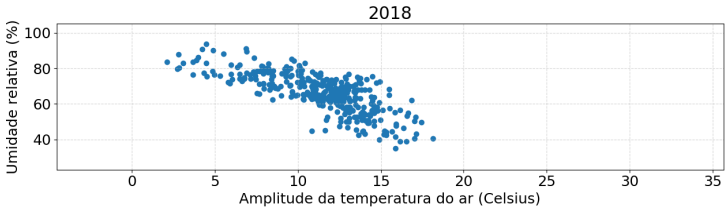


Figure 10: Amplitude térmica e umidade relativa do ar média para o ano de 2018.

Nota-se novamente que há amplitudes térmicas que claramente fogem da maior parte dos valores (Figura 11), indicando a existência de "outliers" a serem investigados.

4.1.3 Pressão atmosférica máxima x temperatura mínima

Foi observado que altas pressões atmosféricas estão relacionadas a baixas temperaturas, como mostram as Figuras 12 e 13, o que de acordo com os especialistas vinculados ao projeto seria o esperado. Também é possível observar que na Figura 12 é evidente que há valores "outlier" de pressão atmosférica, indicando possivelmente que novos intervalos de pressão atmosférica se façam necessários. A correlação para essas

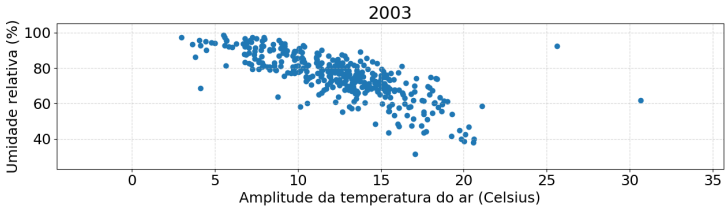


Figure 11: Amplitude térmica e umidade relativa do ar média para o ano de 2003.

variáveis foi $\rho = -0.54$, o que indica uma correlação moderada apesar da clara relação linear evidenciada na Figura 13. A presença de "outliers", notados na Figura 11, influencia o valor da correlação entre as variáveis, fato esse notado ao estimar ρ retirando esses pontos ruidosos e assim aumentando a correlação para o valor de -0.63.

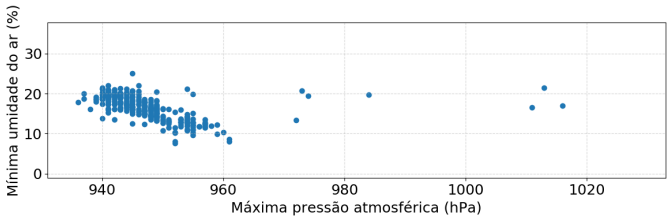


Figure 12: Pressão do ar máxima pela temperatura mínima do ar em 2017.

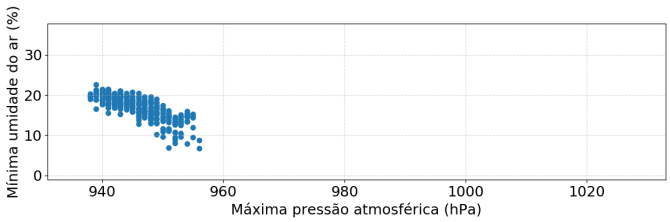


Figure 13: Pressão do ar máxima pela temperatura mínima do ar em 2018.

Observando a Figura 14 nota-se que em anos recentes a distribuição da relação estudada na presente seção está cada vez mais se deslocando para valores menores de pressão atmosférica.

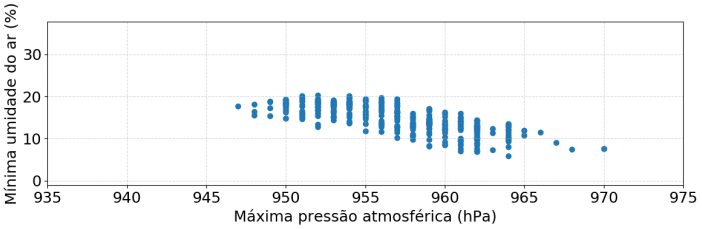
4.1.4 Pressão atmosférica ao longo dos anos

A partir dos indícios observados na seção anterior buscou-se analisar a pressão atmosférica média ao longo dos anos. Observando a Figura 15 nota-se claramente que a pressão atmosférica média nos últimos anos foram as mais baixas do histórico de dados, tendo seu pico em torno dos anos 2000 e desde então tendo valores cada vez menores. Esta evidência indica uma mudança no padrão climático de pressão atmosférica na cidade de Campinas ou um erro de medida do sensor de medição da variável, sendo então tema para trabalhos futuros uma investigação mais aprofundada.

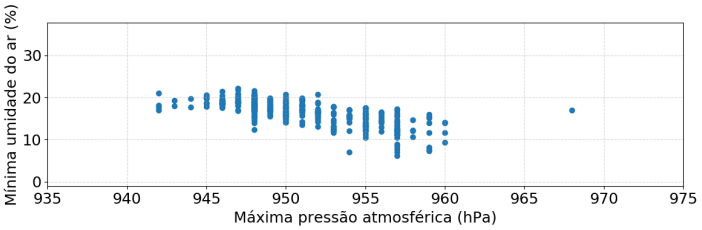
4.2 Ondas de calor

Analisando as ondas de calor ao longo dos anos é possível observar que as ocorrências do fenômeno se concentra nos últimos 10 anos de observação. Nota-se também pela Figura 16 que as ondas de calor mais quentes (em vermelho e preto no gráfico) ocorrem apenas nas estações de verão e primavera, como esperado pois são as estações mais quentes do ano.

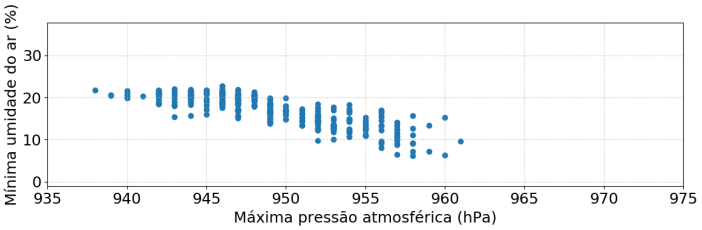
A interpretação do gráfico da Figura 16 é válido para os demais gráficos de estilo semelhante presentes nesse projeto.



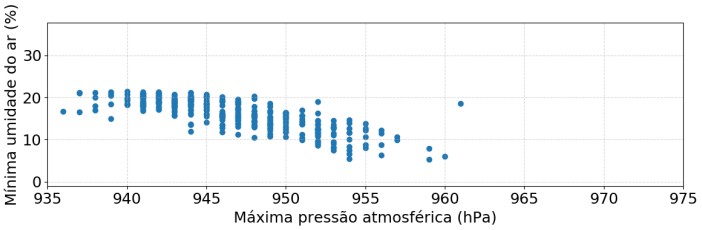
(a) Pressão atmosférica máxima x temperatura mínima em 1999.



(b) Pressão atmosférica máxima x temperatura mínima em 2002.



(c) Pressão atmosférica máxima x temperatura mínima em 2011.



(d) Pressão atmosférica máxima x temperatura mínima em 2016.

Figure 14: Tendência de queda de pressão atmosférica ao longo dos anos.

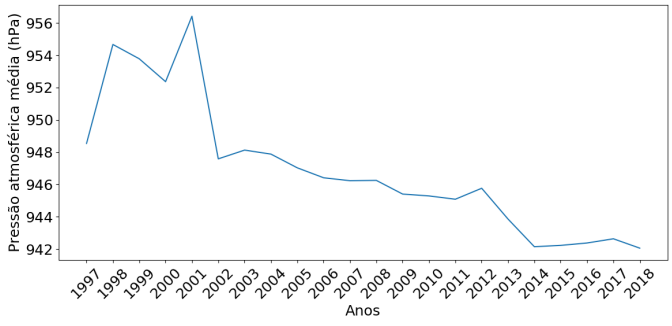


Figure 15: Pressão atmosférica média ao longo dos anos.

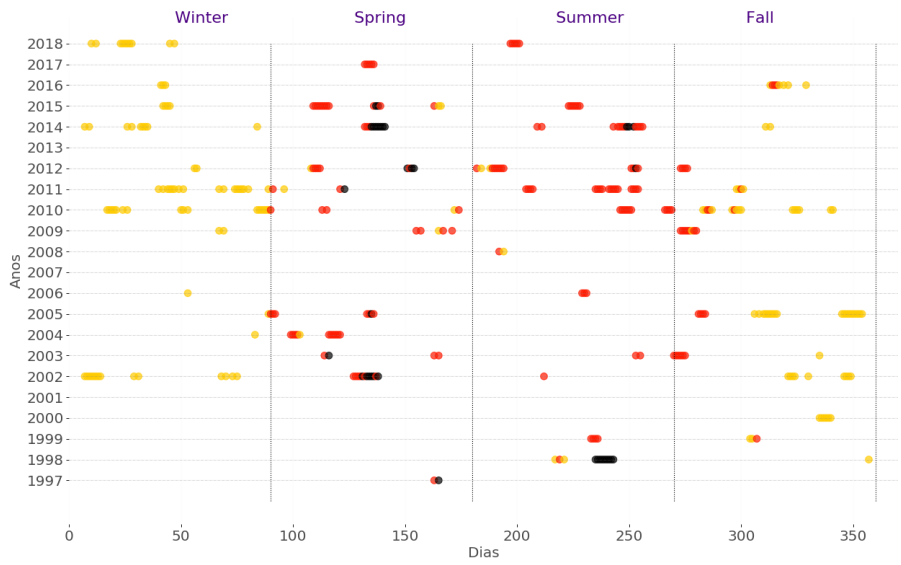


Figure 16: Ondas de calor ao longo dos anos.
Legenda: Amarelo: Ondas de calor abaixo da média de temperatura das ondas de calor; Vermelho: Ondas de calor abaixo do P_{90} de temperatura das ondas de calor; Preto : Ondas de calor acima do P_{90} de temperatura das ondas de calor. Como por exemplo o período da primavera do ano de 2014 onde há alta concentração de dias com temperaturas na faixa de 37°C onde o limitante do P_{90} se encontra na faixa dos 34°C.

Observando a quantidade de dias sob ação de ondas de calor ao longo dos anos (Figura 17) é possível notar que nos últimos anos se concentram as maiores quantidades de dias em onda de calor, tendo seu pico em 2010.

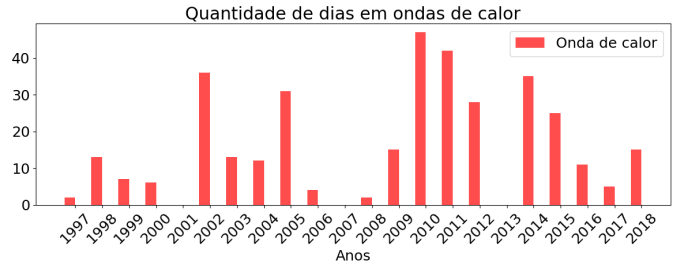


Figure 17: Total de dias em estado de onda de calor por ano.

Podemos observar que através do índice adotado para marcação das ondas de calor o fenômeno ocorre também nas estações mais frias (outono e inverno), no entanto essas ondas são, dentre as temperaturas das ondas de calor, as com temperaturas máximas mais baixas (pontos em amarelo da Figura 16). Isso se torna ainda mais evidente quando observa-se a série de dados dos anos de 2011 e 2014, Figuras 18 e 19, respectivamente, onde as temperaturas máximas são menores no meio do ano (outono e inverno), porém ainda acima da janela de P_{90} do índice calculado, de forma análoga observa-se que as temperaturas máximas são maiores nas bordas da série de dados (primavera e verão) onde se concentram ondas de calor com temperaturas mais elevadas.

Separando-se os dados em dois períodos de aproximadamente uma década, sendo eles 1997-2007 e 2008-2018, a análise de ondas de calor no período indica:

1. O número de ondas de calor na última década foi superior ao da década anterior;
2. A quantidade de dias sob ação de ondas de calor foi superior ao da ultima década.

Destaca-se porém a flutuação anual destes fenômenos e o tamanho da série limitado a duas décadas.

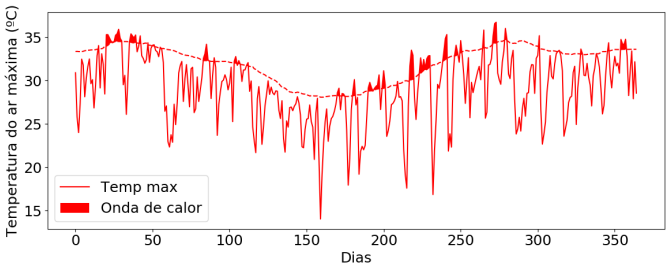


Figure 18: Ondas de calor ao longo de 2011.

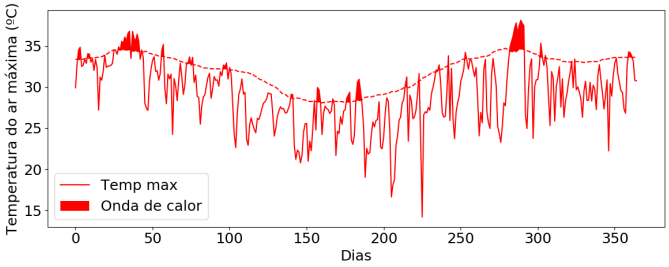


Figure 19: Ondas de calor ao longo de 2014.

Uma análise das demais estações meterológicas de Campinas se mostra promissora como trabalho futuro, visando esclarecer se de fato existe uma tendência de aumento do número de ondas de calor na cidade de Campinas.

4.3 Ondas de frio

Analogamente às ondas de calor, os mesmos estudos foram realizados para as ondas de frio. É possível notar a clara concentração de ondas de frio mais fortes nas estações de outono e inverno, evidenciadas pelas marcações azul escuro e preto na Figura 20.

No entanto, inversamente ao que se observou das ondas de calor, é possível notar a diminuição da frequência de ondas de frio na última década. A Figura 21 mostra como a quantidade de dias sob ondas de frio vêm diminuindo ao longo dos últimos anos, tendo seu pico no ano de 1999.

De forma semelhante as séries de dados de temperaturas máximas mostradas na seção anterior a série de temperaturas mínimas representa bem as estações do ano, tendo seus menores valores nas épocas de outono e inverno, e os maiores valores na primavera e no verão (Figuras 22 e 23).

4.4 Humidex

O *humidex* foi consolidado pelo seu máximo valor para cada dia observado. Nota-se que o *humidex* máximo ocorre em torno das 15:00h do dia, conforme Figura 24, característica esperada de acordo com os especialistas.

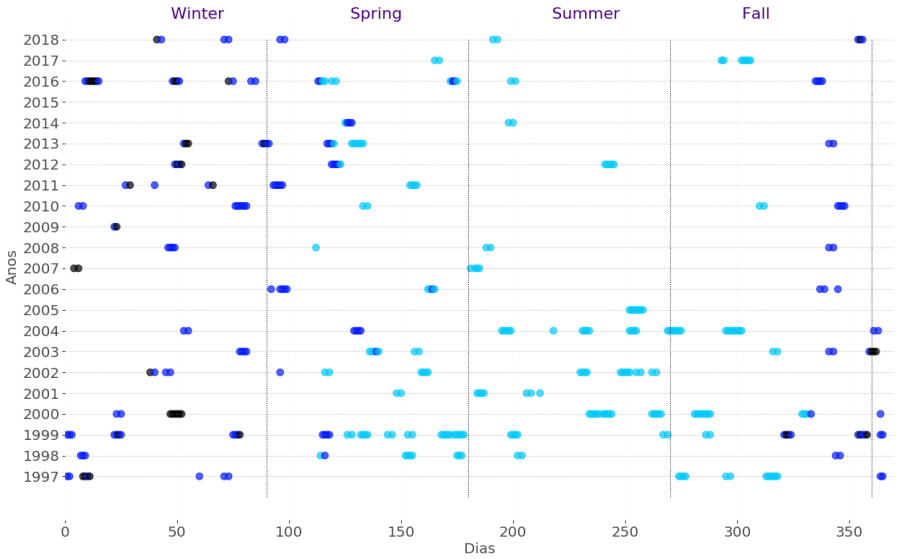


Figure 20: Ondas de frio ao longo dos anos.
Legenda: Azul claro: Ondas de frio acima da média de temperatura das ondas de frio; Azul escuro: Ondas de frio cima do P_{10} de temperatura das ondas de frio; Preto : Ondas de frio abaixo do P_{10} de temperatura das ondas de frio

Figure 21: Total de dias por ano em estado de onda de frio.

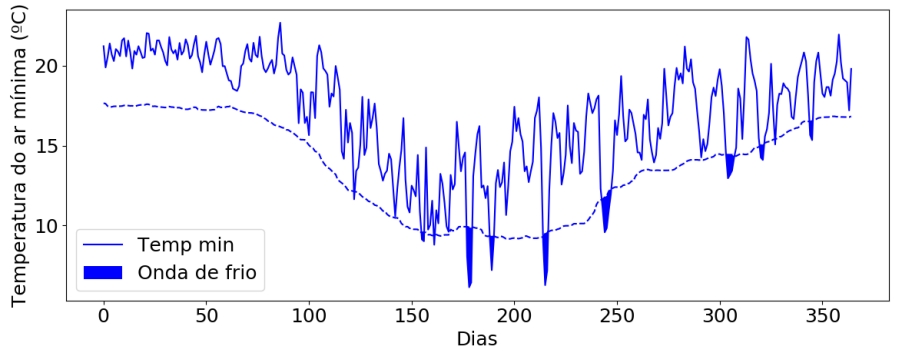


Figure 22: Ondas de frio ao longo de 2011.

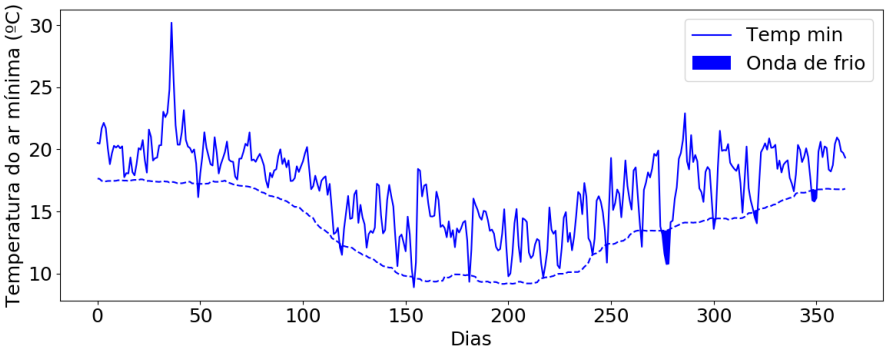


Figure 23: Ondas de frio ao longo de 2014.

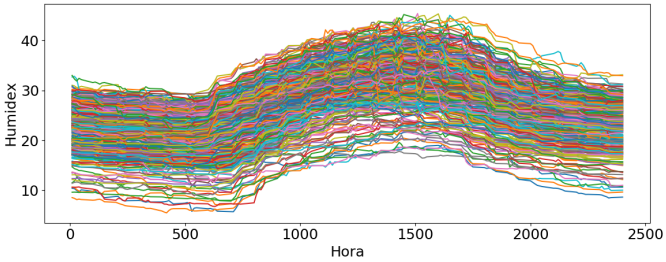


Figure 24: Distribuição do *humidex* de 2018 pelo horário de medição.

De forma semelhante às séries de dados de temperatura, o valor do *humidex* aparenta seguir a mesma tendência de maiores valores nas épocas de primavera e verão, como ilustra as Figuras 25 e 26, onde os valores de *humidex* são maiores nas bordas do gráfico. No entanto, um estudo mais aprofundado se mostra necessário a fim de verificar uma tendência de períodos de desconforto térmico mais frequentes ou maiores nos anos recentes.

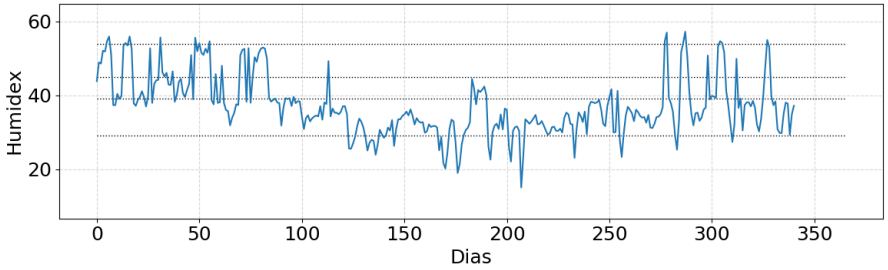


Figure 25: *Humidex* ao longo do ano de 2001.

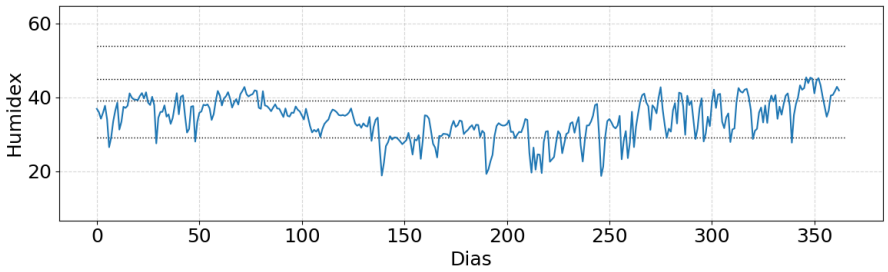


Figure 26: *Humidex* ao longo do ano de 2018.

4.5 Umidade relativa do ar crítica

Analogamente às ondas de frio, refazemos o exercício para umidades relativas mínimas de acordo com a metodologia apresentada na Seção 3.4.4. Como dito na Seção 3.4.4 a umidade do ar é diretamente ligada a saúde humana. Dessa forma, fenômenos críticos envolvendo a umidade relativa do ar são marcadores considerados relevantes para trabalhos futuros do projeto "Clima e saúde". Analisando esse fenômeno, é possível observar que em anos recentes a quantidade de eventos críticos de umidade relativa mínima cresceu consideravelmente, como mostra a Figura 27. Em particular, destaca-se que os dois últimos anos observados (2017 e 2018) tiveram maior ocorrência destes fenômenos.

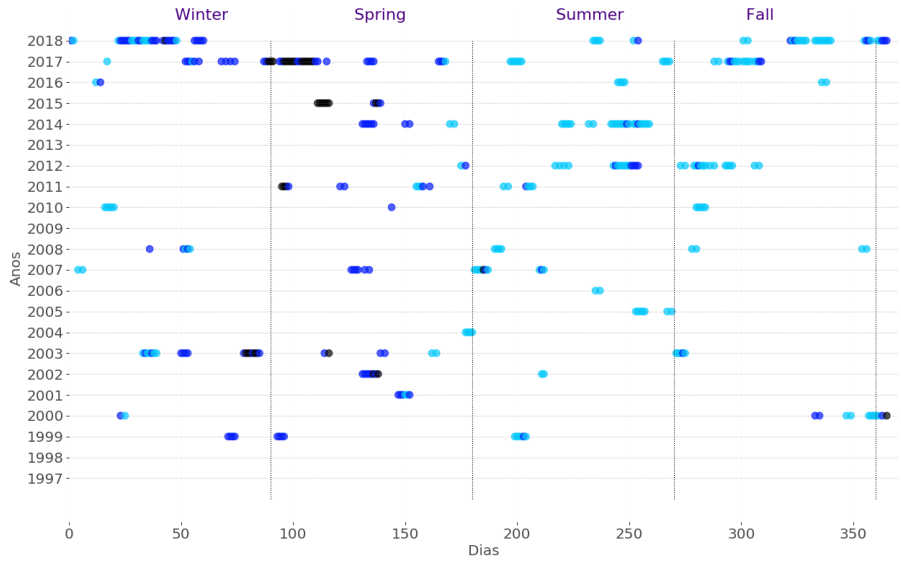


Figure 27: Umidades mínimas críticas ao longo dos anos.
Legenda: Azul claro: Umidades críticas acima da média de umidades críticas; Azul escuro: Umidades críticas cima do P_{10} das umidades críticas; Preto : Umidades críticas abaixo do P_{10} das umidades criticas

A intensidade de ocorrência de umidades mínimas críticas mostra que casos mais amenos encontram-se entre o verão e o outono, fato evidenciado pelas marcações em azul claro na Figura 27, enquanto que casos mais intensos se concentram na transição do inverno para a primavera, fato evidenciado pela alta concentração de marcações azul escuro e pretas também na Figura 27. Na Figura 28 é possível notar claramente como esse efeito aparenta estar ligado a transição de inverno para primavera, assim como eventos críticos de amplitude térmica analisados anteriormente.

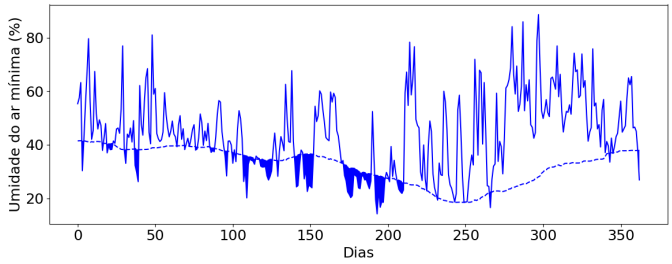


Figure 28: Umidades mínimas críticas ao longo de 2018.

Observando a Figura 29 é possível notar a clara concentração nos últimos dois anos a ocorrência de eventos críticos de umidade.

Umidades muito baixas representam perigo quando se trata da saúde humana, os resultados da presente seção mostram que fenômenos críticos envolvendo umidades relativas mínimas tem se tornado mais frequentes e intensas. Os períodos identificados serão utilizados para estudar correlações com casos médicos do projeto "Clima e Saúde".

Figure 29: Dias em condição crítica de umidade mínima do ar.

5 Conclusão

O tratamento das bases de dados se mostrou uma das maiores dificuldades, visto que os valores inconsistentes encontrados e valores duvidosos nas variáveis exigiam conhecimento maior tanto dos equipamentos da estação meteorológica quanto de especialistas em climatologia.

Através das relações entre variáveis foi possível confirmar padrões típicos propostos pelos especialistas e também detectar padrões não esperados, levantando novas perguntas de pesquisa para trabalhos futuros mais aprofundados, como no caso da queda de pressão atmosférica. Além disso, através das análises gráficas nota-se que novos intervalos de valores podem ser sugeridos para a etapa de limpeza de dados da Seção 3.2.2, gerando menos dados ruidosos, ou esses mesmos valores podem ser estudados mais profundamente em conjunto com dados clínicos para confirmação de casos extremos realísticos. As análises ainda permitiram a detecção de uma possível sensibilidade de medida dos sensores no caso da radiação solar, sendo também um ponto a ser estudado de forma mais aprofundada.

As métricas de eventos extremos se mostraram eficazes, visto que condiziam com conhecimento de especialistas e particularidades da região, como as ondas de calor mais presentes durante o verão e primavera. Com os dados disponíveis (histórico de 22 anos) foi possível observar um indício de crescimento desses cenários extremos, com destaque para períodos críticos de umidades relativa que se mostrou mais frequente em anos recentes, podendo assim se relacionar diretamente com a saúde da população de Campinas. A aplicação dos mesmos testes para períodos mais extensos se mostra promissora para confirmar tendências de crescimento de ondas de calor e umidades relativas críticas, além de confirmar a tendência decrescente de fenômenos de ondas de frio.

O metadados das bases através de arquivos XML se mostrou muito eficaz devido a sua fácil interpretabilidade humana e de máquina, permitindo assim que novas pesquisas sejam realizadas partindo-se das bases climáticas do Cepagri.

Sobretudo, a abordagem de “Ciência de Dados” se mostra eficaz em análises climáticas, permitindo a visualização de relações não óbvias e a flexibilidade para se trabalhar com métricas descritas na literatura. Todo o processo abordado, até a limpeza dos dados se mostrou eficiente em apontar pontos de melhoria, não só em análises como também quanto a falhas sistêmicas de equipamentos.

References

[1] *WHO | Protecting health from climate change: Global research priorities*.

[2] G. Ceccherini, S. Russo, I. Ameztoy, C. Hernandez, and C. Carmona-Moreno, *Magnitude and frequency of heat and cold waves in recent decades: the case of South America*, Natural Hazards and Earth System Sciences Discussions, 3 (2015).

[3] Cgesp, *Cgesp.org. (2019). umidade relativa do ar - cge. [online] disponível em: https://www.cgesp.org/v3/umidade-relativa-do-ar.jsp [acessado 24 jun. 2019].*, 2019.

[4] I. P. O. C. CHANGE, *Climate change 2001 ipcc third assessment report*, Intergovernamental Panel on Climate Change Geneva, IPCC Secretariat, (2001).

[5] U. FAYYAD, G. PIATETSKY-SHAPIRO, AND P. SMYTH, *From data mining to knowledge discovery in databases*, AI magazine, 17 (1996), pp. 37–37.

[6] G. FORZIERI, L. FEYEN, S. RUSSO, M. VOUSDOUKAS, L. ALFIERI, S. OUTTEN, M. MIGLIAVACCA, A. BIANCHI, R. ROJAS, AND A. CID, *Multi-hazard assessment in Europe under climate change*, Climatic Change, 137 (2016), pp. 105–119.

[7] J. L. GEIRINHAS, R. M. TRIGO, R. LIBONATI, AND L. F. PERES, *Climatic Characterization of Heat Waves in Brazil*, Anuário do Instituto de Geociências - UFRJ, 41 (2018), pp. 333–350.

[8] Y. GUO, A. GASPARRINI, S. LI, F. SERA, A. M. VICEDO-CABRERA, M. D. S. Z. S. COELHO, P. H. N. SALDIVA, E. LAVIGNE, B. TAWATSUPA, K. PUNNASIRI, A. OVERCENCO, P. M. CORREA, N. V. ORTEGA, H. KAN, S. OSORIO, J. J. K. JAAKKOLA, N. R. I. RYTI, P. G. GOODMAN, A. ZEKA, P. MICHELOZZI, M. SCORTICHINI, M. HASHIZUME, Y. HONDA, X. SEPOSO, H. KIM, A. TOBIAS, C. ÍÑIGUEZ, B. FORSBERG, D. O. ÅSTRÖM, Y. L. GUO, B.-Y. CHEN, A. ZANOBETTI, J. SCHWARTZ, T. N. DANG, D. D. VAN, M. L. BELL, B. ARMSTRONG, K. L. EBI, AND S. TONG, *Quantifying excess deaths related to heatwaves under climate change scenarios: A multicountry time series modelling study*, 15, p. e1002629.

[9] M. G. LAWRENCE, *The Relationship between Relative Humidity and the Dewpoint Temperature in Moist Air: A Simple Conversion and Applications*, Bulletin of the American Meteorological Society, 86 (2005), pp. 225–234.

[10] W. H. ORGANIZATION, ed., *Climate Change and Human Health - Risks and Responses*, Genève, 2004. OCLC: 254297040.

[11] J. A. OROSA, M. COSTA, RODRÍGUEZ-FERNÁNDEZ, AND G. ROSHAN, *Effect of climate change on outdoor thermal comfort in humid climates*, Journal of Environmental Health Science and Engineering, 12 (2014), p. 46.

[12] S. RUSSO, A. DOSIO, R. G. GRAVERSEN, J. SILLMANN, H. CARRAO, M. B. DUNBAR, A. SINGLETON, P. MONTAGNA, P. BARBOLA, AND J. V. VOGT, *Magnitude of extreme heat waves in present climate and their projection in a warming world*, Journal of Geophysical Research: Atmospheres, 119 (2014), pp. 12,500–12,512.

[13] T. T. SMITH, B. F. ZAITCHIK, AND J. M. GOHLKE, *Heat waves in the United States: definitions, patterns and trends*, Climatic Change, 118 (2013), pp. 811–825.