Projeto de Análise

Andressa Marçal
a262878@dac.unicamp.br

Décio Gonçalves
d226072@dac.unicamp.br

Diego Alyson
d230640@dac.unicamp.br

I. Introdução

Com o grande impacto mundial que a Covid-19 tem causado, em número de infecções e mortes, se faz necessário investigar quais os fatores que corroboram para a disseminação do vírus desta doença. Motivados por esse objetivo de obter trabalhos buscam responder quais os fatores que tem influenciado na propagação do vírus. Para entender quais os fatores tem influenciado este aumento no número de casos, pesquisadores tem relacionado essa variável com algumas outras. Uma das propostas estudadas para tentar modelar as causas da grande quantidade de casos da doença é a relação de influência da temperatura no espalhamento do vírus, seja através do ar ou de superfícies. Também observou-se em outros trabalhos a relação da umidade neste mesmo cenário. Contudo, grande parte destes estudos foram conduzidos em países como China e Irã, ou de modo global sem se atentar as particularidades de cada local. Contudo, os trabalhos não possuem um consenso sobre a influência da temperatura na quantidade de novos casos. Dentre os trabalhos investigados, alguns indicam que com o aumento da temperatura é observada uma redução no número de novos casos [1], [2]. Além disso trabalhos apontam que o vírus decai mais rapidamente quando a umidade ou temperatura aumentam [3]. Este trabalho tem por objetivo investigar a existência dessa correlação no território brasileiro e se baixas ou altas temperaturas junto a variação da umidade provocam variações no número de casos.

II. Questões de pesquisa e Hipótese

Neste trabalho leva-se em consideração a capacidade que o vírus tem de sobreviver em diferentes temperaturas e umidade dentro do território brasileiro. Neste trabalho iremos investigar a influência e a possível correlação entre os valores de temperatura e umidade, com a variação da quantidade de casos de Covid-19 no Brasil.

Para atingir este objetivo, formulamos algumas questões de pesquisa (QP) que irão nos ajudar a responder e validar a existência ou não de correlação de nossa hipótese.

- **QP 1:** Quais fatores ajudam na sobrevivência do vírus no ambiente?
- **QP 2:** Em temperaturas mais baixas é possível observar aumento no número de casos?
- **QP 3:** Em altas temperaturas é possível observar aumento no número de casos?
- **QP 4:** Em quais temperaturas prevalece a maior quantidade de casos?
- **QP 5:** Como as temperaturas influenciam na circulação de pessoas em ambientes públicos? (parques, praias, shoppings)

III. Trabalhos Relacionados

Em seu trabalho, Tobías e Molina [4] realizaram um estudo mundial e histórico sobre o efeito da temperatura. Este trabalho, considerou diferentes variantes do vírus SARS anteriormente observadas para tentar correlacionar os conhecimentos anteriores com esta nova variação encontrada e definida como Covid-19. Observou-se neste estudo que o clima pode afetar a disseminação do vírus, onde baixas temperaturas demonstram uma maior quantidade de infecções diárias.

Como forma de validar os resultados uma regressão Quasi-Poisson foi utilizada, que permite observar características específicas para o contexto de doenças infecciosas. Ajustou-se as tendências lineares e quadráticas, fins de semana e o período de *lockdown*. Considerou-se também a autocorrelação residual. Como o período médio de incubação para COVID-19 é de 5 a 6 dias, foi utilizado um modelo de defasagem distribuída de uma semana.

Outros trabalhos consideram esta mesma observação, Sobral *et al.* [5], analisou as associações entre a transmissão e o número de mortes causadas pelo SARS-CoV-2 e variáveis meteorológicas, temperatura mínima, temperatura máxima e precipitação. Na análise dos dados diários por país utilizou-se um modelo de dados em painel. O trabalho verificou que um aumento na temperatura média diária em 1 grau Fahrenheit reduziu o número de casos em aproximadamente 6,4 casos/dia.

Em um estudo mais localizado, em apenas um país, China, em Peng Shi *et al.* [1], utilizaram a Regressão localmente ponderada e o gráfico de dispersão de suavização (LOESS), modelos não lineares de *lag* distribuído (DLNMs) e meta-análise de efeitos aleatórios para examinar a relação entre a taxa diária de casos confirmados de Covid-19 e as condições de temperatura. Os resultados indicam que com o aumento da temperatura é observada uma redução no número de novos casos.

Malki *et al.* [6] utilizaram vários modelos de regressão para extrair a relação entre diferentes fatores e a taxa de propagação da Covid-19. Os algoritmos de aprendizado de máquina empregados neste trabalho estimam o impacto de variáveis meteorológicas como temperatura e umidade na transmissão do Covid-19, extraindo a relação entre o número de casos confirmados e as variáveis meteorológicas em determinadas regiões. Concluiu-se neste trabalho que a temperatura

e umidade são características importantes para a previsão da taxa de mortalidade da Covid-19.

Os seguintes modelos de aprendizado de máquina, como modelos lineares (regressão linear, regressão de laço, regressão de crista, rede elástica, regressão de ângulo mínimo, regressão de ângulo mínimo de laço, busca de correspondência ortogonal, cume bayesiano, determinação de relevância automática, regressor agressivo passivo, Random Sample Consensus, TheilSen Regressor, Huber Regressor) foram utilizados.

Além disso, foram utilizado modelos baseados em *ensemble* como *Random Forest*, *Extra Trees Regressor*, *AdaBoost Regressor* e *Gradient Boosting Regressor*. *Extreme Gradient Boosting* (XGBoost), *Light Gradient Boosting Machine* (LightGBM) e *CatBoost Regressor*), *Kernel Ridge*, *Support Vector Machine* (SVM), *K-Nearest Neighbours Regressor* (KNN), *Multi-layers Perceptron* (MLP) e *Decision Tree* foram utilizados para a previsão da propagação do coronavírus.

Biryukov *et al.* [3] apresenta um estudo que correlaciona o aumento da temperatura, umidade relativa e a aceleração da inativação do SARS-CoV-2 em superfícies. Os resultados mostram que o SARS-CoV-2 decaiu mais rapidamente quando a umidade ou a temperatura aumentaram, mas o volume da gota (1 a 50 l) e o tipo de superfície (aço inoxidável, plástico ou luva de nitrila) não afetaram significativamente a taxa de decomposição.

Quando na temperatura ambiente (24 ° C), a meia-vida do vírus variou de 6,3 a 18,6 h dependendo da umidade relativa, mas foi reduzida para 1,0 a 8,9 h quando a temperatura foi aumentada para 35 ° C. Estes resultados sugerem um alto potencial de transmissão, por horas a dias, em ambientes internos. Uma análise de regressão foi realizada para determinar um modelo preditivo de decaimento.

No trabalho de Jahangiri *et al.* [2] os autores realizaram uma análise de sensibilidade e especificidade da temperatura ambiente e tamanho da população na taxa de transmissão do novo coronavírus (Covid-19) em diferentes províncias do Irã. Neste trabalho observou-se resultados positivos em relação a confirmação que o número de casos de Covid-19 em climas mais quentes é menor do que em climas moderados ou frios, mas não encontrou-se provas científicas para isto. Estas demonstrações foram feitas através de uma curva ROC da taxa de transmissão do coronavírus, mas observou-se que o parâmetro da temperatura ambiente tem uma relação linear com o número de pessoas afetadas com Covid-19.

Iqbal *et al.* [7] verificou a correlação entre o Covid-19, temperatura e taxa de câmbio na cidade de Wuhan. Os resultados gerais sugerem a insignificância de um aumento na temperatura para conter ou retardar as novas infecções por Covid-19 apresentando resultados contrários a muitos estudos anteriores que sugerem um papel significativo da temperatura em desacelerar a propagação da Covid-19.

Para encontrar tais resultados, foram utilizados, Transformada Wavelet Contínua (CWT), Coerência da Transformada Wavelet (WTC), Coerência Wavelet Parcial (PWC) e Coerência Wavelet Múltipla (MWC) para analisar a associação entre a temperatura média diária de Wuhan, número diário de novos casos de Covid-19 na cidade de Wuhan e a taxa de câmbio RMB.

IV. Bases de Dados

Nesta seção apresentamos as bases de dados que serão utilizadas para embasamento e testes deste projeto. No total exploramos dados provenientes de três bases de dados com o objetivo de obter dados de diferentes fontes que possam ser utilizados para validar as questões de pesquisa do trabalho. Para alcançar o objetivo exploratório de obter informações de possíveis correlações entre o clima e a propagação da Covid-19 utilizamos as bases INMET, Brasil.io e Google Mobility. Estas bases, respectivamente, contêm dados metereológicos, do avanço da covid-19 e de mobilidade, todos referentes ao Brasil.

A. Dataset INMET (Instituto Nacional de Meteorologia)

Com o intuito de obter dados meteorológicos, como variação de temperatura e umidade, para responder os questionamentos iniciais do trabalho, a variação de casos e o impacto do clima, temperatura e umidade nesta variação, buscou-se diferentes *datasets* climáticos das regiões brasileiras. Nesta primeira busca, verificou-se que em sua grande maioria os dados relacionados à clima, nacional e internacional são disponibilizados de forma privada e paga. No INMET (Instituto Nacional de Meteorologia) obteve-se dados gratuitos de históricos em território nacional, por campos de estações meteorológicas do instituto.

Neste trabalho propomos realizar um estudo dos impactos meteorológicos de âmbito nacional e, como observações, algumas cidades selecionadas, considerando-se os diferentes climas e taxas de umidade do país. A base de dados do INMET disponibiliza estes dados por cidade e por histórico temporal, por hora, possibilitando sua utilização para esta pesquisa.

Para realizar a integração, levou-se em consideração três diferentes níveis de observação. Em um primeiro nível, as informações de clima nacional, no segundo nível proposto, observou-se os dados regionalmente, por estados e , por último, para uma melhor observação de localidade, foram selecionadas algumas diferentes cidades, com climas diferenciados. Segue na Figura 1 uma observação geral dos dados para uma única cidade, após tratamento inicial.

B. Dataset Covid-19 (Brasil.IO)

Com o intuito de obter informações sobre a quantidade de novos no Brasil optamos por usar o *dataset* de Covid-19 do site Brasil.IO. Este *dataset* possui os casos confirmados e óbitos obtidos dos boletins das Secretarias Estaduais de Saúde (SES). Os dados foram enriquecidos, de forma que a partir do momento em que um município confirma um caso, ele sempre aparecerá nessa tabela (mesmo que para uma determinada data a SES não tenha liberado o boletim - nesse caso é repetido o dado do dia anterior). A coleta dos dados iniciaram do dia 25 de Fevereiro de 2020. Temos nele 27 estados e 5.294 cidades. Na Figura 2 podemos observar as informações das suas colunas e os tipos dos seus dados após tratamento inicial.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 274 entries, 0 to 273
Data columns (total 18 columns):
 #   Column                                          Non-Null Count  Dtype
---  ------                                          --------------  -----
 0   Data                                            274 non-null    period[D]
 1   PRECIPITAÇÃO TOTAL, HORÁRIO (mm)                274 non-null    float64
 2   PRESSAO ATMOSFERICA AO NIVEL DA ESTACAO, HORARIA (mB)  274 non-null  float64
 3   PRESSÃO ATMOSFERICA MAX.NA HORA ANT. (AUT) (mB)  274 non-null    float64
 4   PRESSÃO ATMOSFERICA MIN. NA HORA ANT. (AUT) (mB)  274 non-null   float64
 5   RADIACAO GLOBAL (Kj/m²)                         274 non-null    float64
 6   TEMPERATURA DO AR - BULBO SECO, HORARIA (°C)    274 non-null    float64
 7   TEMPERATURA DO PONTO DE ORVALHO (°C)            274 non-null    float64
 8   TEMPERATURA MÁXIMA NA HORA ANT. (AUT) (°C)      274 non-null    float64
 9   TEMPERATURA MÍNIMA NA HORA ANT. (AUT) (°C)      274 non-null    float64
 10  TEMPERATURA ORVALHO MAX. NA HORA ANT. (AUT) (°C)  274 non-null  float64
 11  TEMPERATURA ORVALHO MIN. NA HORA ANT. (AUT) (°C)  274 non-null  float64
 12  UMIDADE REL. MAX. NA HORA ANT. (AUT) (%)        274 non-null    float64
 13  UMIDADE REL. MIN. NA HORA ANT. (AUT) (%)        274 non-null    float64
 14  UMIDADE RELATIVA DO AR, HORARIA (%)             274 non-null    float64
 15  VENTO, DIREÇÃO HORARIA (gr) (° (gr))            274 non-null    float64
 16  VENTO, RAJADA MAXIMA (m/s)                      274 non-null    float64
 17  VENTO, VELOCIDADE HORARIA (m/s)                 274 non-null    float64
dtypes: float64(17), period[D](1)
memory usage: 38.7 KB
```

Figure 1. Tipos de Dados Recebidos, Pós-Processamento

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 925921 entries, 0 to 925920
Data columns (total 18 columns):
 #   Column                                          Non-Null Count   Dtype
---  ------                                          --------------   -----
 0   city                                            919869 non-null  object
 1   city_ibge_code                                  922174 non-null  float64
 2   date                                            925921 non-null  object
 3   epidemiological_week                            925921 non-null  int64
 4   estimated_population                            922174 non-null  float64
 5   estimated_population_2019                       922174 non-null  float64
 6   is_last                                         925921 non-null  bool
 7   is_repeated                                     925921 non-null  bool
 8   last_available_confirmed                        925921 non-null  int64
 9   last_available_confirmed_per_100k_inhabitants   908366 non-null  float64
 10  last_available_date                             925921 non-null  object
 11  last_available_death_rate                       925921 non-null  float64
 12  last_available_deaths                           925921 non-null  int64
 13  order_for_place                                 925921 non-null  int64
 14  place_type                                      925921 non-null  object
 15  state                                           925921 non-null  object
 16  new_confirmed                                   925921 non-null  int64
 17  new_deaths                                      925921 non-null  int64
dtypes: bool(2), float64(5), int64(6), object(5)
memory usage: 114.8+ MB
```

Figure 2. Informações do Dataset Covid19

C. Dataset Google Mobility (Google)

Como forma de identificar possíveis tendências de comportamento como a relação entre o isolamento social e a quantidade de casos optamos por analisar dados de mobilidade. Para isso escolhamos utilizar a base Google Mobility[1]. A Figura 3 apresenta as características que a base fornece.

```
RangeIndex: 430432 entries, 0 to 430431
Data columns (total 14 columns):
 #   Column                                                Non-Null Count   Dtype
---  ------                                                --------------   -----
 0   country_region_code                                   430432 non-null  object
 1   country_region                                        430432 non-null  object
 2   sub_region_1                                          430185 non-null  object
 3   sub_region_2                                          423516 non-null  object
 4   metro_area                                            0 non-null       float64
 5   iso_3166_2_code                                       6669 non-null    object
 6   census_fips_code                                      0 non-null       float64
 7   date                                                  430432 non-null  object
 8   retail_and_recreation_percent_change_from_baseline    182567 non-null  float64
 9   grocery_and_pharmacy_percent_change_from_baseline     177081 non-null  float64
 10  parks_percent_change_from_baseline                    156331 non-null  float64
 11  transit_stations_percent_change_from_baseline         132303 non-null  float64
 12  workplaces_percent_change_from_baseline               404785 non-null  float64
 13  residential_percent_change_from_baseline              179990 non-null  float64
```

Figure 3. Informações do Dataset Google Mobility

[1]https://www.google.com/covid19/mobility/

O Google mobility contêm dados que reportam um valor de referência com base na mediana do período pré pandemia. Podemos observar, através de observações iniciais a tendência no Brasil do período em que foi declarado o estado pandêmico, é possível observar que é neste período em que as pessoas estiveram mais tempo em suas casas. Também é possível observar que este índice vem decaindo e está tendendo a se tornar próximo aos valores em períodos normais. Contundo é importante analisar mais informações para tirar conclusões do comportamento dessas curvas.

V. Modelos

Para responder as questões de pesquisa elencadas, precisamos de modelos que auxiliem neste processo. Dentre os modelos que serão utilizados neste trabalho, pretendemos utilizar assim como [4] um regressão Quasi-Poisson dada a sua natureza de conseguir observar características específicas de doenças infecciosas. Além disso alguns outros modelos de regressão podem ser úteis principalmente na interpretação da correlação entre as variáveis analisadas.

VI. Cronograma

- Andressa Marçal:
  - Analisar informações onde foi registrado picos de temperaturas elevadas e baixa umidade, com quantidade de novos casos naquela semana e na semana epidemiológica posterior, para avaliar a possibilidade de sobrevivência do vírus em ambientes mais quentes e secos.
    Data da entrega: 10/11
  - Analisar informações onde foi registrado picos de temperaturas elevadas e alta umidade, com quantidade de novos casos naquela semana e na semana epidemiológica posterior, para avaliar a possibilidade de sobrevivência do vírus em ambientes mais quentes e úmidos.
    Data da entrega: 10/11
  - Analisar informações onde foi registrado picos de temperaturas baixas e alta umidade, com os novos casos registrados naquela semana e na semana epidemiológica posterior, para avaliar a possibilidade de sobrevivência do vírus em ambientes mais frios e úmidos.
    Data da entrega: 17/11
  - Analisar informações onde foi registrado picos de temperaturas baixas e baixa umidade, com os novos casos registrados naquela semana e na semana epidemiológica posterior, para avaliar a possibilidade de sobrevivência do vírus em ambientes mais frios e secos.
    Data da entrega: 17/11
- Décio Gonçalves:
  - Analisar informações de mobilidade em lugares públicos em dias com temperaturas mais elevadas e cruzar com casos registrados na semana epidemiológica posterior;

Data da entrega: 24/11

- Analisar informações de mobilidade em lugares públicos em dias com temperaturas mais baixas e cruzar com casos registrados na semana epidemiológica posterior;
  Data da entrega: 24/11
- Analisar informações de mobilidade na semana epidemiológica que mais teve registro de novos casos
  Data da entrega: 24/11
- Analisar informações de mobilidade na semana epidemiológica que menos teve registro de novos casos
  Data da entrega: 24/11
- Analisar informações de mobilidade na semana epidemiológica em que teve a reabertura do comercio, podemos usar como exemplo, São Paulo e cruzar essa informação com a semana epidemiológica posterior.
  Data da entrega: 30/11
- Analisar informações de mobilidade em feriados e cruzar com a informação de registro de novos casos na semana epidemiológica posterior.
  Data da entrega: 30/11

- Diego Alyson:
  - Analisar a temperatura da semana epidemiológica que mais teve registros de novos casos
    Data da entrega: 17/11
  - Analisar a temperatura da semana epidemiológica que menos teve registros de novos casos
    Data da entrega: 17/11
  - Modelagem (Treino/Teste/Validação)
    Data limite: 15/12
  - Avaliação do desempenho e métricas do algoritmo utilizado
    Data limite: 30/12
  - Coleta dos resultados finais
    Data limite: 04/01
- Escrita do Artigo

REFERENCES

[1] P. Shi, Y. Dong, H. Yan, C. Zhao, X. Li, W. Liu, M. He, S. Tang, and S. Xi, "Impact of temperature on the dynamics of the covid-19 outbreak in china," *Science of The Total Environment*, p. 138890, 2020.

[2] M. Jahangiri, M. Jahangiri, and M. Najafgholipour, "The sensitivity and specificity analyses of ambient temperature and population size on the transmission rate of the novel coronavirus (covid-19) in different provinces of iran," *Science of The Total Environment*, p. 138872, 2020.

[3] J. Biryukov, J. A. Boydston, R. A. Dunning, J. J. Yeager, S. Wood, A. L. Reese, A. Ferris, D. Miller, W. Weaver, N. E. Zeitouni *et al.*, "Increasing temperature and relative humidity accelerates inactivation of sars-cov-2 on surfaces," *MSphere*, vol. 5, no. 4, 2020.

[4] A. Tobías and T. Molina, "Is temperature reducing the transmission of covid-19?" *Environmental Research*, vol. 186, p. 109553, 2020.

[5] M. F. F. Sobral, G. B. Duarte, A. I. G. da Penha Sobral, M. L. M. Marinho, and A. de Souza Melo, "Association between climate variables and global transmission of sars-cov-2," *Science of The Total Environment*, vol. 729, p. 138997, 2020.

[6] Z. Malki, E.-S. Atlam, A. E. Hassanien, G. Dagnew, M. A. Elhosseini, and I. Gad, "Association between weather data and covid-19 pandemic predicting mortality rate: Machine learning approaches," *Chaos, Solitons & Fractals*, vol. 138, p. 110137, 2020.

[7] N. Iqbal, Z. Fareed, F. Shahzad, X. He, U. Shahzad, and M. Lina, "Nexus between covid-19, temperature and exchange rate in wuhan city: New findings from partial and multiple wavelet coherence," *Science of The Total Environment*, p. 138916, 2020.