Data Visualization: Assignment 2

Vishruth Vijay
IMT2022507
vishruth.vijay@iiitb.ac.in

Shreyas Arun Saggere
IMT2022006
shreyas.saggere@iiitb.ac.in

Sathvik S Rao
IMT2022082
sathvik.rao@iiitb.ac.in

***Abstract*—The document consists of Scientific Visualization of Gridmet data for the time period July 2018 to September 2018, Network visualization of Subreddit hyperlink data using graph layout algorithms and Interactive visualization of OxCGRT data.**

I. Scientific Visualization

A. Scalar Field visualization- Color mapping

Color Mapping is a scalar field visualization technique which is used to depict the variation of a scalar quantity like temperature, pressure etc. spatially. Color mapping translates scalar values into colors using defined color maps. Sequential colormaps and diverging colormaps are usually used for such tasks.

Colormaps have been used to determine the major contributing factors of burning index in the United States during the period July 2018 - September-2018

1) ***Dataset****:* The dataset used in this study is derived from GridMet data [3] for the year 2018. This dataset provides meteorological data across a spatial grid consisting of 585 latitude and 1386 longitude values. For the period under analysis, from July 2018 to September 2018, various meteorological features such as temperature, wind speed, relative humidity, and potential evapotranspiration were collected. Each feature was initially stored in a separate file, with daily records available for each variable.

2) ***Data Preparation****:* To ensure the dataset was manageable and representative, a subset of 10 days was linearly sampled from the July to September period using xarray[13] and pandas[7]. This sampling method allowed for efficient temporal analysis while maintaining key variability within the dataset. The data was then processed to merge individual files into a new sampled dataset, with each day represented as .nc file. Each .nc file consisted of all data variables collected. The maximum air temperature (tmmx) and minimum air temperature values(tmmn) were averaged out. All the data variables were studied after local/global min-max normalization.

3) ***Study of Burning Index****:* The burning index is a critical indicator used in fire weather forecasting to assess fire potential. This is used to study the patterns of forest fires. Based on the data given, a correlation matrix was plotted to study the dependence between different data variables.

Based on the Figure 1, we see precipitation amount, wind speed, wind velocity, air temperature etc. To have correlation with the data. Using the correlation value, a fire hazard score was calculated, plotted and compared with the actual burning index value as shown in Figure 2.

Figure 2 depicts that not all correlated values actually contribute to the burning index. After performing an ablation study, the color map closest to the burning index score was achieved by:

$$\text{Fire hazard score} = \frac{0.34 \cdot \text{PE} + 0.22 \cdot \text{WS} + 0.1 \cdot \text{T} - 0.38 \cdot \text{RH}}{0.34 + 0.22 + 0.1 + 0.38}$$

PE : potential evapotranspiration
WS : wind speed
T : air temperature
RH : relative humidity

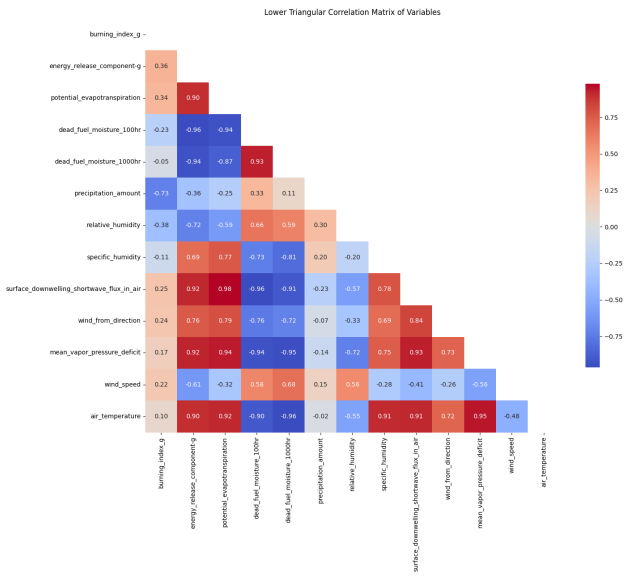

Fig. 1: A correlation matrix of the different meteorological data variables. The second column depicts the relation of different data variables with the burning index.

This concludes that PE,WS, T and RH were the key factors influencing forest fire risk.

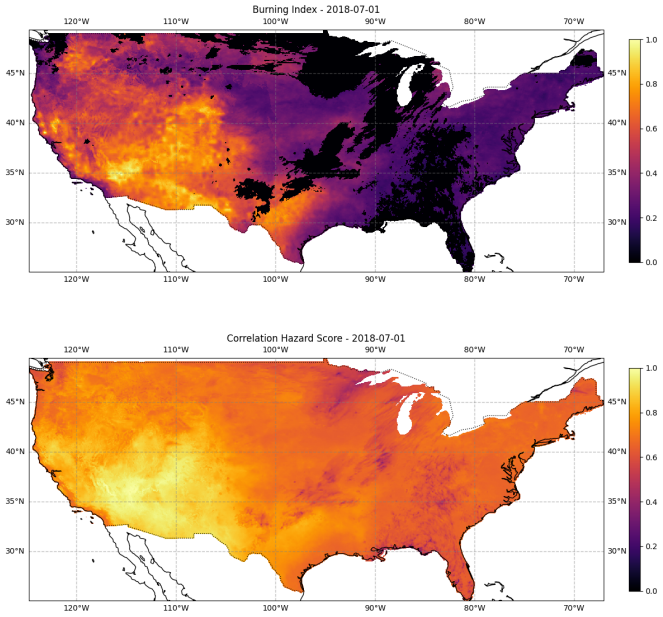

Fig. 2: The figure highlights the dissimilarity between the correlated values and burning index.

*4) **Strategies used for plotting**:* Experiments were done on multiple color mapping techniques, using different sequential color maps, scaling and normalizing techniques.

- **Sequential Color Maps**: Color maps like *viridis*, *inferno*, and *YlOrRd* were tested to represent the burning index data, with each map providing a smooth progression from low to high risk.
- **Diverging color maps**: Color maps like *coolwarm* were compared with the sequential color maps. Diverging color maps were not used for further analysis due to no meaningful middle value. But I-C3 uses diverging color maps as it use the middle value as the mean temperature.
- **Continuous Scaling**: This scaling applies a seamless gradient to the burning index data, helping capture subtle variations across a continuous risk spectrum.
- **Discrete Scaling**: Discrete scaling divides the burning index into set color bands, making distinct risk levels more visually apparent and easier to interpret.
- **Logarithmic Scaling**: This scaling compresses higher values, enhancing the visibility of lower burning index values, useful when data varies greatly.
- **Normalization** The local normalization technique gave more insight into the data than the global counterpart. Most of the figures plotted using global normalization, gave very similar plots across timestamps.

At the end, the combination of *inferno discrete scaling* and *local normalization* techniques proved most effective for clearly distinguishing values of different scores.

*5) **Observations and inferences**:* The data provided was a time-series data over the months of July 2018 to September 2018. For the study of forest fires, a few days have been sampled. Figures 9 to 16 depict the burning index and the calculated fire hazard score corresponding to the features discussed above. The plots illustrate that July and August were devastating months for the western United States, with California, Colorado, and New Mexico being the hardest-hit states. Figure 9 highlights peak burning index levels in these regions. Major fires ignited in early July, including the Pawnee Fire in California [11], the Spring Creek Fire in Colorado [12]. The wind direction, seen in Figure 19, may have contributed to the spread of these fires towards the east. The cause of these forest fires were analysed to be due to higher temperatures, lesser humidity and other human activity.

California continued to show a slightly higher burning index due to the Ferguson Fire (Yosemite National Park) [10] which was contained by mid-August. Multiple fires continued to emerge in the state of California, in the month of September, while the overall fires in the country had reduced.

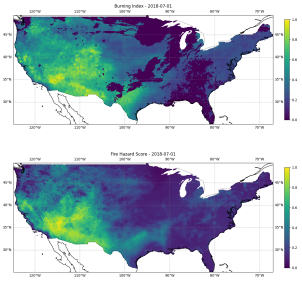

Fig. 3: Plotting the calculated fire hazard score and burning index score using the *viridis* continuous color map.

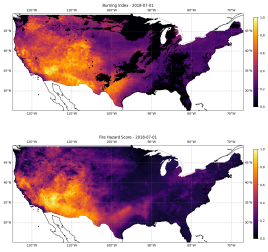

Fig. 4: Plotting the calculated fire hazard score and burning index score using the *inferno* continuous color map.

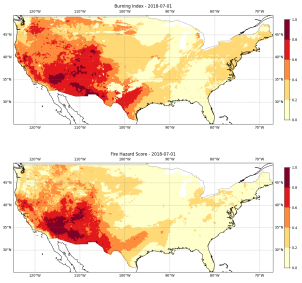

Fig. 5: Plotting the calculated fire hazard score and burning index score using the *YlOrRd* discrete color map.

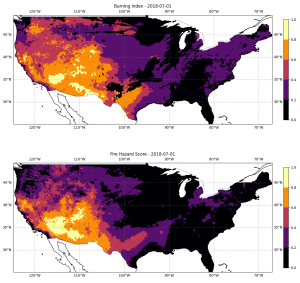

Fig. 6: Plotting the calculated fire hazard score and burning index score using the *inferno* discrete color map.

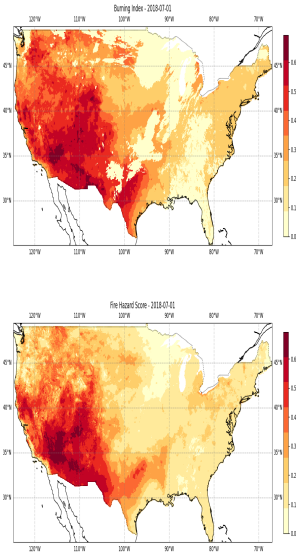

Fig. 7: Plotting the calculated fire hazard score and burning index score using the *YlOrRd* log discrete color map.

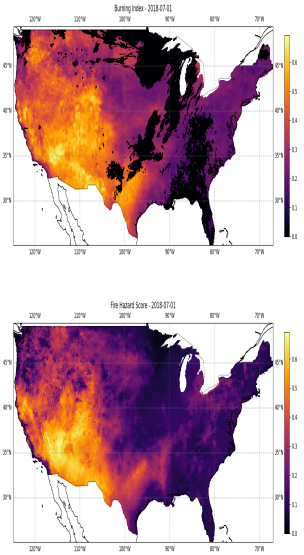

Fig. 8: Plotting the calculated fire hazard score and burning index score using the *inferno* log continuous color map.

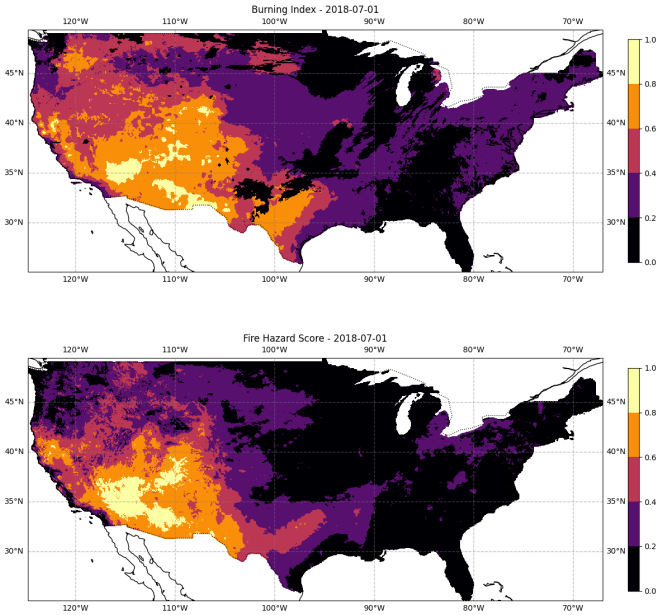

Fig. 11: Burning Index data and calculated fire hazard score of the USA on July 1, 2018

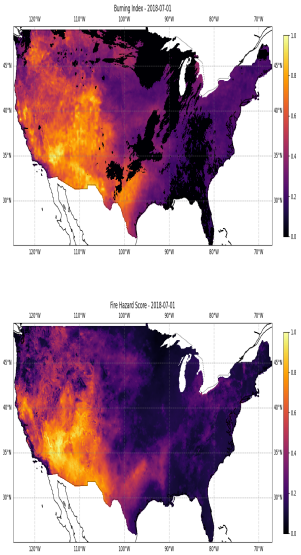

Fig. 9: Plotting the calculated fire hazard score and burning index score using the *inferno* continuous color map with global normalization

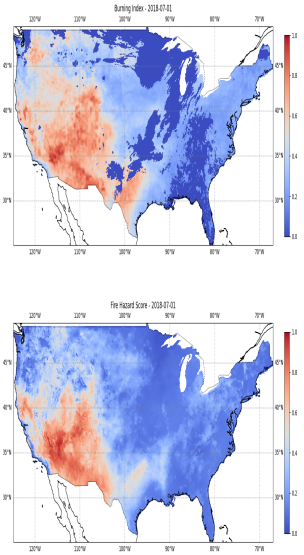

Fig. 10: Plotting the calculated fire hazard score and burning index score using the *diverging* color map.

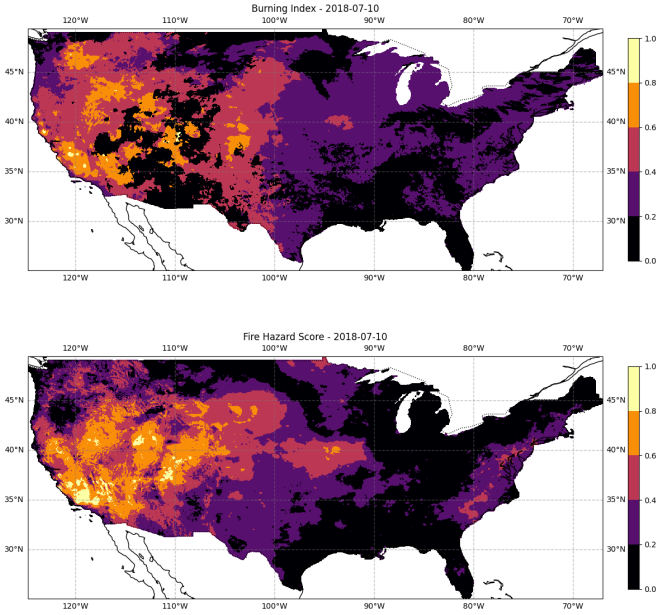

Fig. 12: Burning Index data and calculated fire hazard score of the USA on July 10, 2018

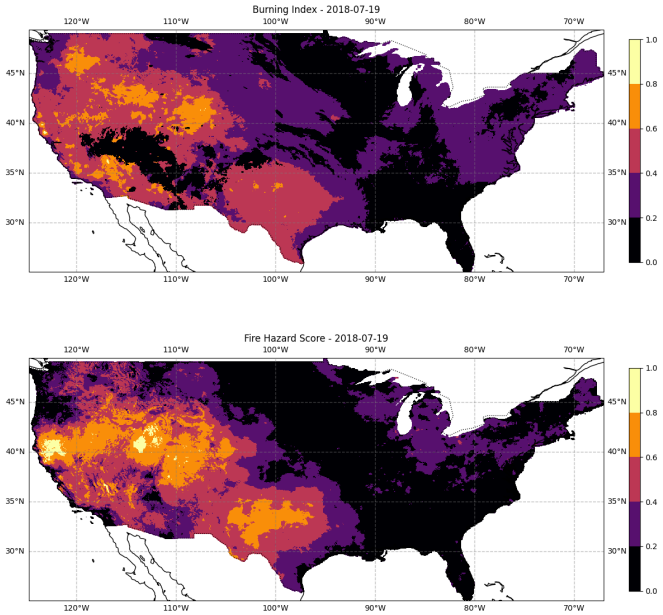

Fig. 13: Burning Index data and calculated fire hazard score of the USA on July 19, 2018

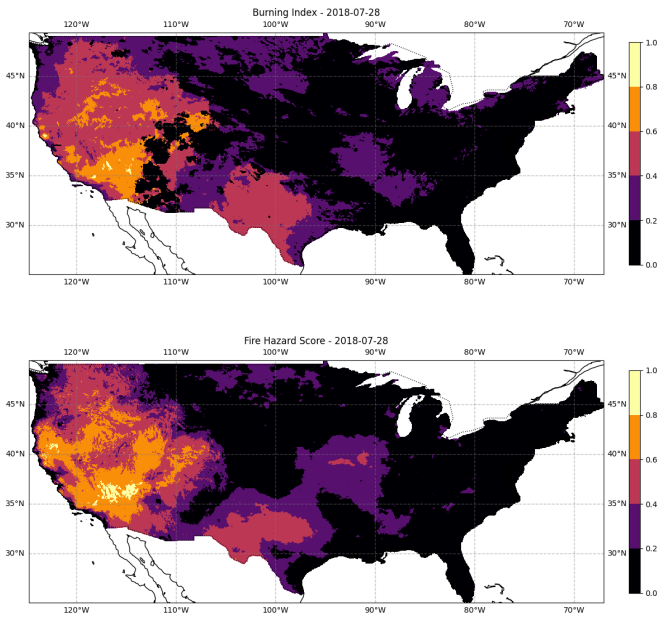

Fig. 14: Burning Index data and calculated fire hazard score of the USA on July 28, 2018

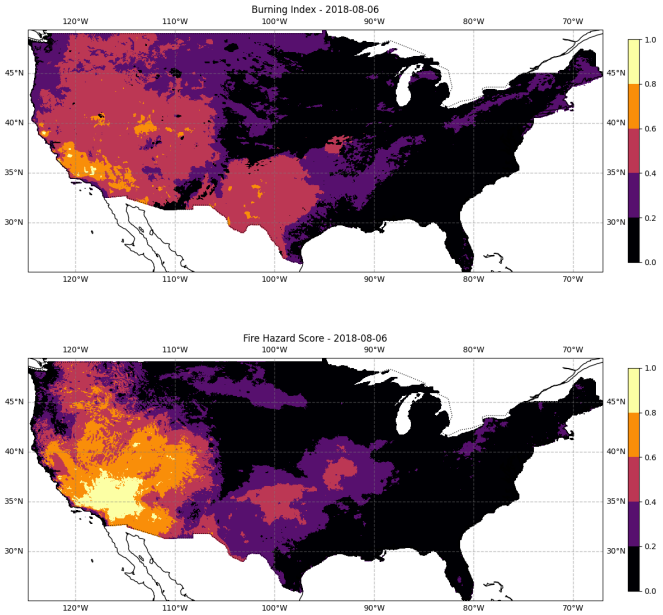

Fig. 15: Burning Index data and calculated fire hazard score of the USA on August 6, 2018

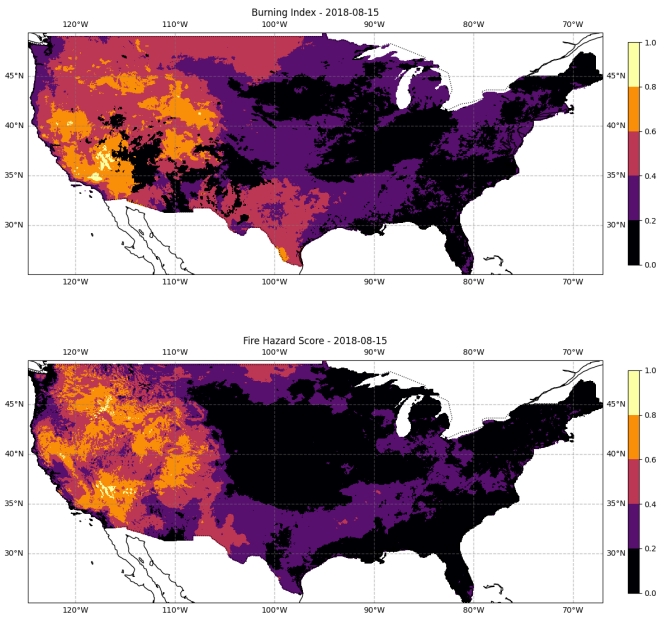

Fig. 16: Burning Index data and calculated fire hazard score of the USA on August 15, 2018

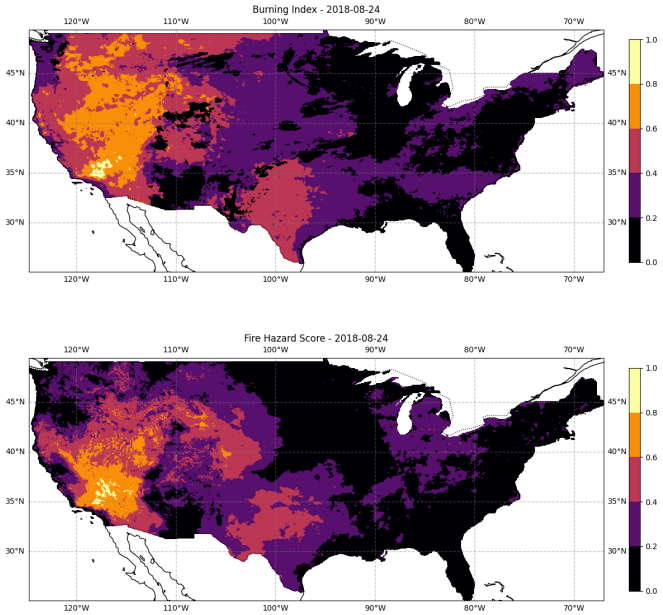

Fig. 17: Burning Index data and calculated fire hazard score of the USA on August 24, 2018

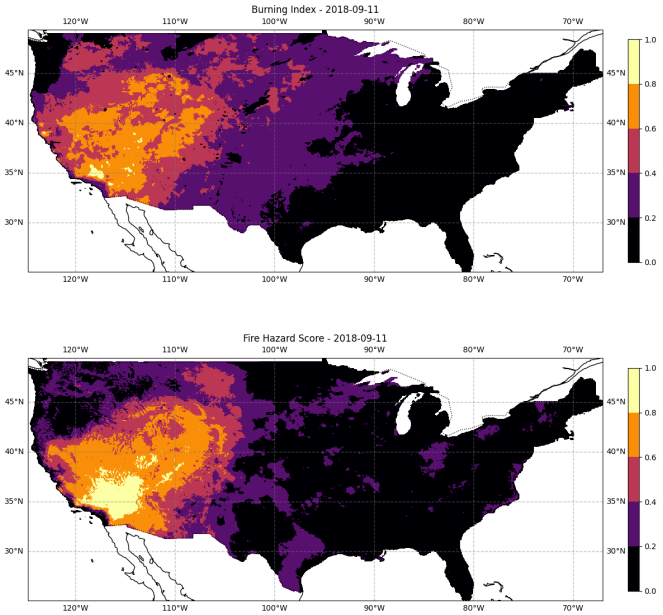

Fig. 19: Burning Index data and calculated fire hazard score of the USA on September 11, 2018

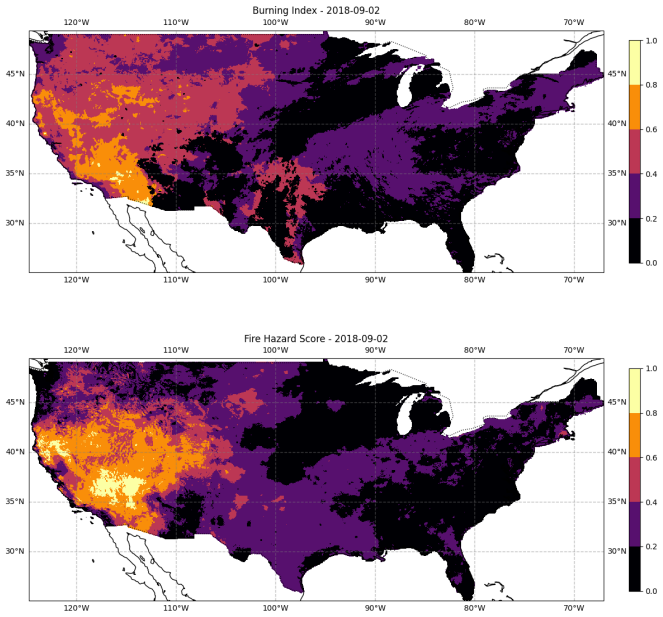

Fig. 18: Burning Index data and calculated fire hazard score of the USA on September 2, 2018

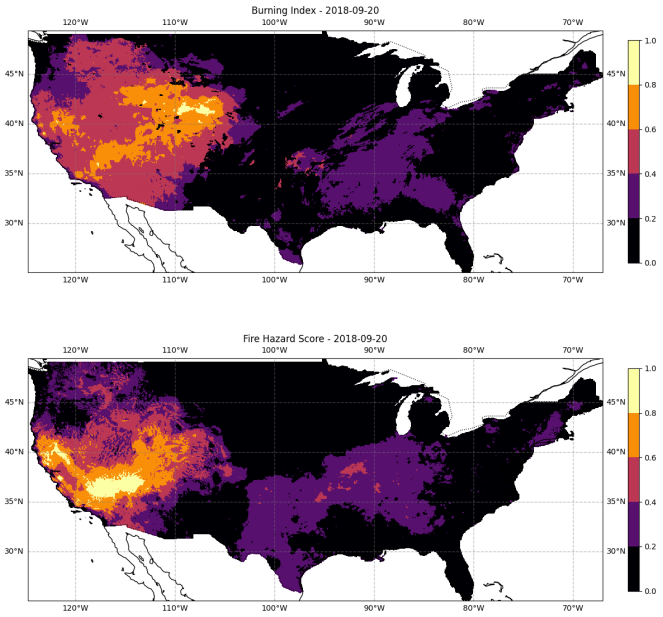

Fig. 20: Burning Index data and calculated fire hazard score of the USA on September 20, 2018

B. Quiver-Plot

Quiver plots are a powerful technique for visualizing vector fields, where both the direction and magnitude of a vector quantity (such as wind velocity or ocean currents) are represented graphically. In these plots, arrows are drawn to indicate the direction of the vector field at specific spatial points, with the length and/or color of the arrows encoding the magnitude of the vector. Quiver plots provide an intuitive way to represent physical phenomena like wind patterns, ocean currents, and electromagnetic fields.

This study uses quiver plots to represent wind speed(in m/s) and direction over a geographical area using data from July 2018 to September 2018. Two types of quiver plots were produced: one in which the arrow length represents wind speed, and another in which the arrow color represents wind speed, with constant arrow length.

*1) **Dataset**:* The same dataset that as mentioned in the I-A1 was used for the study of wind direction and wind speed with the help of a quiver plot

*2) **Data Preparation**:* The same subset of days as mentioned in I-A2 was plotted for to maintain consistency and to analyze the correlation between different data variables

*3) **Study of Wind speed and direction**:* As shown in [9] and [8], wind speed and wind direction have a correlation with the potential spread of wildfires. Therefore, analyzing wind speed and wind direction is also a crucial component in fire weather forecasting

*4) **Strategies Used For Plotting**:* To reduce the complexity and improve computational efficiency, the data is binned into larger latitude and longitude intervals (2° bins). This binning process averages the wind speed(in m/s) and direction within each spatial bin, providing a smoother and more interpretable dataset for visualization.

The wind speed(in m/s) and direction are then converted into vector components (u and v), where u represents the eastward (or westward) component of the wind, and v represents the northward (or southward) component. Two versions of the vector field are computed:

- Where the arrow length corresponds to the wind speed(in m/s)
- Where the arrow length is fixed, but the wind speed(in m/s) is represented by color using different color maps

This process allows for clear and informative quiver plots that can visualize both wind magnitude and direction across the region. By using the global minimum and maximum wind speeds across all files, a consistent color scale is applied to ensure uniform interpretation across different time periods.

Different color maps as shown in 21 were experimented with to inspect which color-map gave a more intuitive representation of the data. Viridis color map to be the most visually appealing and a more accurate representation of the data. Hence, Viridis color maps was chosen to visualize the quiver plots.

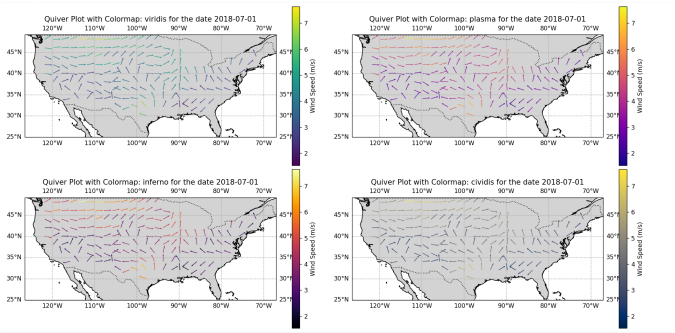

Fig. 21: Quiver Plot using different color-maps to indicate the magnitude of wind speed(in m/s)

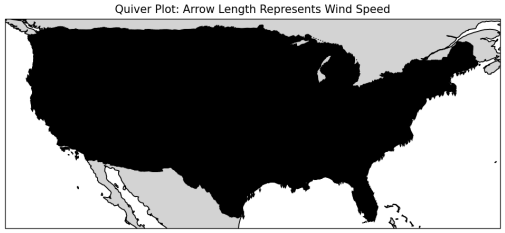

Fig. 22: Plotting the quiver plots without binning where the length of the arrow indicates the magnitude of wind speed(in m/s).

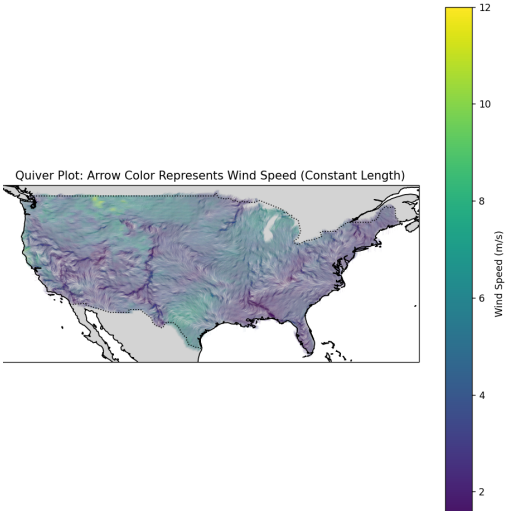

Fig. 23: Plotting the quiver plots without binning where the length of the arrow indicates the magnitude of wind speed(in m/s).

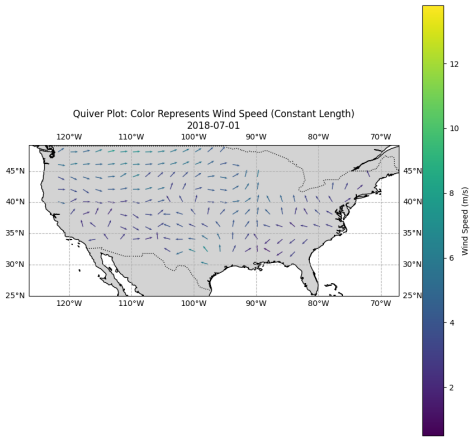

Fig. 24: Plotting the quiver plot for wind direction and speed and using the viridis color map to depict the magnitude of wind speed(in m/s) for 2018-07-01.

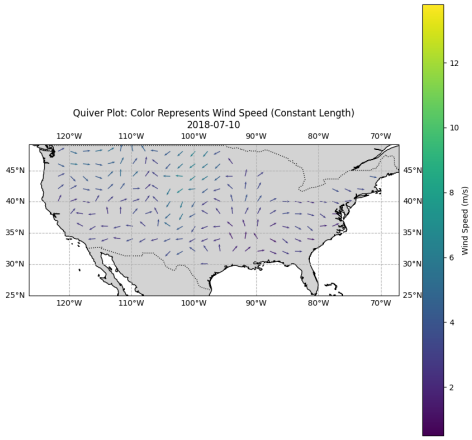

Fig. 25: Plotting the quiver plot for wind direction and speed and modelling the length of the arrow to be proportional to the magnitude of the wind speed(in m/s) for 2018-07-10.

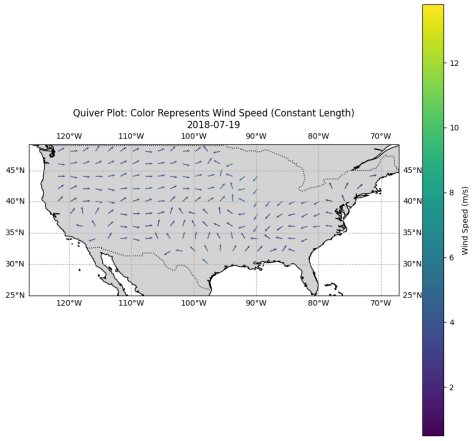

Fig. 26: Plotting the quiver plot for wind direction and speed and modelling the length of the arrow to be proportional to the magnitude of the wind speed(in m/s) for 2018-07-19.

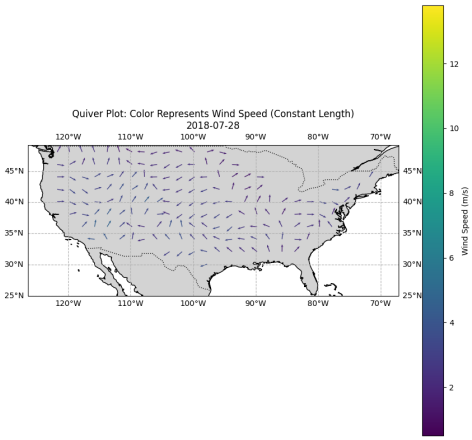

Fig. 27: Plotting the quiver plot for wind direction and speed and modelling the length of the arrow to be proportional to the magnitude of the wind speed(in m/s) for 2018-07-28

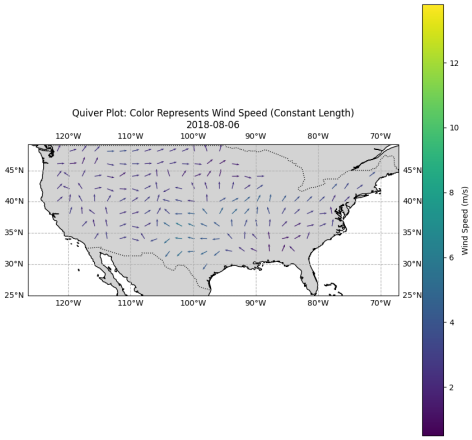

Fig. 28: Plotting the quiver plot for wind direction and speed and modelling the length of the arrow to be proportional to the magnitude of the wind speed(in m/s) for 2018-08-06

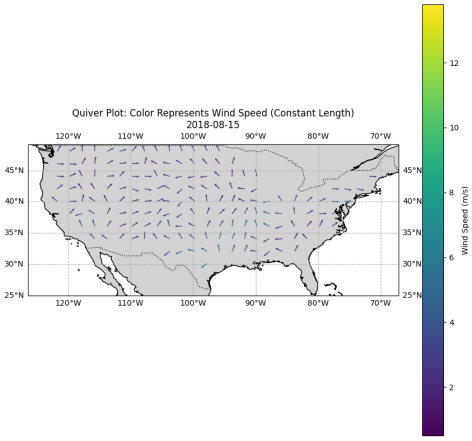

Fig. 29: Plotting the quiver plot for wind direction and speed and modelling the length of the arrow to be proportional to the magnitude of the wind speed(in m/s) for 2018-08-15

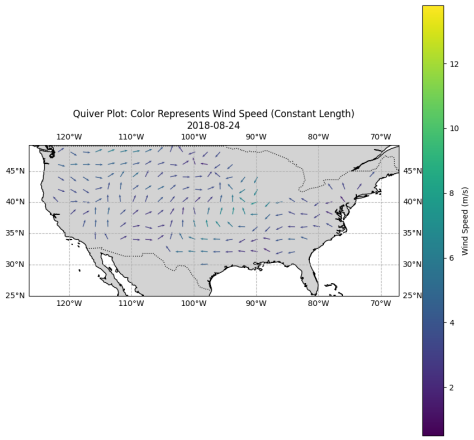

Fig. 30: Plotting the quiver plot for wind direction and speed and modelling the length of the arrow to be proportional to the magnitude of the wind speed(in m/s) for 2018-08-24

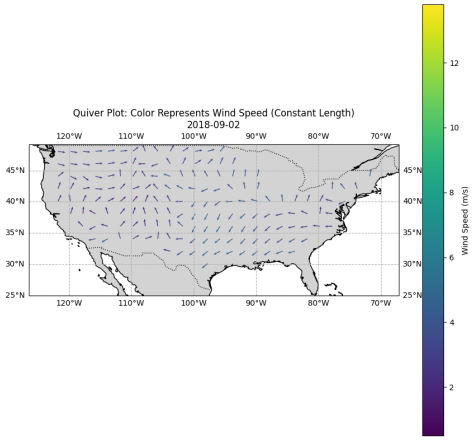

Fig. 31: Plotting the quiver plot for wind direction and speed and modelling the length of the arrow to be proportional to the magnitude of the wind speed(in m/s) for 2018-09-02

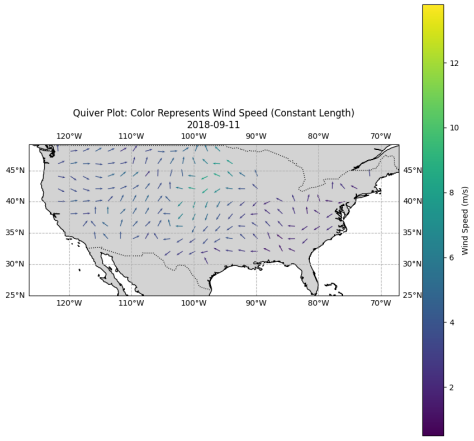

Fig. 32: Plotting the quiver plot for wind direction and speed and modelling the length of the arrow to be proportional to the magnitude of the wind speed(in m/s) for 2018-09-11

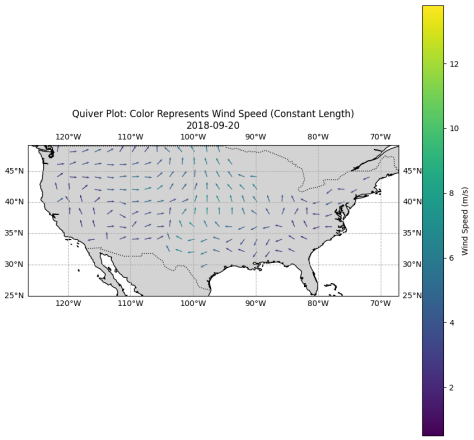

Fig. 33: Plotting the quiver plot for wind direction and speed and modelling the length of the arrow to be proportional to the magnitude of the wind speed(in m/s) for 2018-09-20

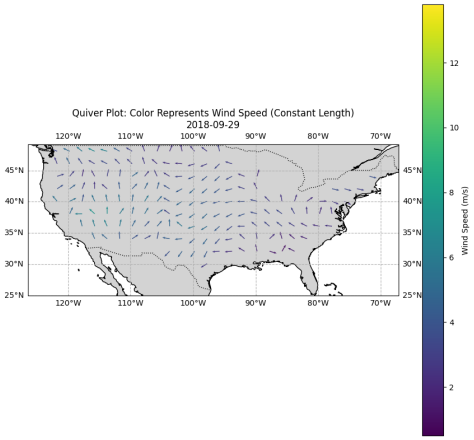

Fig. 34: Plotting the quiver plot for wind direction and speed and modelling the length of the arrow to be proportional to the magnitude of the wind speed(in m/s) for 2018-09-29

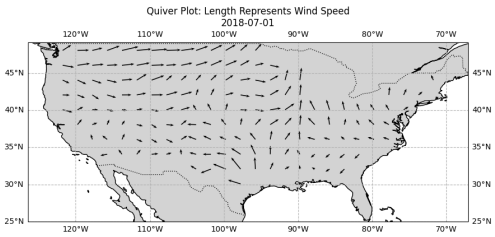

Fig. 35: Plotting the quiver plot for wind direction and speed and modelling the length of the arrow to be proportional to the magnitude of the wind speed(in m/s) for 2018-07-01.

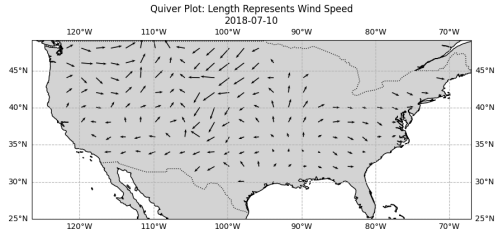

Fig. 36: Plotting the quiver plot for wind direction and speed and modelling the length of the arrow to be proportional to the magnitude of the wind speed(in m/s) for 2018-07-10.

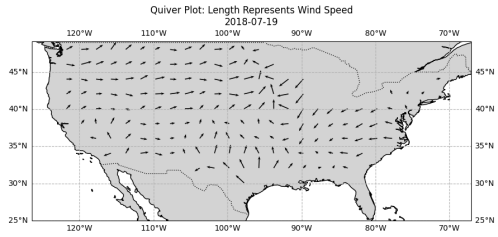

Fig. 37: Plotting the quiver plot for wind direction and speed and modelling the length of the arrow to be proportional to the magnitude of the wind speed(in m/s) for 2018-07-19.

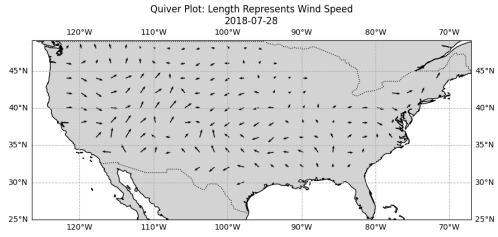

Fig. 38: Plotting the quiver plot for wind direction and speed and modelling the length of the arrow to be proportional to the magnitude of the wind speed(in m/s) for 2018-07-28.

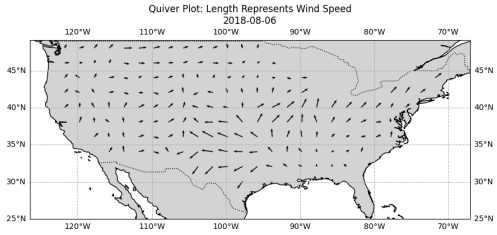

Fig. 39: Plotting the quiver plot for wind direction and speed and modelling the length of the arrow to be proportional to the magnitude of the wind speed(in m/s) for 2018-08-06

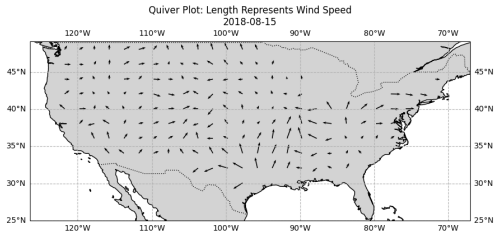

Fig. 40: Plotting the quiver plot for wind direction and speed and modelling the length of the arrow to be proportional to the magnitude of the wind speed(in m/s) for 2018-08-15

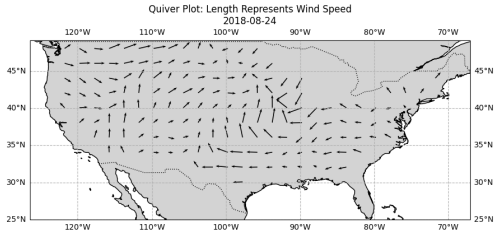

Fig. 41: Plotting the quiver plot for wind direction and speed and modelling the length of the arrow to be proportional to the magnitude of the wind speed(in m/s) for 2018-08-24

C. Contour Plots

Contour maps are graphical representations of three-dimensional data on a two-dimensional plane, where lines or color gradients connect points of equal value. They enable easy identification of regions with similar values, highlight gradients or transitions between high and low values, and make complex data easier to interpret at a glance. This study uses contour maps to analyze air temperature patterns across the United States. By representing temperature variations through contour lines or color gradients, these maps provide a clear visualization of temperature distribution and gradients over space.

1) ***Dataset****:* For the contour plots, the same GridMet dataset described in Part I-A of the report is utilized. Only the daily air temperature data from the pre-processed dataset are used in this analysis to study temperature patterns over time and across different geographic locations.

2) ***Data preparation****:* The same subset of days as described in Part I-A2 was used to plot the contour plots. This was done to maintain consistency and derive collective inferences at a later stage.

3) ***Approach****:* In this study, the contour and contourf functions were used to visualize air temperature data using two distinct methods: the marching squares algorithm and the contour fill method respectively.
Basic approach that was followed in this study to plot the contour plots:

- The netCDF4 library was used to read and write data in netCDF format, which is a standard for storing multi-dimensional scientific data, including variables such as temperature, humidity, and wind speed(in m/s). Here, netCDF4 enabled access to GridMet data files, containing a grid of latitude and longitude values with meteorological data for the July to September 2018 study period.
- xarray was used to load data from netCDF files and convert it into a more manageable format for plotting.
- The contour function was used to implement the marching squares algorithm, producing contour lines that represent areas of equal air temperature, while contourf was applied for the contour fill method, creating a color-filled gradient to visualize temperature distributions across geographic space.

The following methods were used for the plots:

1) **Marching Squares Algorithm :** Using the marching squares algorithm with contour, we generated precise contour lines representing temperature intervals over the dataset. The levels were defined using np.linspace() to create a range of evenly spaced intervals, ensuring consistent visual representation of different temperature ranges. Unlike traditional contour lines in black, this implementation used lines colored according to the specified color map (e.g., viridis, coolwarm, jet), allowing each line to correspond visually with its temperature range. This approach enhanced the map's continuity and made it easier to trace gradual temperature changes across regions.
2) **Contour fill algorithm :** This plot used the same levels as the marching squares plot, keeping interval values consistent across both visualizations. Various color maps, including jet, rdbu, coolwarm, and viridis, were tested to evaluate their ability to represent temperature differences clearly. To complement the filled contours, contour lines with a finer line width of 0.2 were added using the same color scheme, aligning with the filled regions and enhancing the gradient's clarity.

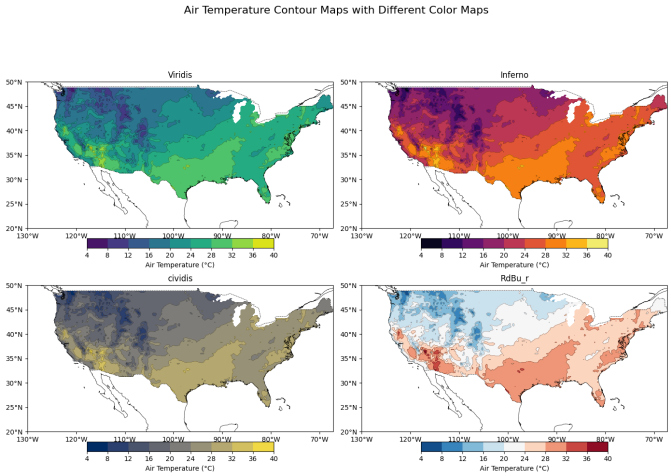

Fig. 42: The different color maps used for the contour fill method of creating contour plot - from top left to bottom right : viridis, inferno, plasma and redblue

In selecting an optimal color map for visualizing air temperature patterns, we evaluated many different options including viridis, inferno, redblue, cividis (as shown in fig.42), rainbow and coolwarm. After comparing these options, we chose the "coolwarm" color map for its effectiveness in conveying temperature-related insights. "coolwarm" provides a balanced, perceptually intuitive transition from blue to red, highlighting both cooler and warmer regions clearly. While the rainbow color map offered vibrant distinctions, it can sometimes create misleading visual artifacts. Therefore, "coolwarm" was the most suitable colour map for this study.

In experimenting with different line widths for air temperature contour maps, we aimed to determine the optimal thickness. Figure 43 shows contour maps created without contour fill at various line widths (0.2, 0.5, 1.0, and 1.5). After further experimentation with different line widths, we found 1.2 to be optimal for contour plots without contour fill. However for contour maps with contour fill, a thinner line width of around 0.2 was ideal, as it complemented the color-filled regions by providing subtle boundary definitions.

After experimenting with various number of levels for the

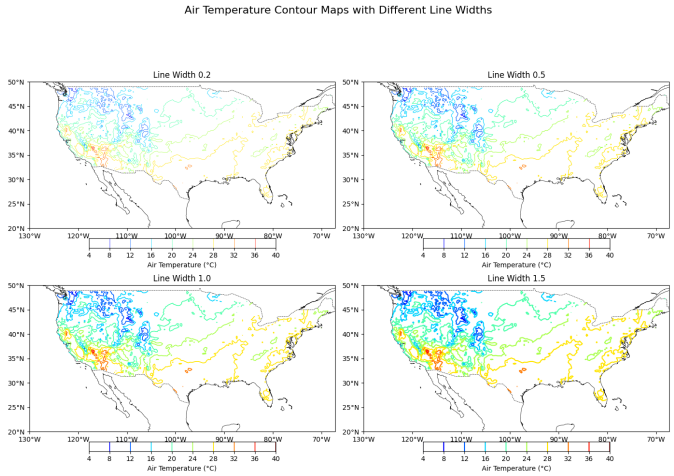

Fig. 43: The plots show the different line widths ranging from 0.5 (top left) to 2.0 (bottom right) that were tested out to find the most appropriate line width for visually appealing contour plots and not cause visual clutter

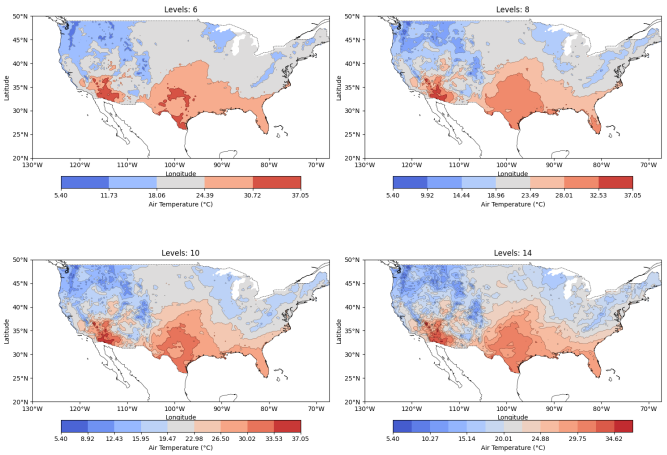

Fig. 44: The plots show the different levels ranging from 5 (top left) to 20 (bottom right) that were tested out to find the most appropriate line width for visually appealing contour plots and not cause visual clutter

contour plots (Fig. 42), we found that 10 was the optimal number of levels. Anything more, like 14, caused visual clutter, with overlapping contour lines that reduced readability. Lesser number of levels reduced the resolution of information displayed.

We compared contour maps with and without contour fill to evaluate which better represents air temperature patterns. Contour fill proved to be more effective, as it uses continuous color gradients to represent temperature variations across regions, making patterns and transitions visually clear and easy to interpret. The comparison is shown in Figure 45.

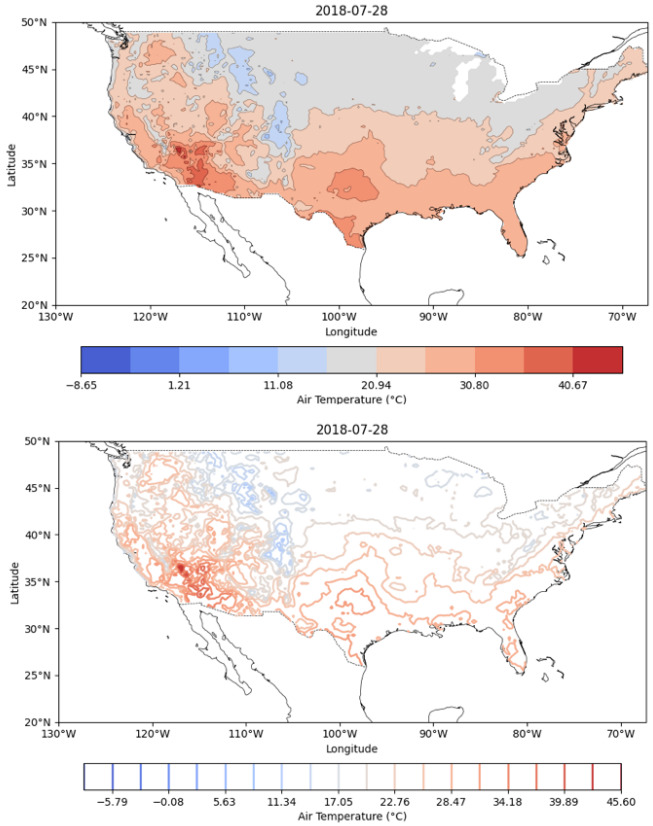

Fig. 45: The figure displays two types of contour maps representing air temperature over the United States for the date 2018-07-28. The top plot using contour fill method and the bottom plot using contour lines only (marching squares algorithm utilized).

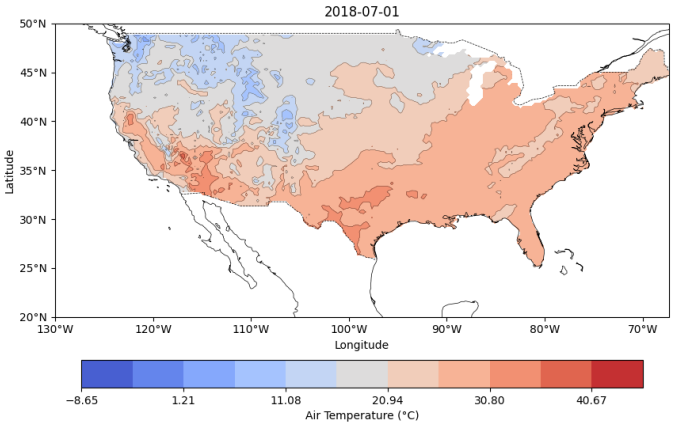

Fig. 46: Contour plot using the contour fill method for July 1, 2018, with the coolwarm color map

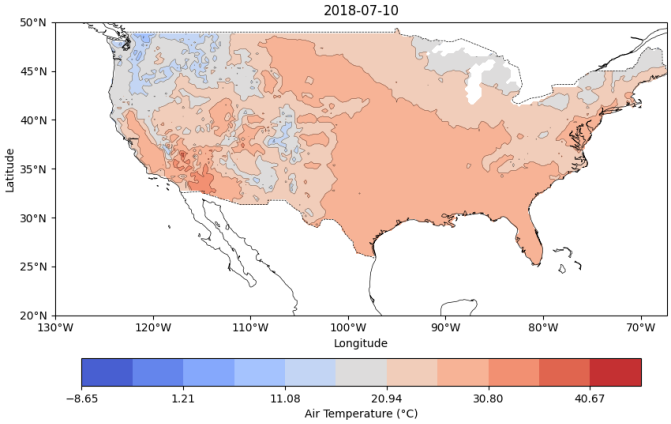

Fig. 47: Contour plot using the contour fill method for July 10, 2018, with the coolwarm color map

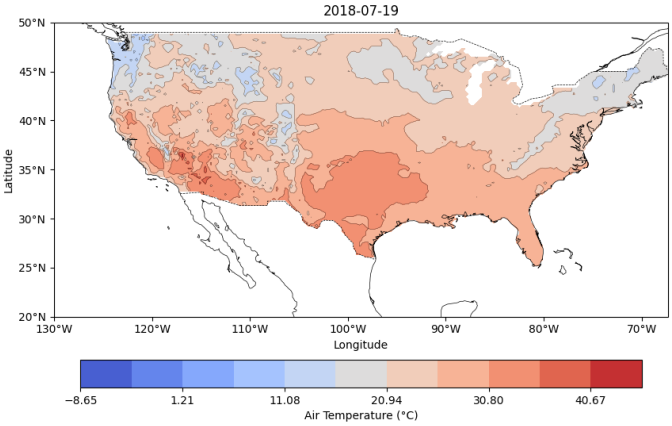

Fig. 48: Contour plot using the contour fill method for July 19, 2018, with the coolwarm color map

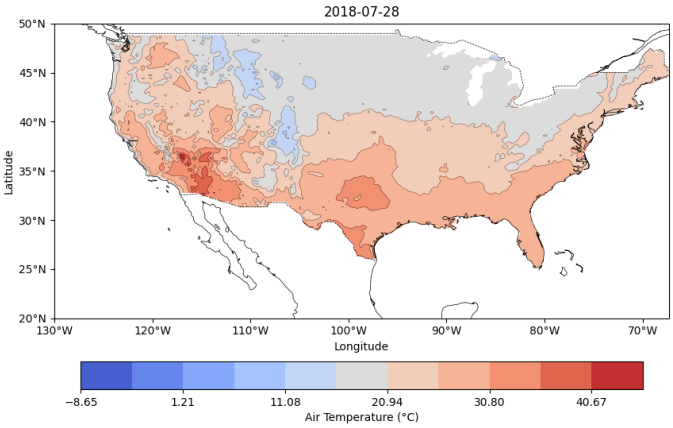

Fig. 49: Contour plot using the contour fill method for July 28, 2018, with the coolwarm color map

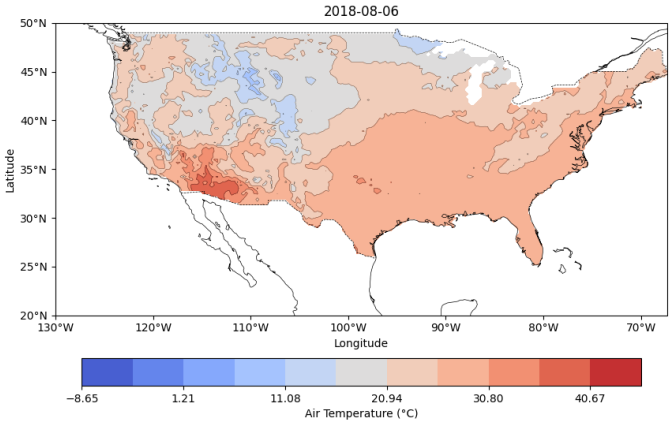

Fig. 50: Contour plot using the contour fill method for August 6, 2018, with the coolwarm color map

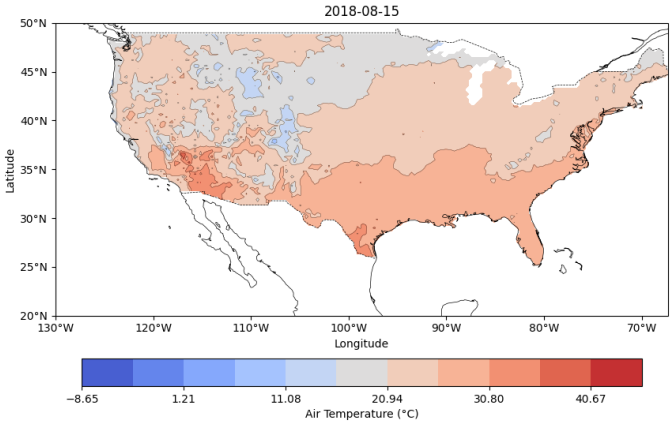

Fig. 51: Contour plot using the contour fill method for August 15, 2018, with the coolwarm colormap

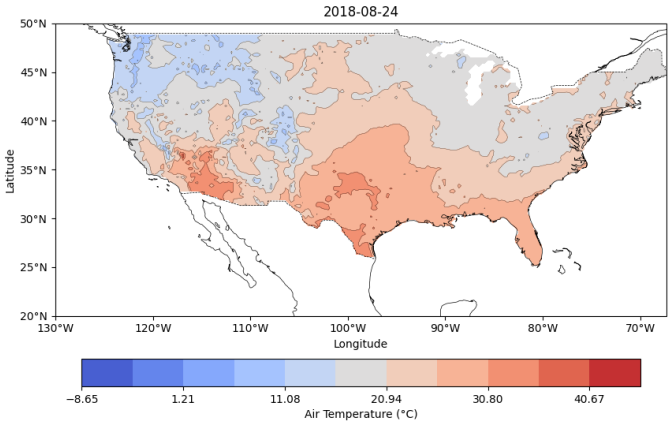

Fig. 52: Contour plot using the contour fill method for August 24, 2018, with the coolwarm colormap

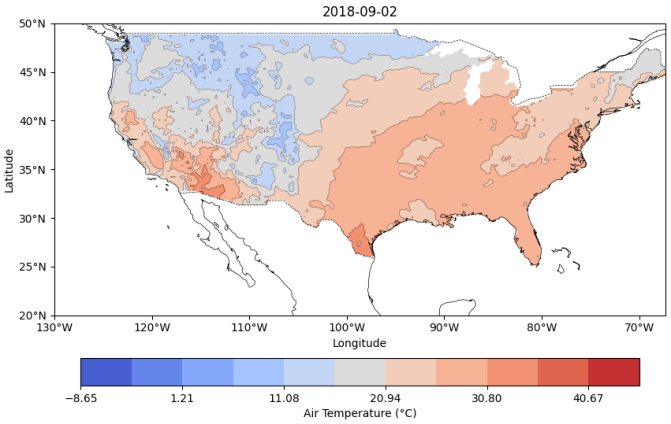

Fig. 53: Contour plot using the contour fill method for September 2, 2018, with the coolwarm colormap

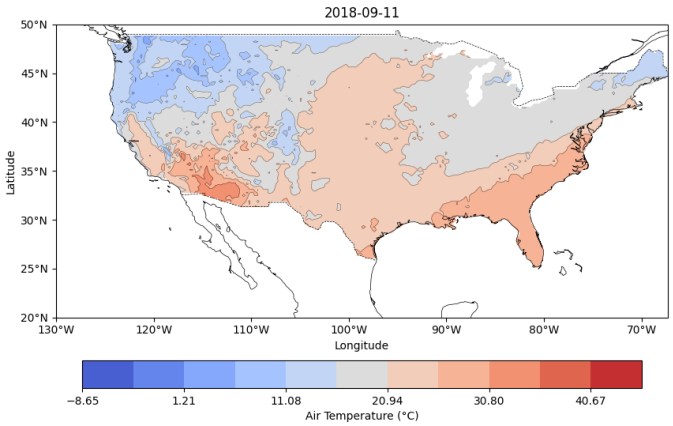

Fig. 54: Contour plot using the contour fill method for September 11, 2018, with the coolwarm colormap

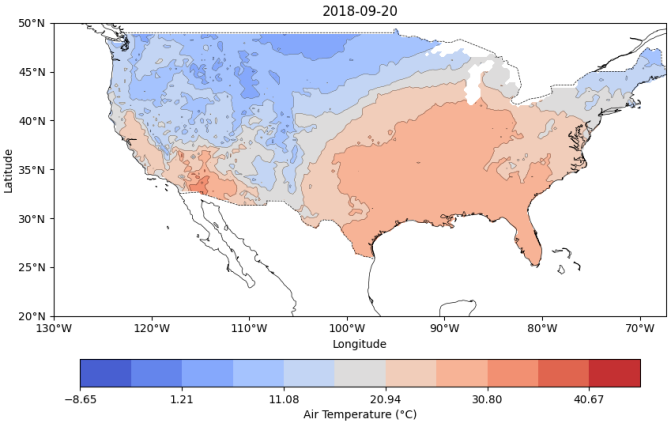

Fig. 55: Contour plot using the contour fill method for September 20, 2018, with the coolwarm color map

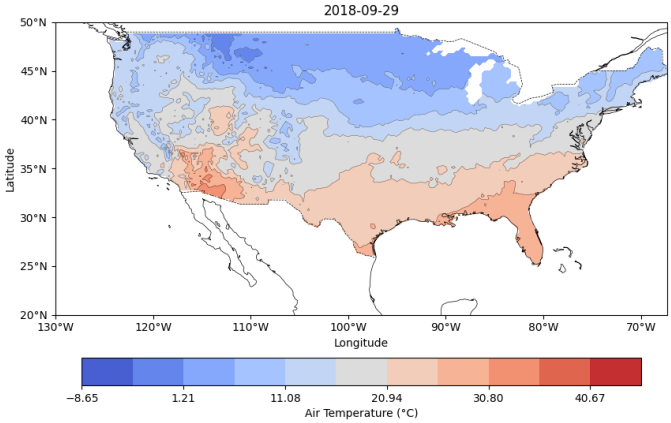

Fig. 56: Contour plot using the contour fill method for September 29, 2018, with the coolwarm colormap

D. Observations and inference

Figures 47 to 57 represent the contour plots generated for the period from July to September 2018. These contour plots have been generated using the contour fill method and coolwarm color map. They reveal notable temperature variations across the United States. A consistent observation is that the southwestern region around the end of July, particularly around California, shows significantly higher temperatures compared to the rest of the country. This temperature anomaly can be attributed to the intense California forest fires that occurred during this period, which led to localized heating. These contour maps effectively highlight the spatial distribution of air temperatures, confirming the warmer-than-average conditions in this area.

Similar inferences were made in Part IA (heatmaps) and Part IB (quiver plots), supporting the conclusion that the southwestern U.S. experienced elevated temperatures likely due to the impact of these forest fires.

II. Information Visualization

A. Network Visualization

Network visualization is a technique used to illustrate relationships between entities in a dataset, often revealing patterns, clusters, or community structures. For this study, network visualization was applied to analyze interactions and connections on social media, specifically using hyperlink data from Reddit.

*1) **Dataset**:* The dataset used for this visualization was obtained from "SNAP Social Network:Reddit Hyperlink Network" [4], which contains records of hyperlinks between various subreddits. Each record includes information about interactions such as the type and weight of hyperlink connections, timestamps, and attributes of the source and target subreddits. The dataset consisted of 55,863 nodes and 858,490 edges. To handle such large data, filtering techniques were used to reduce it to a manageable size. The nodes represent subreddits and directed edges are present if the source subreddit contains a hyperlink to the target subreddit.

*2) **Data Preparation**:* To prepare the data for network visualization in Gephi[2], several steps were taken to convert and enhance the dataset. Initially, the data was in TSV (Tab-Separated Values) format. This format was converted into CSV (Comma-Separated Values) to ensure compatibility with Gephi, a popular tool for network visualization. Additionally, attributes within the dataset were renamed to match Gephi's expected field names.

New fields were calculated from the raw data to get better insights. For instance, impact score was calculated from the attribute information given in the dataset.

$$\text{Pos\_Imp} = \frac{\text{VP} + \text{LP} + \text{LA} + \text{LS} + \text{LF}}{\text{NW}}$$

$$\text{Neg\_Imp} = \frac{\text{VN} + \text{LN} + \text{LAnx} + \text{LAng} + \text{LSad}}{\text{NW}}$$

$$\text{Imp\_Score} = \text{Pos\_Imp} - \text{Neg\_Imp}$$

- **VP** (**VADER_Pos**): Positive sentiment score from VADER (Valence Aware Dictionary and sentiment Reasoner) analysis.
- **LP** (**LIWC_Posemo**): Positive emotion score from the LIWC (Linguistic Inquiry and Word Count) tool.
- **LA** (**LIWC_Affect**): Overall affect (emotional expression) score from LIWC.
- **LS** (**LIWC_Social**): Social orientation score from LIWC, representing social word usage.
- **LF** (**LIWC_Friends**): Friendship-related word score from LIWC, indicating references to friends.
- **VN** (**VADER_Neg**): Negative sentiment score from VADER analysis.
- **LN** (**LIWC_Negemo**): Negative emotion score from LIWC.
- **LAnx** (**LIWC_Anx**): Anxiety-related word score from LIWC, reflecting expressions of anxiety.
- **LAng** (**LIWC_Anger**): Anger-related word score from LIWC, indicating expressions of anger.
- **LSad** (**LIWC_Sad**): Sadness-related word score from LIWC, representing expressions of sadness.
- **NW** (**Num_Words**): Total number of words in the text or document, used to normalize the scores.
- **Pos_Imp** (**Positive_Impact_Score**): Composite score representing positive impact, calculated as the average of positive, social, and friendship-related expressions.
- **Neg_Imp** (**Negative_Impact_Score**): Composite score representing negative impact, calculated as the average of negative emotions, anxiety, anger, and sadness.
- **Imp_Score** (**Impact_Score**): Net impact score, calculated by subtracting the Negative Impact Score from the Positive Impact Score.

*3) **Algorithms used**:*

- **Fruchterman-Reingold Algorithm:** A force-directed layout algorithm where nodes repel each other and edges act as springs to attract connected nodes. It balances attractive and repulsive forces to create an organized layout, highlighting clusters and reducing edge crossings.
- **Force Atlas 2:** An optimized force-directed algorithm for large networks. It emphasizes repulsion to avoid node overlap and uses adaptive repulsion based on node density, making it efficient for complex, large-scale networks.
- **Radial Axis Layout:** This layout arranges nodes in concentric circles around a central node. Nodes are placed based on their hierarchical level or distance from the center, making it suitable for visualizing radial or hierarchical structures.
- **Yifan Hu Layout:** A hybrid layout combining force-directed and multi-level approaches, designed for speed and scalability. It combines repulsive and attractive forces with a hierarchical approximation, producing clear and visually pleasing layouts for large networks.
- **Isometric Layout** The isometric layout is a visualization technique that arranges nodes in a manner that preserves the relative distances between them, typically creating a 2D representation of a higher-dimensional graph. This layout is particularly useful for ensuring that the structure of the graph is maintained while facilitating easy interpretation of node relationships.

*4) **Observations and inferences**:* Figure 57 depicts 3 major communities which were detected by modularity classes [5]. The orange community represents gaming and technology related subreddits, the gray community represents sports related subreddits like NBA, Hockey etc, the green community represents a general category of subreddits like *askreddit, iama* etc. The size of the nodes is scaled according to their Betweenness Centrality[1]. Nodes representing subreddits like *iama, askreddit, subredditdrama* are larger as they influence the flow of information within their respective communities

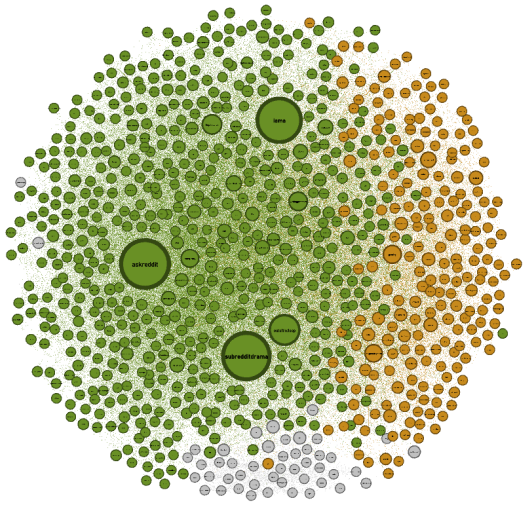

Fig. 57: The figure depicts detected communities of the subreddits with the layout given by Force Atlas 2 algorithm. The size of the nodes represent the betweenness centrality score and the color of the nodes represent the community it belongs to.

and across the network. This indicates their crucial role as intermediaries, facilitating connections between disparate groups and serving as hubs of activity. The same figure plotted with Fruchterman Reingold layout is shown in Figure 58. Unlike in figure 57, the communities aren't seen with clarity, with the sports community not being visible.

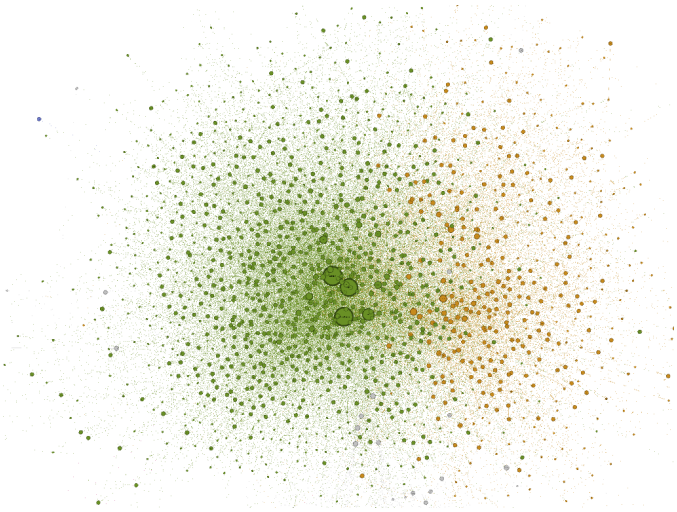

Fig. 58: The Node link diagram depicts the detected communities of the subreddits with the layout given by the Fruchtermann Reingold algorithm. The gray community(sports community) is no longer visible.

Figure 59 provides valuable insights into the impact of various subreddits on the broader community. The impact score, calculated as outlined in Section II-A2, is represented by node colors on a diverging color map ranging from red to blue.

An edge between the subreddits represents a relationship marked by negative sentiment, indicating that interactions or references between these subreddits often involve criticism, disagreement, or conflict. Isometric Layout was used to achieve evenly spaced distribution of nodes, in contrast to figure 60, which presents a more cluttered view. Notably, the number of subreddits with a positive impact score is relatively low.

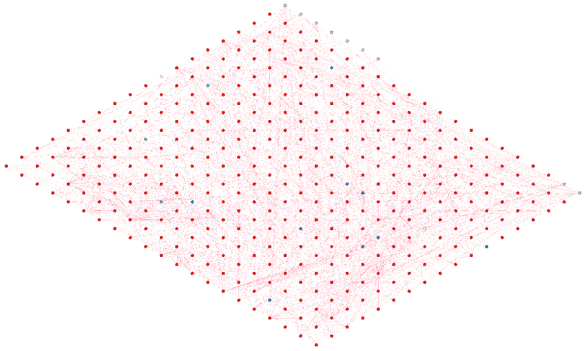

Fig. 59: The Node link diagram represents the impact of the subreddits on the community. The nodes are colored based on a diverging color map (red to blue).

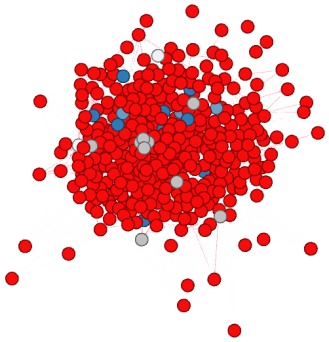

Fig. 60: The Node link diagram represents the impact of the subreddits on the community using the Yifan-Hu layout algorithm.

Figure 61 portrays a zoomed in version of figure 59. It focuses on the *dataisbeautiful* subreddit which has a positive score. The nodes that aren't linked to the subreddit are suppressed. The majority of its links to the subreddits having

negative impact are of negative sentiment (signified in red), while the link to the *math* subreddit is of positive sentiment.

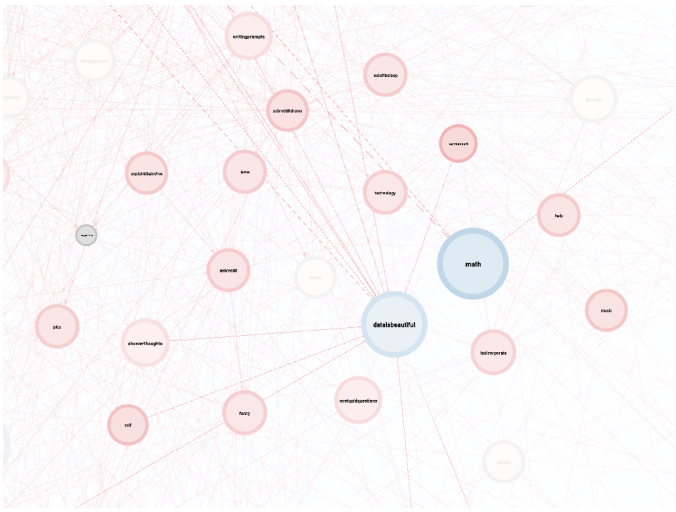

Fig. 61: A zoomed in version of Figure 42

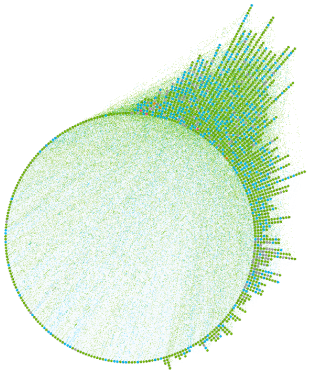

Fig. 62: Radial Axis layout, grouped and ordered clockwise based on the degree of the nodes. The color of the node represents the community it belongs to.

Figure 62 displays the degree distribution in the graph in a sequential, clockwise arrangement. Nodes are positioned in increasing order of degree, indicating that the majority of nodes have a lower degree. The nodes that are grouped along the same angle about the center have the same degree. This distribution suggests that most subreddits tend to maintain a limited number of hyperlinks to other subreddits, typically ranging from 30 to 100 with the majority of them belonging to the general community(depicted in green). The nodes here are filtered to be having a degree of at least 30.

B. Parallel Coordinate Plot

1) ***Dataset****:* The dataset used was the Oxford Covid-19 Government Response Tracker (OxCGRT) [6] which is a comprehensive dataset that recorded how governments around the world responded to the Covid-19 pandemic from 2020 to 2022. The data tracks various policy measures, including containment actions like school closures and travel restrictions, as well as economic and health system responses. It aggregates this information into indices that reflect the strictness and extent of government actions. Developed by the Blavatnik School of Government at Oxford University, the dataset provides valuable insights for researchers and policymakers to analyze and prepare for future health emergencies.

2) ***Data Preparation****:* This section outlines the data preparation processes undertaken to visualize COVID-19 containment policies and case counts through parallel coordinate plots (PCPs). The objective was to structure the datasets to facilitate comparative analysis across selected countries and to derive meaningful insights from the visualizations.

For the first set of PCPs(see Figure 63- 65), which focuses on the containment policies implemented by various countries, the dataset was refined to include the key measure of School Closing Index. The date column was converted into a month-year format, allowing for a clear aggregation of policy measures over time. The dataset was then grouped by country and month-year, enabling the calculation of average values for the specified containment measures. Selected countries were filtered based on their response measures, and the data was sorted accordingly to facilitate comprehensive analysis.

In the preparation of the second set of PCPs(see Figure 66-67), which visualizes the progression of confirmed COVID-19 cases across selected countries, the dataset was filtered to include the top 20 countries with significant case counts. The date column was similarly transformed into a monthly format to capture the number of confirmed cases at the end of each month. To address missing values in the case data, interpolation methods were applied, ensuring a continuous dataset throughout the specified period. This structured approach provided a robust foundation for comparative analysis and effective visualization of the COVID-19 case counts across the selected nations over time.

3) ***Plotting and Interactability****:* This section discusses the plotting techniques and interactive features implemented in the visualizations of COVID-19 containment policies and case counts. By utilizing Plotly.js and D3.js libraries, we were able to create dynamic and interactive plots that facilitate user engagement and exploration of the data. The interactive nature of the plots allows for brushing and selection, enhancing the analytical experience by enabling users to focus on specific countries or time periods.

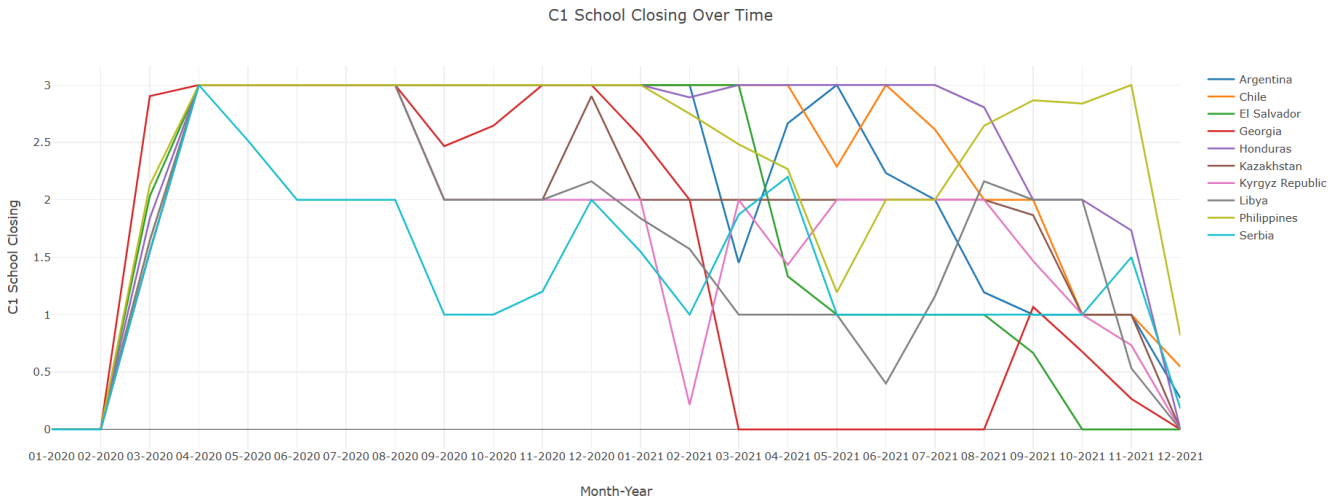

Fig. 63: The figure is a PCP that plots the average C1 School Closing index for a specific month and a specific country.

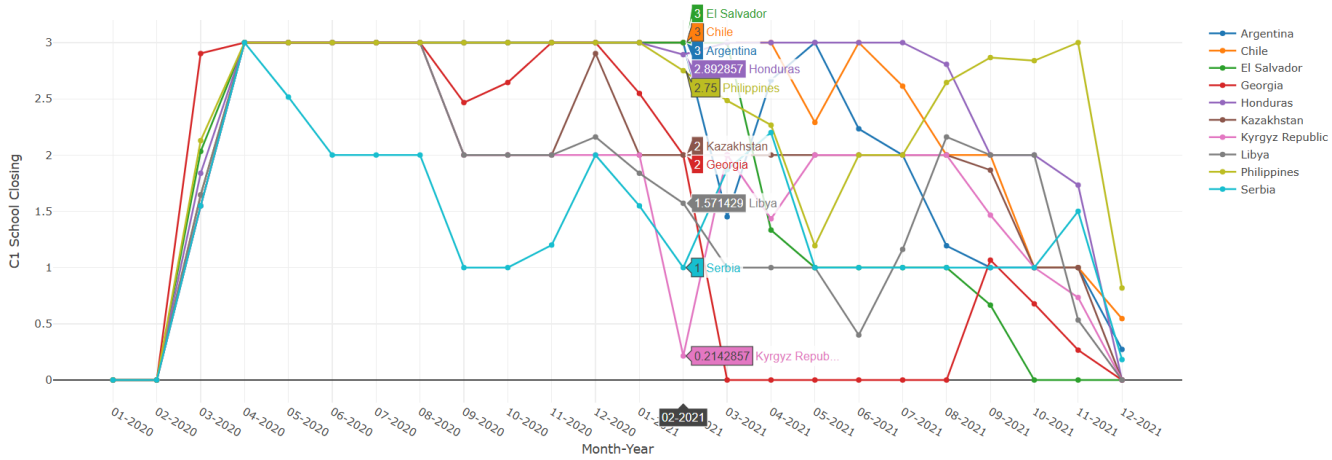

Fig. 64: The figure is a PCP that shows the average C1 School Closing value for each country when hovering the cursor over a specific month.

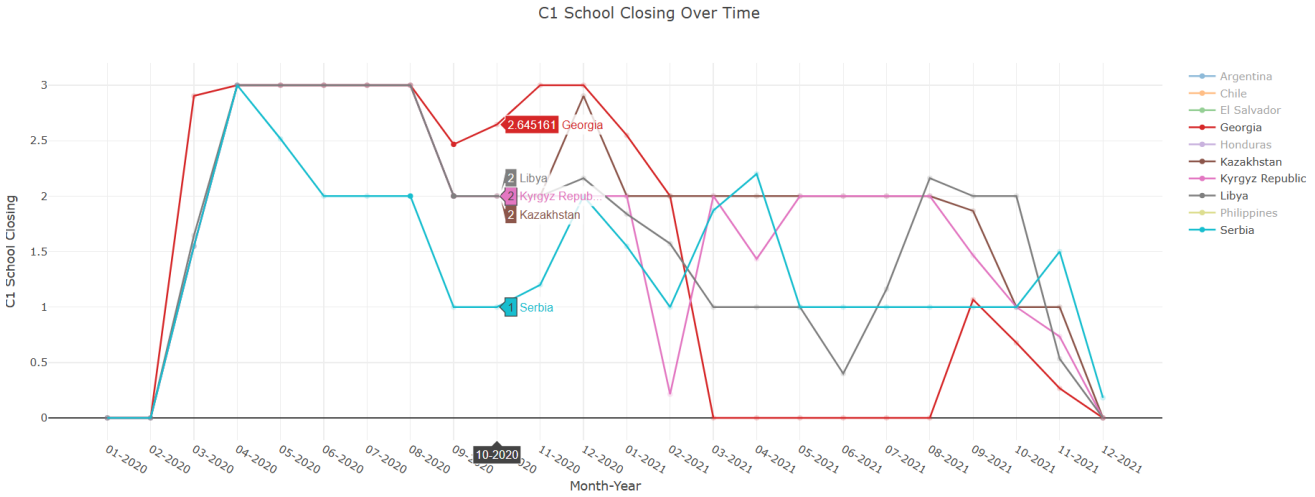

Fig. 65: The figure is a PCP that shows the average C1 School Closing index for only selected countries when selected (brushing) with the cursor.

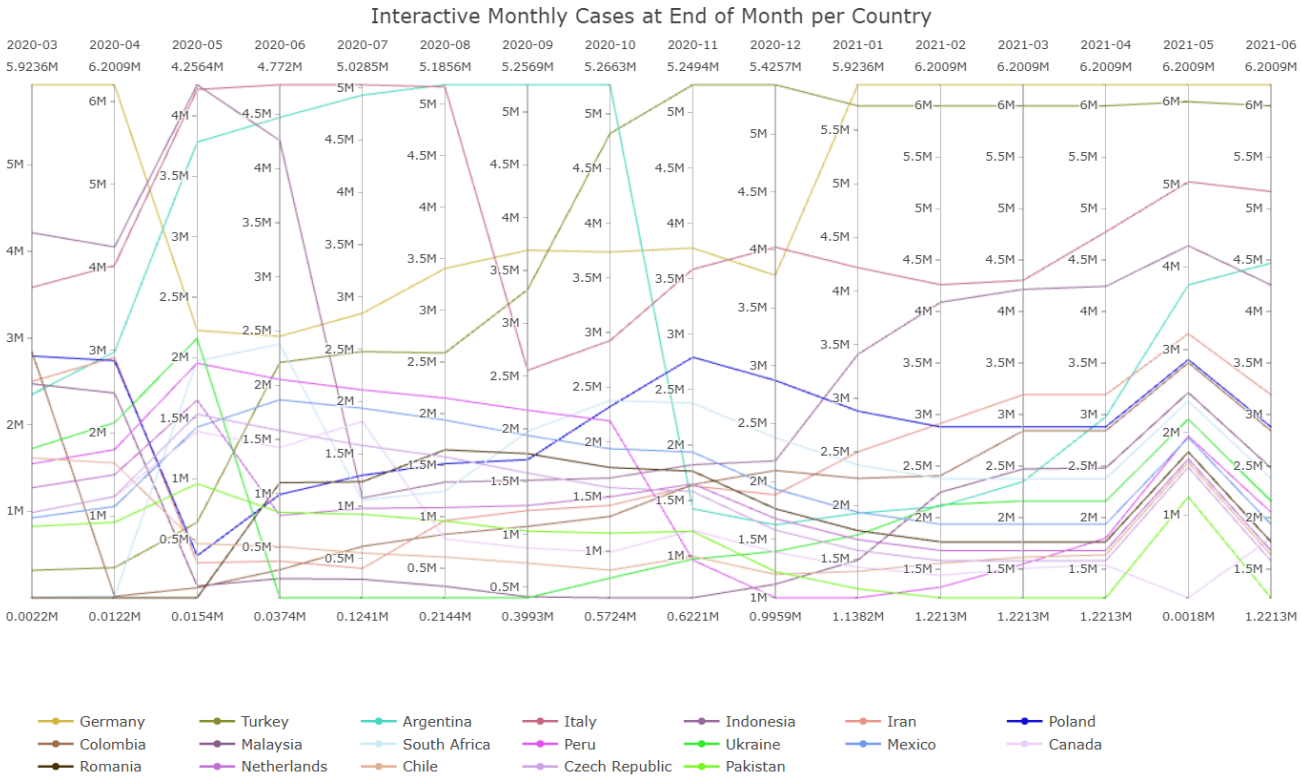

Fig. 66: The figure is a PCP that plots the cases recorded in a specific month for a specific country. 'M' indicates that values of cases are in the order of 10^6

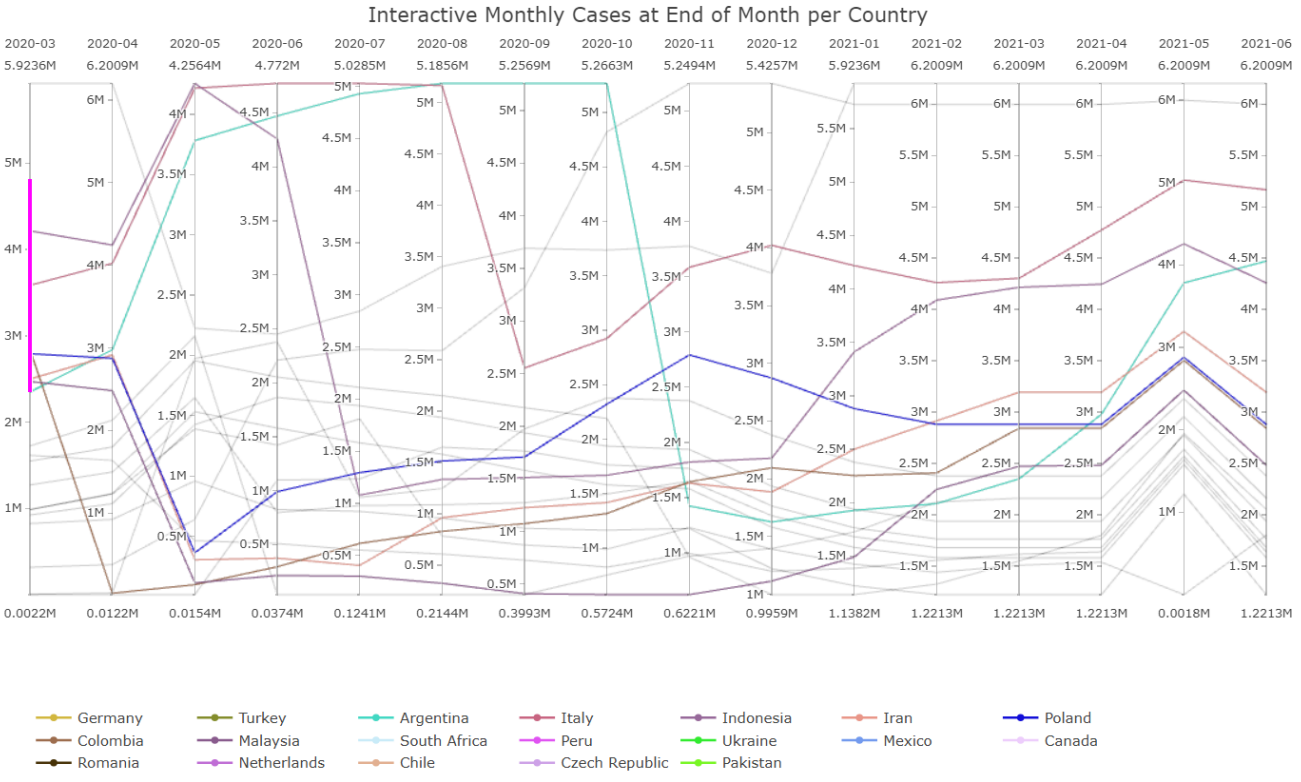

Fig. 67: The figure is a PCP that shows the cases recorded in a specific month for selected countries(using brushing). 'M' indicates that values of cases are in the order of 10^6

The first set of PCP plots(see Figure 63- 65) visualizes the C1 School Closing policies over time, enabling users to track the evolution of school closure measures implemented by different countries. By leveraging D3.js for data loading and manipulation, the script extracts unique countries and months from the dataset, creating traces for each country with corresponding C1 values. The Plotly library facilitates the creation of a line plot with markers, enhancing visibility and interactivity. Users can select specific countries through a brushing mechanism, which filters the displayed data to focus on the selected countries, thereby allowing for a clearer comparative analysis of policy responses.

In the second set of plots(see Figure 66- 67), which illustrates the interactive monthly confirmed COVID-19 cases across selected countries, we employed parallel coordinate plotting techniques to visualize the data. The dimensions of the plot correspond to different months, while the lines represent confirmed case counts for each country. A color scale is applied to enhance the distinction between countries, and the inclusion of a legend further supports the identification of data series. The plot is designed for user engagement, with tool tips providing additional information on hover, thus enhancing the ability to interpret the case count data over the specified period. Together, these visualizations not only display critical information but also invite users to interact with the data in meaningful ways.

*4) **Inference and Observation**:* For the first set of PCPs(see Figure 63- 65)

- There is a noticeable increase in the School Closing Policy index in February 2020. This indicates that the countries depicted in the plot implemented swift and immediate responses to the initial wave of COVID-19 by closing schools.
- After the initial months, significant variation in the School Closing Policy index can be observed. For instance, some countries, like Honduras, maintained consistently high index values throughout 2020 and a large part of 2021, with a decline only after July 2021. On the other hand, countries like Serbia experienced fluctuations in the school closing index, alternating between high and low values.

For the second set of PCPs(see Figure 66- 67)

- Throughout 2020 and a significant portion of 2021, a general upward trend can be observed in the number of COVID-19 cases across all the countries represented in the dataset. This trend reflects the global spread of the virus and its sustained transmission during this period.
- Some countries, such as Colombia, experienced a steady and gradual increase in the number of reported cases over time, with no dramatic spikes or sudden shifts, indicating a more controlled or consistent transmission rate.
- In contrast, certain countries like Argentina exhibited significant fluctuations in their case numbers. For instance, Argentina saw a sharp spike in cases around May 2020, reaching approximately 5 million cases. Following this peak, the number of cases remained relatively stable until October 2020, before declining through December 2020, reaching around 1.5 million cases. However, the trend reversed again in mid-2021, with cases rising sharply to approximately 4.5 million by June 2021.

C. Treemap Plot

A treemap plot is often used to visualize hierarchical data using nested rectangles. This visualization technique enables easy comparison of proportions within categories, making it highly suitable for displaying large amounts of data where different items can be grouped by common attributes. In this study treemaps were used to visualize both hierarchical and non-hierarchical data.

*1) **Dataset and Data Preparation**:* The dataset used for this study is a subset of the dataset used in the A1 assignment - Oxford Covid-19 Government Response Tracker (OxCGRT) Dataset. Data cleaning was done keeping in mind the amount of data that can be represented using a treemap without causing any visual clutter. For this, certain rows were carefully chosen and excluded from the dataset. To explore both hierarchical and non-hierarchical relationships in COVID-19 data, two distinct subsets were employed:

- **Hierarchical Data:** For U.S.-based analysis, COVID-19 cases were organized region-wise by state and analyzed across multiple months.
- **Non-Hierarchical Data:** For a broader, country-level view, only confirmed case counts by country were considered.

*2) **Implemetation**:* The implementation of the COVID-19 cases treemap uses HTML, FusionCharts, and JavaScript libraries to create interactive visualizations. HTML was used to structure the page and provide container to the chart. FusionCharts was used for the creation of interactive treemaps, allowing for dynamic visualization fo the COVID-19 data with both hierarchical and non-hierarchical structures. Javascript managed the data processing and chart rendering while PapaParse enabled CSV parsing, extracting relevant data for visualization from the dataset.

*3) **User Interaction**:* The treemap visualization was designed to represent COVID-19 cases distribution. In both hierarchical and non-hierarchical views, the size of each rectangular area corresponds to the number of COVID-19 cases—larger rectangles indicate higher case counts. Additionally, colour is used as a visual channel, with darker color representing higher number of cases. The intuitive combination of both size and color makes it easier to compare case levels across regions.

For enhanced interactivity, the visualization includes a hover feature. When users hover over a region, a tooltip appears displaying details such as month and the specific number of cases, as shown in figures 68, 69. In the hierarchical treemap, the users can click on a region to view the next level of data. This interaction was chosen to enable users to view multiple

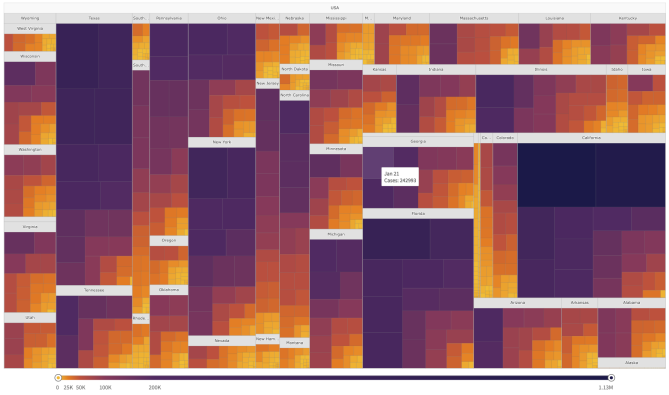

Fig. 68: Image of Treemap plot with squarified layout for hierarchical data, particularly showing first level of the hierarchy and demonstrating user interaction such as hover and click.

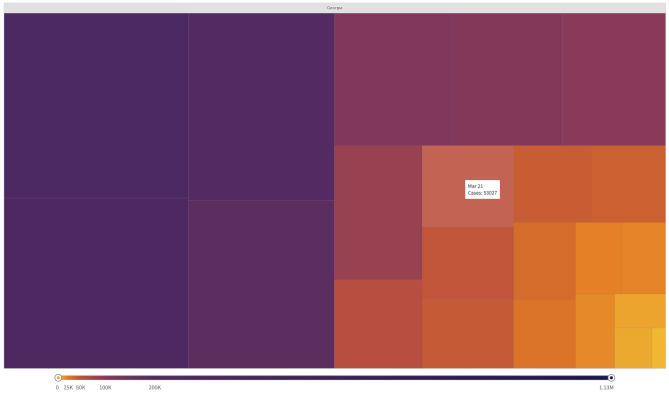

Fig. 69: Image of a Treemap plot with squarified layout for non-hierarchical data showing the second level of the hierarchy, expanded from Fig. 69

levels of data without overwhelming the user. This feature is shown in all figures that involve visualization for hierarchical data, namely figures 69, 76, 78 and 80.

To further enhance the user experience, a slider has been added below the treemap, as demonstrated in the figure 70. The slider acts as a filter allowing users to set a minimum and maximum thresholds for case counts. As the slider is adjusted, countries/regions with cases below or above the respective thresholds are grayed out. It helps users focus on areas with specific case counts.

D. Plots

The figures have been categorized into treemaps plotted for hierarchical data and non-hierarchical data. Figures 68-70, 75-80 represent treemaps for hierarchical data while figures 71-74 represent treemaps for non-hierarchical data. For both hierarchical and non-hierarchical data, 4 layouts have been used, namely squarified, slice and dice - horizontal, vertical and alternate.

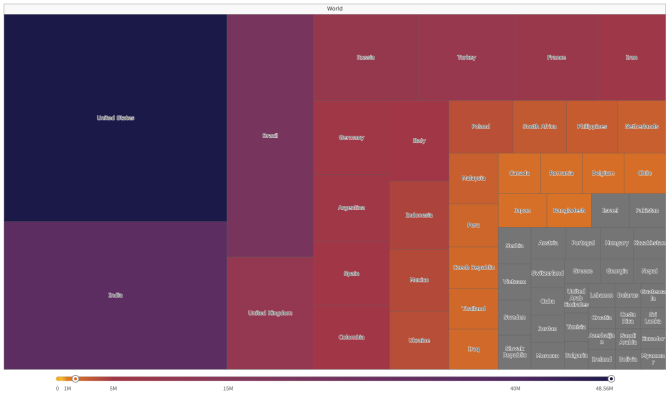

Fig. 70: Image of Treemap plot with squarified layout for hierarchical data to demonstrate the functionality of the slider.

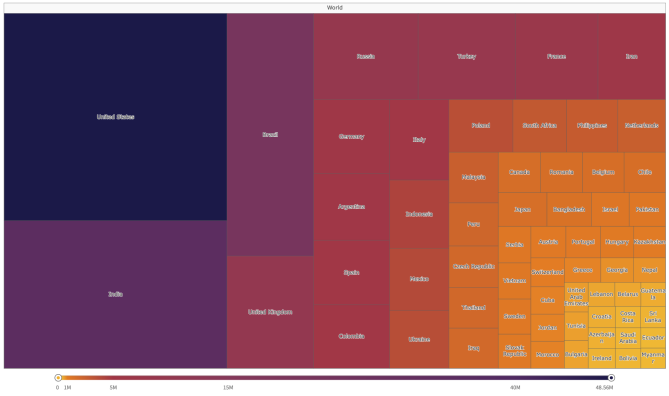

Fig. 71: Image of Treemap Plot for non-hierarchical data using squarified layout.

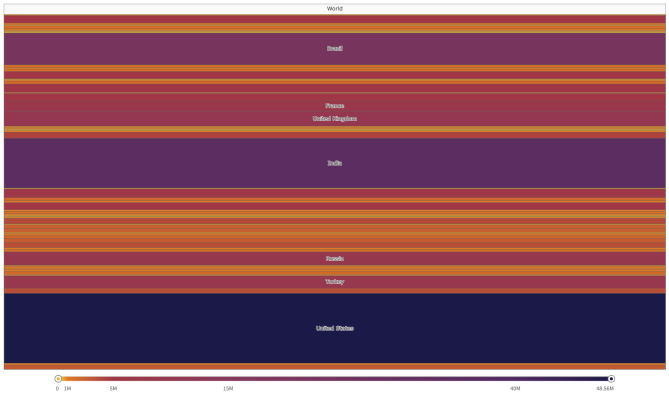

Fig. 72: Image of Treemap Plot for non-hierarchical data using slice and dice - horizontal layout.

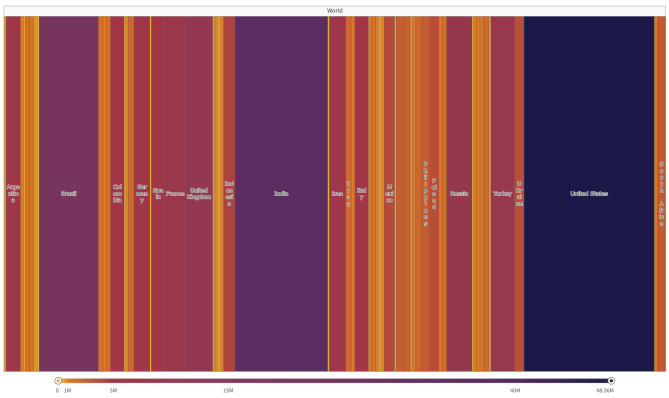

Fig. 73: Image of Treemap Plot for non-hierarchical data using slice and dice - vertical layout.

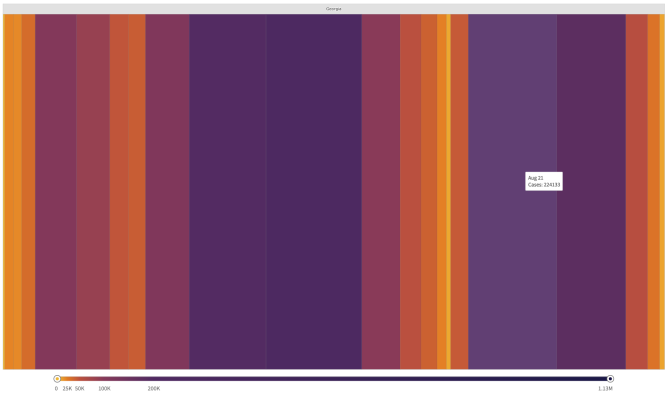

Fig. 76: Image of second level of Treemap Plot for hierarchical data using slice and dice - horizontal layout, expanded from Fig. 76.

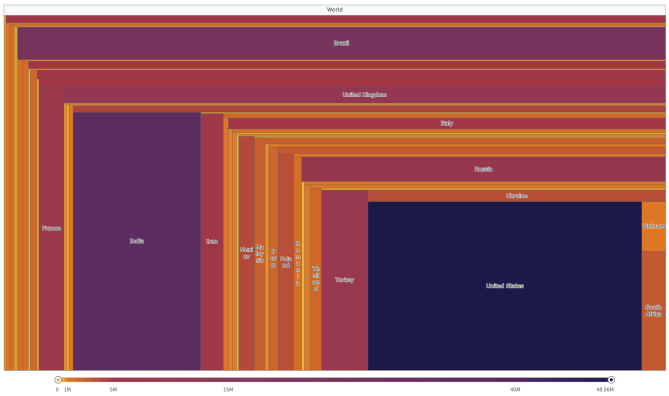

Fig. 74: Image of Treemap Plot for non-hierarchical data using slice and dice - alternate layout.

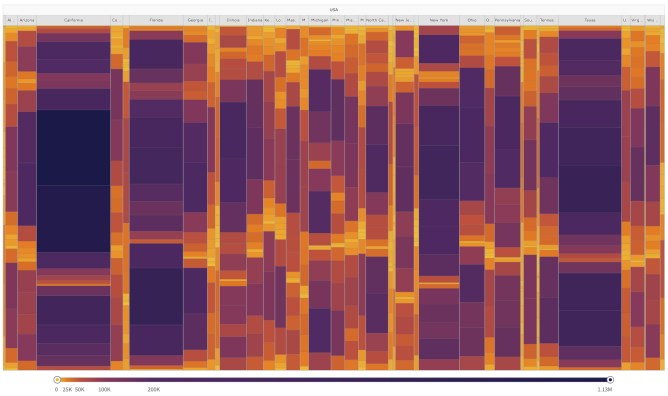

Fig. 77: Image of first level of Treemap Plot for hierarchical data using slice and dice - vertical layout.

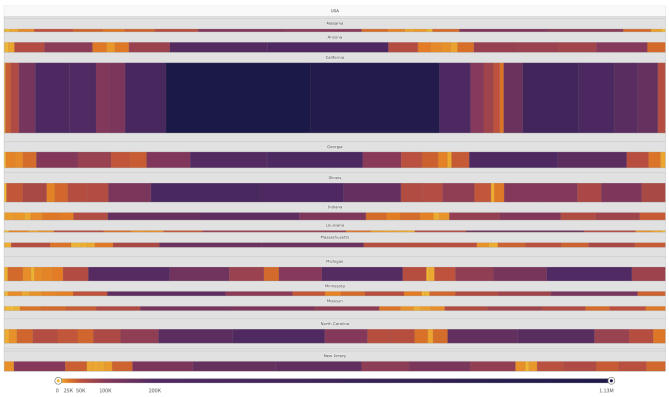

Fig. 75: Image of the first level of Treemap Plot for hierarchical data using slice and dice - horizontal layout.

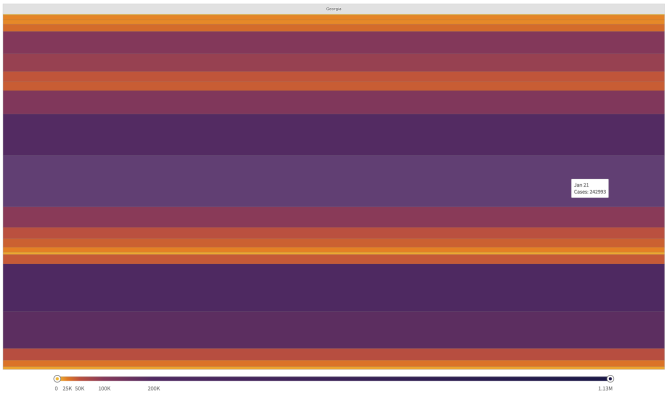

Fig. 78: Image of second level of Treemap Plot for hierarchical data using slice and dice - vertical layout, expanded from Fig. 78

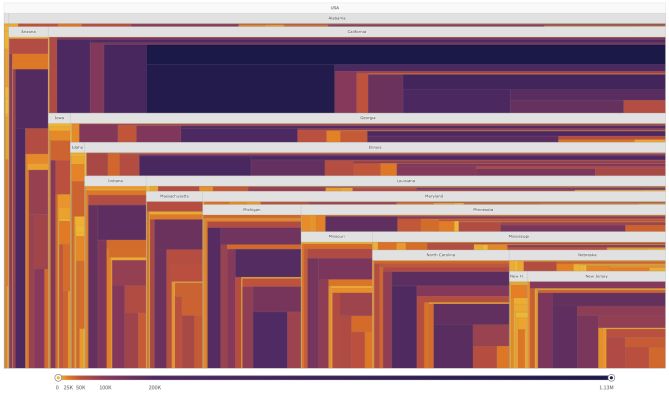

Fig. 79: Image of first level of Treemap Plot for hierarchical data using slice and dice - alternate layout.

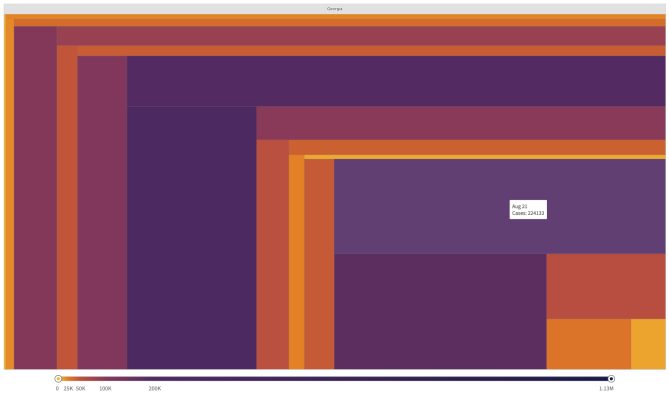

Fig. 80: Image of second laevel of Treemap Plot for hierarchical data using slice and dice - alternate layout, expanded from Fig. 80

E. Observations and Inferences

1) ***Hierarchical Data (Region-wise and Month-wise Analysis for the USA)****:* The hierarchical treemap visualization allowed for a more elaborate and deeper examination fo the COVID-19 cases across different regions in the United States and by month, thereby giving us more insights in regional spread and seasonal fluctuations. Larger areas in the treemap, representing regions with higher case counts, were observed mainly in regions with more urban density. The month wise data within each region demonstrated notable spikes corresponding to the different waves of COVID-19 observed during 2020-2021.

The hierarchical view provides insight into both spatial and temporal trends, helping users understand more about the pandemic's progression across the United States. The clustering of high cases near urban centers and information of peak periods can help inform health authorities for planning and resource allocation.

2) ***Non-Hierarchical Data (Country-wise Analysis)****:* In the non-hierarchical treemap of COVID-19 cases by country, it was observed that majority of the cases were contributed by a few countries, with significant clusters in countries with high populations. Countries with darker colors and larger sizes, such as the United States, India, and Brazil stand out prominently due to the large number of confirmed cases. This non-hierarchical visualization effectively highlights the disparities in COVID-19 impact across countries.

The use of slider control enabled a clearer view of countries that fall within specific thresholds. This particularly helped study either only high-impact or low-impact countries. The non-hierarchical perspective is useful for identifying global patterns and can help international organizations channel their efforts efficiently.

Both hierarchical and non-hierarchical observations reveal distinct aspects of COVID-19's spread, showing the value of varied treemap visualizations for layered data analysis.

REFERENCES

[1] *Betweeness Centrality*. URL: https://en.wikipedia.org/wiki/Betweenness_centrality.

[2] *Gephi: The Open Graph Viz Platform*. URL: https://gephi.org/features/.

[3] Climatology Lab. *Gridded Meteorological Data for the United States*. https://www.climatologylab.org/gridmet.html. 2023.

[4] Jure Leskovec and Andrej Krevl. *SNAP Datasets: Stanford Large Network Dataset Collection*. http://snap.stanford.edu/data/soc-RedditHyperlinks.html.

[5] *Modularity classes*. URL: https://parklize.blogspot.com/2014/12/gephi-clustering-layout-by-modularity.html.

[6] *OxCGRT: Oxford COVID 19 Dataset*. URL: https://www.kaggle.com/datasets/ruchi798/oxford-covid19-government-response-tracker.

[7] *Pandas: Python Data Analysis Library*. URL: https://pandas.pydata.org/.

[8] O. Rios, W. Jahn, and G. Rein. "Forecasting wind-driven wildfires using an inverse modelling approach". In: *Natural Hazards and Earth System Sciences* 14.6 (2014), pp. 1491–1503. DOI: 10.5194/nhess-14-1491-2014. URL: https://nhess.copernicus.org/articles/14/1491/2014/.

[9] P. H. THOMAS. "Rates of Spread of Some Wind-driven Fires". In: *Forestry: An International Journal of Forest Research* 44.2 (Oct. 1971), pp. 155–175. ISSN: 0015-752X. DOI: 10.1093/forestry/44.2.155. eprint: https://academic.oup.com/forestry/article-pdf/44/2/155/6742840/44-2-155.pdf. URL: https://doi.org/10.1093/forestry/44.2.155.

[10] Wikipedia. *Ferguson Fire — Wikipedia*. URL: https://en.wikipedia.org/wiki/Ferguson_Fire.

[11] Wikipedia. *Pawnee Fire — Wikipedia*. URL: https://en.wikipedia.org/wiki/Pawnee_Fire.

[12] Wikipedia. *Spring Creek Fire — Wikipedia*. URL: https://en.wikipedia.org/wiki/Spring_Creek_Fire_(2018).

[13] *Xarray Documentation*. URL: https://docs.xarray.dev/en/stable/index.html.