EPA Air Quality Consequent to Forest Fire Consumption of Townships

Project Proposal for Data Visualization Analysis

Nathan S. Hicks

University of Indiana

Certificate in Data Visualization

October 18, 2020

EPA Air Quality Consequent to Forest Fire Consumption of Townships

Nathan Hicks 10/18/20

Stakeholder Insight Needs

After reviewing your prior work, provide a revised statement that describes your visual analytics problem and the stakeholders that would use your work in a decision-making process. Identify the most important visual analytics insight needs and discuss why this project is important.

The visual analytics project proposed here is intended to give visibility to recent forest fire activity and consequent air quality pollutants that may have negatively impacted the health and safety of humans and other living creatures in unknown quantities and pollutant varieties. Of particular interest to this analysis are the actual consumption of townships by various forest fires primarily along the western coastal states of the United States of America. Air quality in numerous locations quickly became hazardous for many groups of individuals that lasted in some instances for months.

It is said that the rapid onset of forest fire activity during the month of September 2020 was resultant from natural cause (e.g. lightning or similar). A portion of this project proposal then, is to evaluate publicly available data that classifies forest fire causes, whether they be human, natural, or of a similar category. Furthermore, some forest fires indeed had consumed entire townships. In itself, this is an alarming outcome. It is conceivable then that the pollutants produced from a typical forest fire could vary greatly from the pollutants deployed with the consumption of said entire townships. Thus, another important visualization outcome is to understand geospatial distributions of the various Environmental Protection Agency (EPA) air quality parameters, to determine any variance from what might be expected.

Visualizing the various EPA parameters within the identified data can primarily help 2 stakeholder groups: 1) members of government, and 2) private individuals. Generally speaking, both stakeholder groups share insight needs, with an emphasis on geospatial identification of air quality pollutants. For example, knowing the constituents of air quality at a specified location and if there are specific toxins resultant from townships burning, may be very important to 1) protecting one’s own health, and 2) as from the perspective of government, to know if a certain population of individuals is in danger. In the latter instance, knowing where and when to direct emergency resources can thus be leveraged.

EPA Air Quality Consequent to Forest Fire Consumption of Townships

Nathan Hicks 10/18/20

Data Acquisition

Provide a short description of the data set(s) you plan to use in order to address your visual analytics problem and provide simple descriptive statistics of the data (e.g., number of records, a simple analysis of important data fields).

This project proposes to initially utilize 2 primary datasets for data visualization analysis. The prescribed datasets are publicly available from the following websites, that each permit data retrieval using an application program interface (API):

- National Interagency Fire Center (NIFC)
  https://data-nifc.opendata.arcgis.com/datasets/incident-3/data?geometry=-152.132%2C28.047%2C-47.982%2C51.327&orderBy=FireDiscoveryDateTime&orderByAsc=false
- United States Environmental Protection Agency (EPA)
  https://docs.airnowapi.org

The currently to-date NIFC data provides 404 rows and 97 columns of raw data representing recent or ongoing forest fire activity within the borders of the United States. Table 1 below highlights the more important of the attributes intended for this visualization analysis.

Table 1 – Highlighted attribute descriptions of the National Interagency Fire Center Dataset.

NIFC Data Attributes	Attribute Description
CalculatedAcres	calculation of how many acres are burned since fire discovery
ContainmentDateTime	date/time of when the fire is contained
DailyAcres	how many burned acres result each day; acres per day
FireCause	cause of the fire
FireCauseGeneral	cause of the fire; general cause
FireCauseSpecific	cause of the fire; specific cause
FireCode	specific code assigned to the fire
FireDiscoveryDateTime	date/time of the fire discovery
FireOutDateTime	date/time of the extinguished fire
IncidentName	the name of the fire event
IncidentShortDescription	a short description of the fire event
IncidentTypeCategory	category assigned to the fire event
IncidentTypeKind	the specific category of the fire event
InitialLatitude	latitude of the fire event
InitialLongitude	longitude of the fire event
IsTrespass	flag if the fire event relates to trespassing activity; Boolean
IsValid	flag if the fire event is a valid record
POOCity	US city of the originating fire event
POOCounty	US county of the originating fire event
POOJurisdictionalAgency	the US agency jurisdiction of the fire event
POOLandownerCategory	the category label of the land owner relating to the fire event

POOLandownerKind	the specific category label of the land owner of the fire event
POOState	US state of the fire event
PredominantFuelGroup	indication of the underlying fuel used to maintain the fire event
x	additional coordinate values; x-axis
y	additional coordinate values; y-axis

The currently to-date EPA data provides 1,042,144 rows and 29 columns of raw data representing various air quality sensor measurements and related attributes for the US state of Washington only, spanning the year of 2020 to-date. Table 2 below highlights the more important of the attributes intended for this visualization analysis.

Table 2 – Highlighted attribute descriptions of the National Interagency Fire Center Dataset.

EPA Data Attributes	Attribute Description
county	the US county in which the sensor station is located
date_gmt	the Universal Coordinate Time 'date' of the sensor data captured
latitude	latitude of the sensor station taking samples
longitude	longitude of the sensor station taking samples
method	the method used to obtain the sensor data sample
method_type	the specific method used to obtain the sensor data sample
parameter	the name of the sensor attribute captured; the air quality of interest
qualifier	how to qualify the magnitude sampled
sample_duration	how long of a duration the sensor data was sampled
sample_frequency	how frequently the sensor data is sampled
sample_measurement	the magnitude of the sensor data sampled; the actual value of interest
state	the US state in which the sensor station is located
time_gmt	the Universal Coordinate Time 'time' of the sensor data captured

Analysis and Visualization

Describe potential data analysis and visualization methods that you have or may apply to the data that you identified for the project. Make sure to discuss any related work that you have found and used in developing your visual analytics workflow, which should be *cited as a reference* at the end of the report.

Provide a set of prototype visualizations (either a hand-drawn sketch that you envision or computer mock-ups from data) and describe the alignment of **graphic symbols and graphic variables** used in each visualization to the analysis and data sets used in the project.

The necessary datasets are acquired via API (see above) and manipulated by way of the Project Jupyter Jupyter Notebook (https://jupyter.org/) tool. Python is the programming language used therein (https://www.python.org). Only the most basic results were derived of the datasets, such that the project constructs could begin to take form. Specifically, the attributes previously given (above) are perceived to contain optimal data for initial visualization analysis, such the project goals are achieved.

Aside of the aforementioned dataset results, multiple sketched visualizations are established for further consideration. Instinctually, one map type to analyze for air pollutants is a geospatial map overlaid with an isoline map depicting pollutant magnitudes. In fact, many categorical types of data within the EPA dataset can be visualized using isolines in this manner, as this dataset contains 1,363 available possible measurement parameters. The EPA employs the following 7 parameters to assess overall air quality, and these are most commonly known to the general public, typically reported as a value of integer magnitude:

Table 3 – EPA established parameters used to assess air quality.

Parameter Code	Parameter Name
88502	Acceptable PM2.5 AQI & Speciation Mass
42101	Carbon monoxide
42602	Nitrogen dioxide (NO2)
44201	Ozone
81102	PM10 Total 0-10um STP
88101	PM2.5 - Local Conditions
42401	Sulfur dioxide

Most certainly, these are important values to evaluate, though more impacting results are expected from using other additional parameters. Most importantly here, is that the parameters shown in Table 3, facilitate the initial visualization results, that are independent of which parameter is chosen, for geospatial maps in any case. For additional consideration then, are the following air quality parameters associated with toxic combustion biproducts of dwelling structures as shown in Table 4 (https://www.hpac.com/association-solutions/article/20929620/what-is-so-dangerous-about-smoke-in-building-fires#:~:text=The%20nitrogen%20and%20sulfur%20compounds,hundred%20known%20toxins%20are%20generated.):

Table 4 – Additional EPA parameters not commonly identified for air quality evaluation (gray cells indicate standard 7 air quality parameters, above).

Parameter Code	Parameter Name
42604	Ammonia
62604	Ammonia (precip)
22604	Ammonia (SP)
42101	Carbon monoxide
42607	Hydrogen cyanide

42402	Hydrogen sulfide
42602	Nitrogen dioxide (NO2)
42603	Oxides of nitrogen (NOx)
42600	Reactive oxides of nitrogen (NOy)
12169	Sulfur (TSP) STP
42401	Sulfur dioxide
85169	Sulfur PM10 LC
82169	Sulfur PM10 STP
86169	Sulfur PM10-2.5 LC
83169	Sulfur PM10-2.5 STP
88169	Sulfur PM2.5 LC
84169	Sulfur PM2.5 STP
42306	Sulfuric acid
12402	Sulfuric acid (TSP) STP
43911	Total reduced sulfur
42269	Total sulfur

The complexity to this approach is that not all air quality reporting stations observe each and every one of these identified parameters (Tables 3, 4), and a reduced volume of data should be the expected outcome here.

With respect to the remaining dataset derived from the NIFC API, regarding forest fire activities, the geospatial locations of each fire event can map geospatially as well. Doing so will help correlate between specific air quality pollutants and forest fire events across the United States. Although the EPA dataset here collected is constrained to the US state of Washington, the API data retrieval scope can expand accordingly depending upon the geospatial location under consideration. Both datasets contain the required geospatial coordinates and date and time stamps. In this way each data set can combine to make more impactful analysis results. Very specifically, the total destruction of various west coast townships due to fire can be evaluated for toxic air particulates that significantly degrade quality of life.

Thus, it is important to identify today's publicly accessible data visualizations, so as to help establish a baseline to measure this project's results. As such, Figure 1 below employs blue circles [geometric symbol – (Area) and (Point)] / [graphic variable – (Retinal-Form-Shape), (Retinal-Color-Hue)] upon a US map to indicate existing forest fire events. Each of the geospatial maps (all Figures 1-3) depicting the US and its states borders, employ graphical lines [(geometric symbol – (Line)] and [graphic variable – (Retinal-Form-Shape)], primarily.

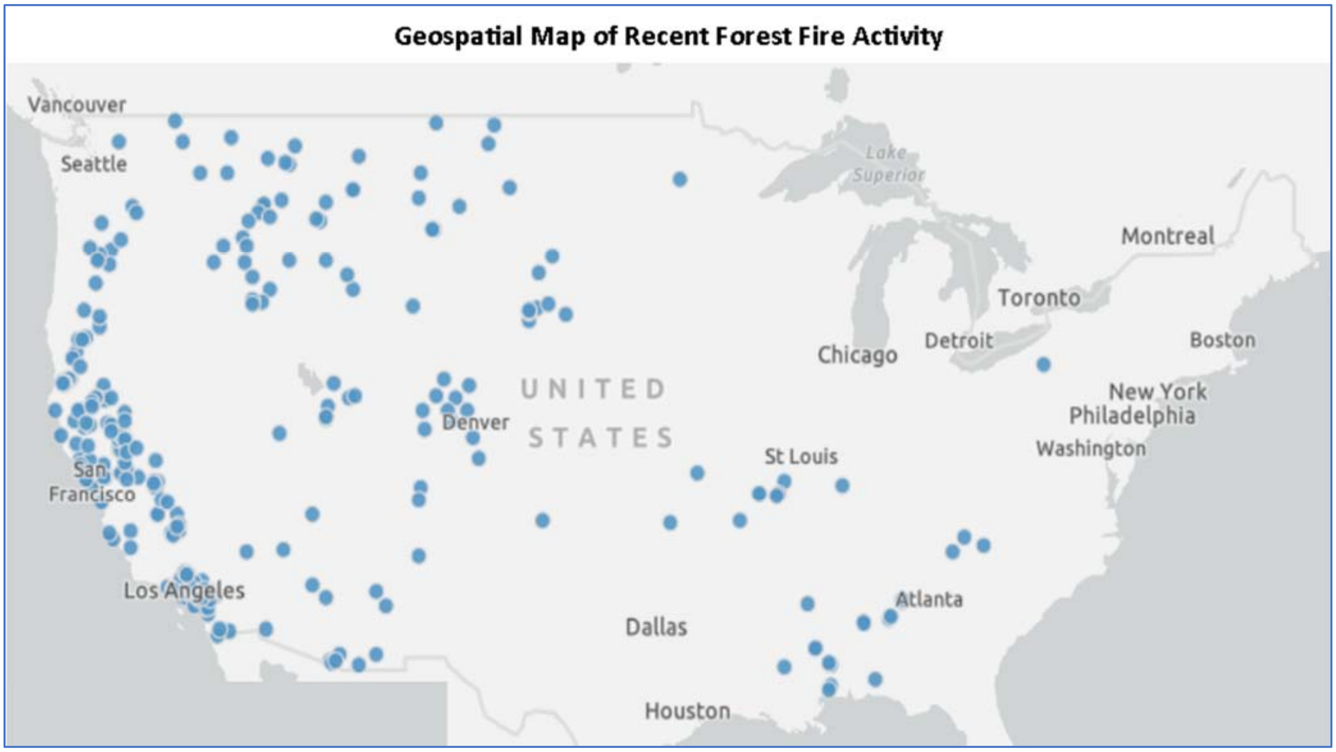

Figure 1 – National Interagency Fire Center geospatial map of existing forest fire activities across the United States.

https://data-nifc.opendata.arcgis.com/datasets/incident-3/data?geometry=-140.786%2C27.127%2C-36.636%2C50.673&orderBy=FireDiscoveryDateTime&orderByAsc=false

For Figure 2, below, the EPA provides a geospatial map with monitoring sensor station locations across the United States. The graphic symbols and graphic variables are very similar to those identified within the forest fire map of Figure 1. One variance of the EPA map below is the use of multi-colored circles to map each sensor station locations [geometric symbol – (Area) and (Point)] / [graphic variable – (Retinal-Form-Shape), (Retinal-Color-Hue)] that matches the air quality health scale (Fig. 2, right side).

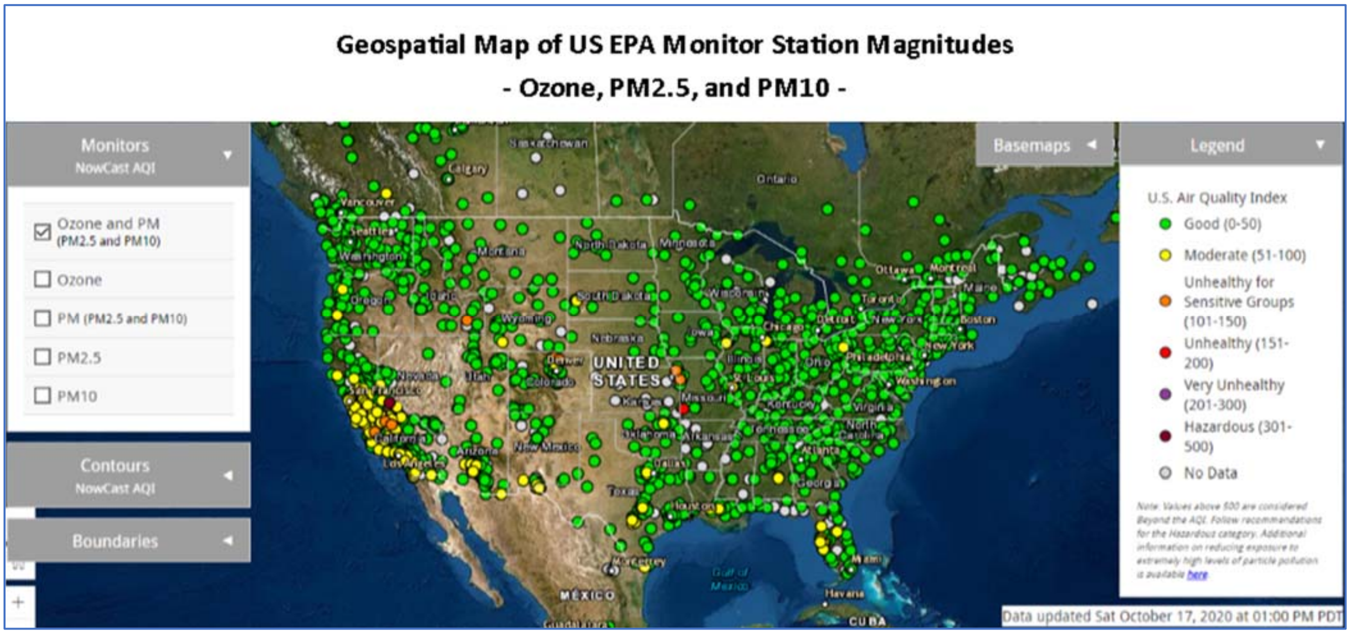

Figure 2 – United States Environmental Protection Agency geospatial map of monitoring sensor stations for the primary types of air pollutants ozone and the PM particulates class. https://gispub.epa.gov/airnow/#

Figure 3 below, is an isoline map of the air quality scores [geometric symbol – (Surface)] / [graphical variable – (Spatial-x, -y, -z), (Retinal-Form-Curvature), (Retinal-Color-Hue)] across the United States.

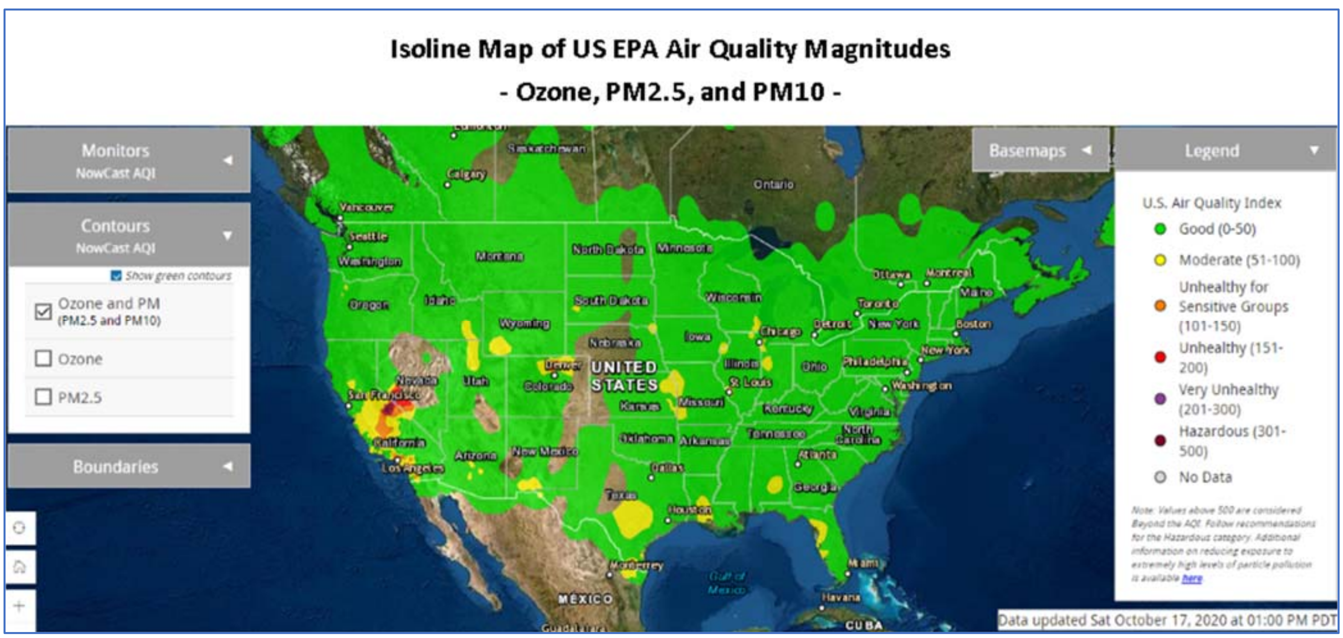

Figure 3 – United States Environmental Protection Agency geospatial map with an isoline map overlay representing air quality pollutant scores ozone and the PM particulate class, used to assess the regional impact of air quality. https://gispub.epa.gov/airnow/#

EPA Air Quality Consequent to Forest Fire Consumption of Townships

Nathan Hicks 10/18/20

For this project, a few simple improvements to existing data visualizations would yield a more granular and correlative analysis between forest fire activity, regional air quality, and the more toxic but less observed air quality pollutant constituents. To implement such improved data visualization upon the forest fire data, as shown in Figure 4, it would improve the visualizations meaning if the circular dots conveyed categorical information. This is achieved as shown, by varying the circle shape fill color [graphic symbol – (Area), (Point)] / [graphic variable – (Retinal-Form-Shape), (Retinal-Color-Hue)]. An additional variable that could be depicted here, would relate to a quantitative variable, and the circle diameter could relate therein.

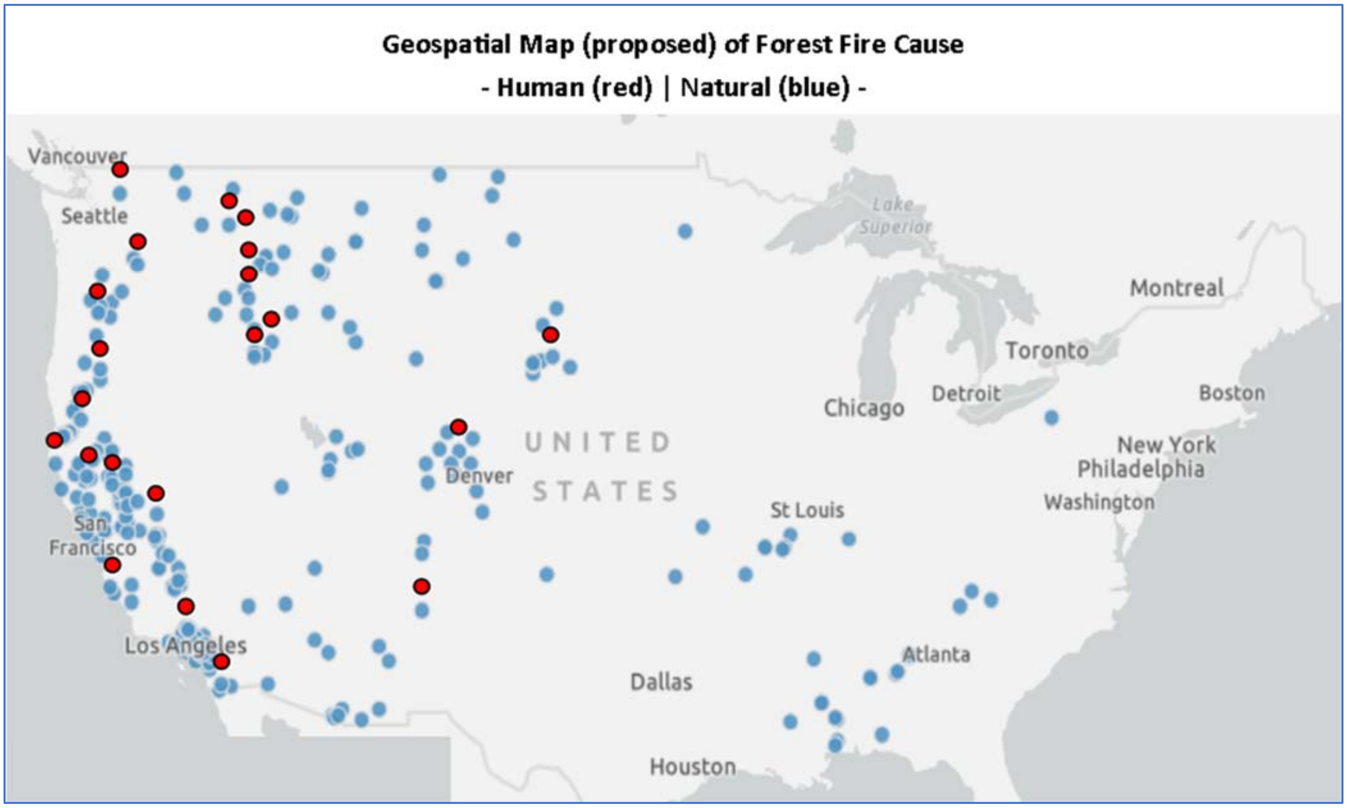

Figure 4 – A proposed geospatial map depicting forest fire cause between 'Human' in red and 'Natural' in blue.

To accomplish improved data visualizations for Figures 2 and 3 above, the additional EPA parameters as outlined in Table 4 would be evaluated in the same graphical fashion as is already done with respect to the parameters of Table 3. This improvement will provide a more granular analysis of the constituent air quality components that are considered most unhealthy (e.g. combustion bi-products correlated to the burning of townships). However, these are simply ways of improving upon existing data visualization methods. For a new perspective, Figure 5 depicts a circular line plot of the annual forest fire activity (red, inner plot) and the associated air quality pollution score (blue, outer plot) for that segment of time. This type of plot can easily filter to any prescribed date range, and as the center of the plot indicates, any air quality parameter can be presented in this manner. It is easy to imagine Figure 5 with a different time scale, where perhaps a monthly scale is used in-place of the annual scale shown.

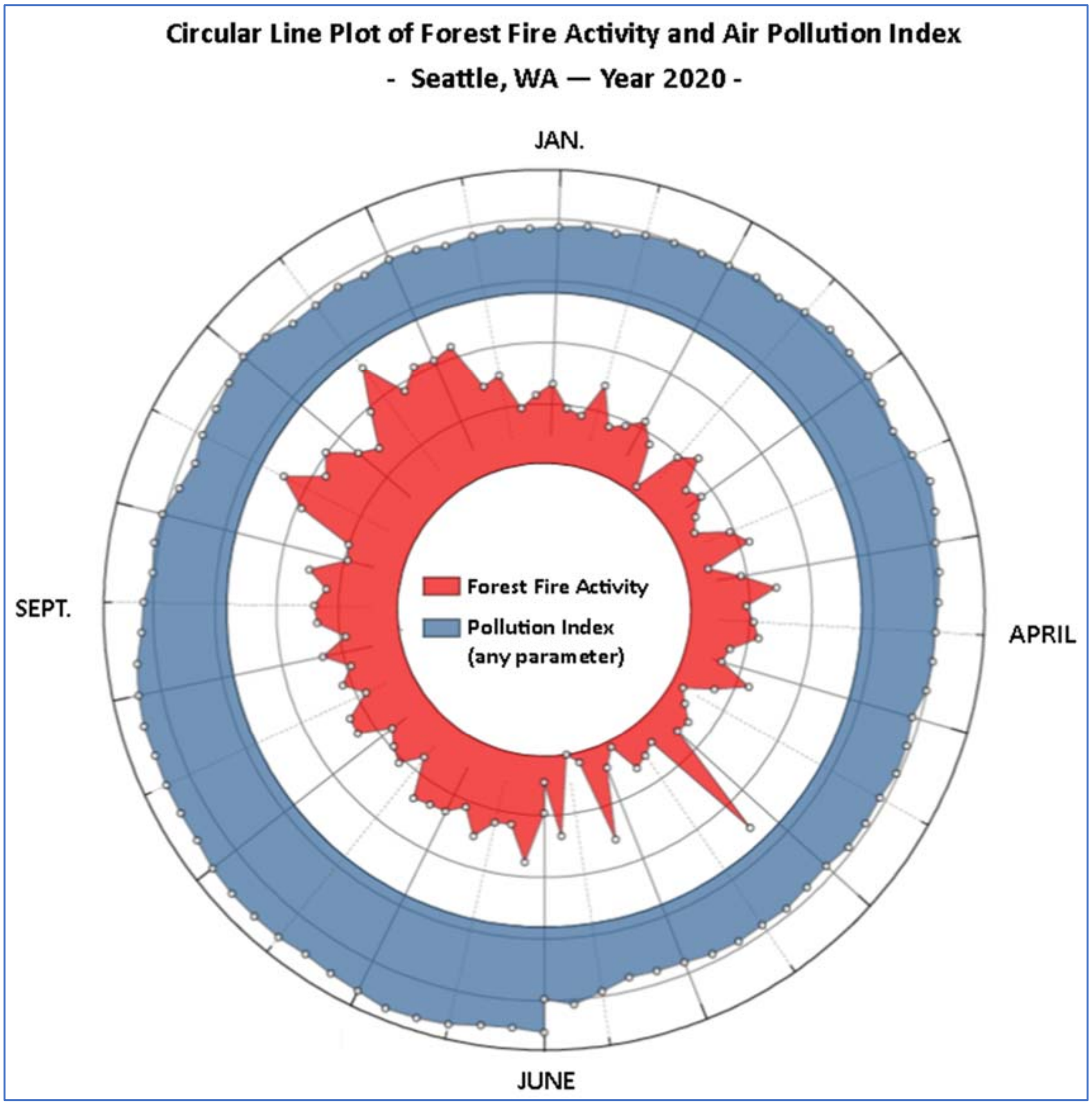

Figure 5 – A circular line plot example comparing forest fire activity with air quality scores for the year 2020. https://www.originlab.com/doc/Origin-Help/Stacked-Radial-Plot

Of course, these additional visualizations (Figures 5 and onward) do not represent our actual dataset values but instead are here given to convey the fundamental visualization format intended to meet this projects' goals. For figure 5 then, the 2 categories of graphic symbols and graphic variables are identified as: [graphic symbols – (Point), (Line), and (Area)] and [graphic variables – (Retinal-Form-Rotation), (Retinal-Form-Curvature), (Retinal-Form-Closure), and (Retinal-Color-Hue)]. The remaining 2 visualizations, Figures 6 and 7, depict hive plots for each of the 2 underlying NIFC and EPA datasets.

In Figure 6, a very interesting hive plot can be established depicting the 3 related variables of pollutant type, pollutant toxicity, and pollutant size. This is valuable, as for any given time period a quick assessment of these complex relationships (or similar) can effectively convey important air quality pollutant outcomes.

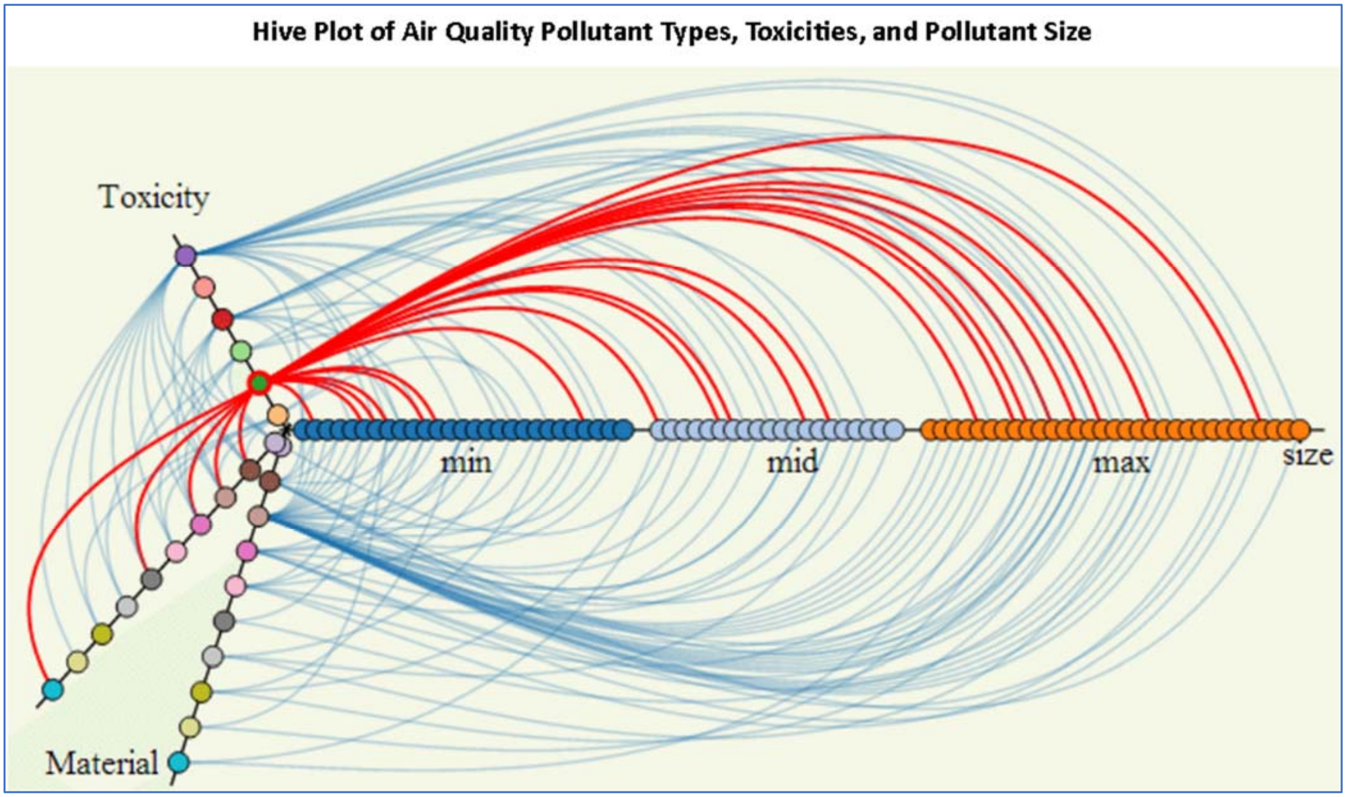

Figure 6 – An example hive plot depicting 3 qualitative variables as identified within the EPA dataset.
https://www.researchgate.net/figure/Hive-plot-view-for-nano-toxicity-type-material-and-particle-size_fig8_304337470

For this figure, we see that [graphic symbols – (Line) and (Area)] and [graphic variables – (Retinal-Color-Hue), (Retinal-Form-Curvature), (Retinal-Form-Size) and (Retinal-Color-Saturation)]. Using these variable types helps to clearly observe the trends and categorizations that are visualized for improved interpretation.

Figure 7 is of similar fashion as the prior and reflects the forest fire dataset provided by NIFC. We see the chosen axis here are 'Fire Code', 'Daily Acres Burned', and 'Predominate Fuel Group'. Alternate axis labels could also be 'Calculated Acres', 'POOState', and 'POOJurisdictionAgency'. These are attributes that are identified in Table 1 for the underlying dataset. In this hive plot, the following graphic elements are observed: [graphic variables – (Area)] and [graphic symbols – (Retinal-Form-Curvature) and (Retinal-Color-Hue)]. Each of these variable types helps to make a more meaningful visualization.

Hive Plot of Specified Forest Fire Event Parameters

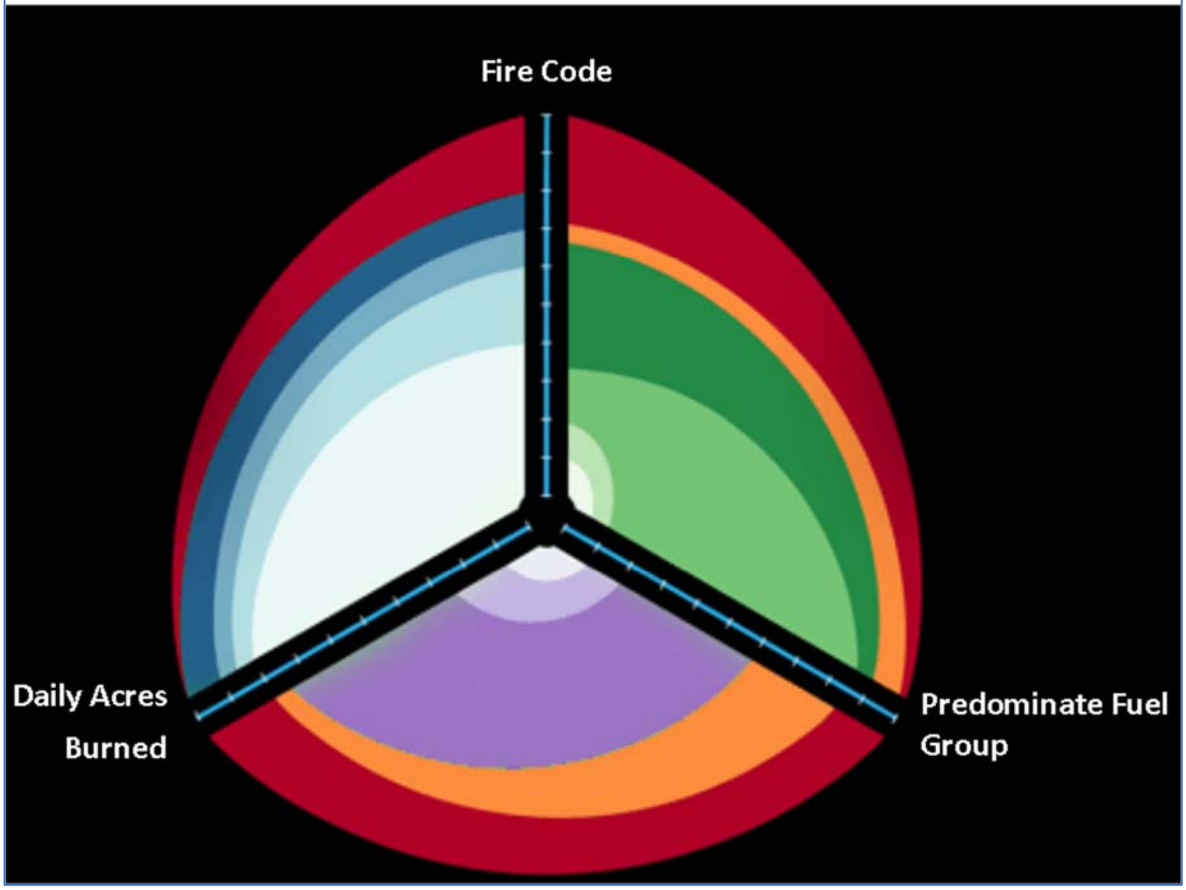

Figure 7 – An example hive plot pertaining to specified attributes of the NFIC dataset. https://seeingcomplexity.wordpress.com/2011/02/05/hive-plots-and-hairballs/

Interpretation of Results

Discuss the key insights that you hope to gain from the visual analytics project proposal. Try to tie these insight gains to decisions that your stakeholders will make based on this analysis. How will you evaluate the success of your work to the visualization and decisions made as a result?

Discuss any challenges and opportunities you foresee during the project. Are there any problems that have surfaced in your initial discussions or that you might test during validation and evaluation of your project?

The key insights gained from this project are focused upon the toxic US forest fire bi-products that are created resultant from the recent (September 2020) total destruction of townships by way of forest fire events, particularly at the west coast of the United States. The most common way to assess such bi-products is to evaluate sensor data related to environmental parameters that are scored and categorized such that the general public can make decisions and maintain awareness of air quality. However, the commonly reported air quality scores are categorized into only 3 types (see Figure 3): ozone, PM2.5, and PM10. Therefore, in order to uncover the more intricate details of this visual analysis project, additional air quality parameters (see Table 4) should be assessed. Once this level of awareness is achieved and communicated to the 2 stakeholder groups (government and private individuals), perhaps, they each can then more effectively make reliable decisions of existing conditions. For example, the hive plot shown in Figure 7 is effective at concluding what magnitudes a certain parameter is trending (increase or decrease). Similarly, Figure 4 depicts the classification between a forest fire caused by a human versus one that is ignited by natural means. Leveraging this visualization allows government authorities to observe geospatial locations of criminal activities, and an ability to foresee future problem locations as well.

A successful project outcome will convey a higher level of detailed information in a visualization format, as already discussed. Measures of project success beyond this threshold could be that organizations and individuals would come to more frequently use the visualizations in day to day activities, say, of where to recreate or not for an individual stakeholder. However, it is expected that available datasets may not contain enough useful information in order to convey the appropriate visualization recommendations. Mainly, not all air quality monitor stations contain sensors (nor data) that will support those parameters given in Table 4. Thus, not having sufficient data across a geographical area will not be useful enough in achieving the full project potentials. In this regard, the project may become unfeasible, however, there are potentially additional data sources yet to be discovered, such as Purple Air (https://www2.purpleair.com), that is a network of private individual air quality monitor stations different than the EPA system employed in this project.