SFU SCHOOL OF COMPUTING SCIENCE

Forest Fire Risk Prediction & Analyses

CMPT 732 Project Report

Aidan Vickars (301255806)

Rishabh Kaushal (301443177)

Anant Sunilam Awasthy (301467798)

Problem Definition

Due to global warming, forest fires are becoming an increasingly common and dangerous norm in today's world. This was particularly evident this past summer in British Columbia (BC) where temperatures reached over $40^{o}C$ that kicked off one of the worst fire seasons on record. This resulted in hundreds of fires across the province that displaced thousands of people, caused millions of dollars in property damage, and destroyed huge swarths of forest. While this problem is particularly local to us it is not limited to just BC and is a major problem around the world. Therefore, the aim of this project is to analyze factors in relation to forest fires over time then leverage our analyses into a machine learning model that predicts the risk of a forest fire daily. These analyses include air quality, forest characteristics, and weather data.

There are two key challenges associated with this problem. The first is integrating data from a variety of different sources and formats. For example, a key problem that will be elaborated on below was efficiently extracting the perimeter of every forest fire from a raw geojson format into a parallelizable format and subsequently linking it to different regions of BC in an efficient manner. The other challenge was downloading, storing, computing, and analyzing the big Air Quality datasets of BC that are around 85 GB.

Methodology

For data analyses, we leveraged Spark, AWS S3, AWS Redshift, AWS PostgreSQL (RDS) and Dash. Spark was used due to its significant parallelization abilities that allowed us to manipulate large data sets. A notable example was processing the "Forest Cover Inventory" dataset[1] that contained approximately 18 GB forest data. Another significant example was the air quality dataset[2] like CO.csv for 1980-2008 years was around 1 GB – 1.5 GB and there were 14 big csv files like this. PySpark was able to reduce these files to just around 1MB – 2MB on average.

The problem we are trying to solve for the Air Quality datasets is to find trends to see if the air quality is affected over the years. To do this, once the air quality csv datasets were reduced, we stored them on an AWS S3 bucket. From the S3 bucket, they were copied to both AWS Redshift and AWS PostgreSQL (RDS) databases for further analyses. AWS databases will be able to scale easily once more data is available in the future. Dash then used the air quality data in an AWS PostgreSQL (RDS) database and local files as well to show various analytical plots.

In a different facet, Dash was used due to its abilities in two aspects. The first was its ability to break our entire dashboard into components that allowed all members of the project to work on the dashboard concurrently. The second was its ability to allow for the use of other python packages. To be specific, it allowed us to incorporate both GeoPandas and Plotly that together enabled us to visualize longitude and latitude coordinates in a polygon format. Furthermore, for air quality analyses, Dash directly uses the air quality data stored in AWS PostgreSQL (RDS) database.

Problems

Partitioning

A significant problem we encountered was how to examine areas of BC on a granular level while maintaining scalability. Our initial approach was to examine areas BC using predefined polygons[3] made up of longitude and latitude coordinates. However, this created a significant problem while attempting to determine over lapping areas that have experienced forest fires. To be succinct, to identify areas that have experienced

[1] British Columbia Data Catalogue. "Results – Forest Cover Inventory." Accessed Nov 24, 2021. https://catalogue.data.gov.bc.ca/dataset/results-forest-cover-inventory.

[2] British Columbia Data Catalogue. "Air Quality Monitoring." Accessed Nov 24, 2021. https://catalogue.data.gov.bc.ca/dataset/air-quality-monitoring-verified-hourly-data/.

[3] British Columbia Data Catalogue. "Legally Defined Administrative Areas of BC – Internal." Accessed Nov 24, 2021. https://catalogue.data.gov.bc.ca/dataset/legally-defined-administrative-areas-of-bc-internal.

multiple fires we used the Ray Casting Algorithm[4] that given a point and a polygon with *n* sides will determine if the point is internal or external to the polygon by iterating through each edge of the polygon in *O(n)* time. However, to determine if a forest fire is at least partially contained in the area of another involves *O(nm)* comparisons where *m* and *n* are the number of coordinates in the polygons of each fire respectively. Our dataset[5],[6] contained approximately 20 000 forest fires. The time complexity of this quickly became too large.

To solve this, we took two approaches. The first was to divide BC into a 10×10 grid of square partitions. This solved the time complexity issue from above because determining if a coordinate was internal or external to a partition required exactly four comparisons. To be succinct, this required checking if a coordinate was between the left and right sides, and between the top and bottom sides of the partition. In formal terms, determining if a forest fire was at least partially contained in a region now took *O(4m) = O(m)*; a reduction by a factor of *n.* Finally, we subsequently filtered out the partitions that were not inside BC by applying the Ray Casting Algorithm on every corner and center of each partition with respect to the perimeter of BC. The resulting partitions are shown in Figure 2 below.

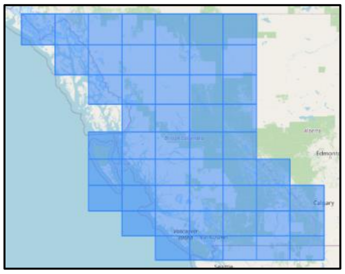
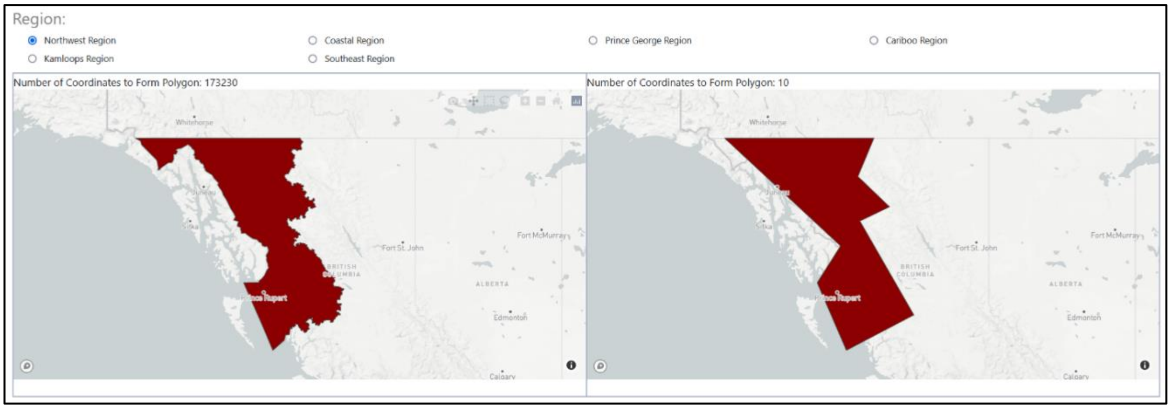


Figure 2: Partitions of BC

The second, approach we took was to use the predefined polygons of BC[7] that could contain well over 100 000 coordinates, and reduce the number of coordinates in each polygon using the "Ramer-Douglas-Peucker" Algorithm[8]. After applying this algorithm on each polygon, we were able to produce representative polygons that contained a dramatic reduction in the number of coordinates they contained. The resulting time complexity is like that of above. An example of this is shown in Figure 3 below.



Region:

◉ Northwest Region ○ Coastal Region ○ Prince George Region ○ Cariboo Region

○ Kamloops Region ○ Southeast Region

Number of Coordinates to Form Polygon: 173230

Number of Coordinates to Form Polygon: 10

Figure 3: Sample Results of Ramer-Douglas-Peucker" Algorithm

[4] Wikipedia. "Point in Polygon." Accessed Nov 24, 2021. https://en.wikipedia.org/wiki/Point_in_polygon.

[5] British Columbia Data Catalogue. "Fire Perimeters – Current." Accessed Nov 24, 2021. https://catalogue.data.gov.bc.ca/dataset/fire-perimeters-current.

[6] British Columbia Data Catalogue. "Fire Perimeters – Historical." Accessed Nov 24, 2021. https://catalogue.data.gov.bc.ca/dataset/fire-perimeters-historical.

[7] British Columbia Data Catalogue, "Legally Defined Administrative Areas of BC – Internal."

[8] Wikipedia. "Ramer-Douglas-Peucker algorithm." Accessed Nov 24, 2021. https://en.wikipedia.org/wiki/Ramer%E2%80%93Douglas%E2%80%93Peucker_algorithm.

BC Air Quality Datasets

One of the most substantial problems we faced was processing and reducing the BC Air Quality Data Set[9] from an unmanageable 84 GB to a manageable 8.5 MB. The following is the approach we took to do this. First, instead of analyzing all the air quality related parameters, we decided to only use a subset of them. Secondly, the datasets contained hourly values from various air monitoring stations. However, we did not require that level of detail. As a result, we decided to calculate monthly and annual averages for the stations using PySpark (data-frame/SQL) scripts and used these averages instead of the hourly values. This significantly reduced the overall size of the datasets.

One of the issues that we faced while calculating monthly and annual averages for various air monitoring stations occurred during the "JOIN" queries. The main problem was that we were joining the annual and monthly averages in a single query using a "With" subquery. We found this query to be inefficient because it was first calculating the annual average from a large data-frame with approximately 5.6 million rows. Then, from the same large data-frame, it calculated the monthly averages. Performing all of these calculations in a single query made the computation time extremely long. To solve this, we calculated the monthly and yearly averages separately, and then joined these significantly smaller data-frames together which reduced the overall computation time from about 1 hour to about 2 minutes per CSV dataset. The final CSV files were then loaded to an AWS S3 bucket and then copied to an AWS Redshift database.

Now, the subsequent problem that we faced occurred while importing the cleaned csv files from AWS S3 bucket to the AWS Redshift database. The main issue was that the rows that contained NA/NULL in the datasets were causing column type mismatch errors during the import. For instance, an integer column in an AWS RedShift database didn't know how to treat a null or a NA value. To resolve this, we coded air_extract_df.py which removed all the rows containing NA and NULL values from all the final datasets.

We then wanted to use that Redshift Database to connect to our Dash server for air quality analyses and visualizations. We tried various connectors, but nothing worked. To solve this, we had to shift the data from the AWS Redshift database to an AWS PostgreSQL (RDS) database. We were then able to use the RDS database to connect to Dash by creating our own API. Thus, in the future if we have data that is live in some capacity, Dash will be able to use that dynamic data directly from the RDS database.

Once we connected Dash to the AWS RDS database, we realized that just calculating annual averages and monthly averages by the stations was not a promising idea as there were more than 550 distinct stations. This meant that we would have to plot 550 different plots to visualize how the air quality varied for each station. To solve this, we decided to instead calculate annual averages and monthly averages for the regions since there were only 7 distinct regions. Thus, we created visualizations for just 7 regions instead of 550 stations.

Results

Forest Data

Research Question: Are there particular forest characteristics that make a forest more susceptible to forest fires?

To analyze the relationship between forest fires and the forests they destroy, we examined the "Forest Cover Inventory" dataset[10] that contains 18 GB of forest characteristics like density, age, tree type etc. and the coordinates of each forest respectively. We analyzed this at a granular level by mapping each forest contained in the dataset to the partitions of BC as defined in the previous section. The result of this mapping is shown in Figure 4 below. As can be seen in Figure 4, we found that there was no data for well over fifty percent of BC. Because of

[9] British Columbia Data Catalogue, "Air Quality Monitoring."

[10] British Columbia Data Catalogue, "Results – Forest Cover Inventory."

this, and because forest fires are not limited to any particular regions of BC, we did not perform any further analysis using this dataset. We felt that because there was no data for the majority of BC, extracting any meaningful findings would be difficult and possibly misleading because the insights would be limited to a small subset of BC.

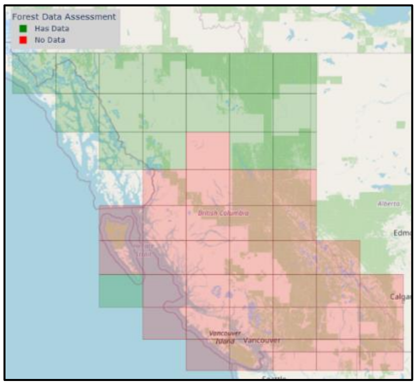


Figure 4: Forest Data Assessment

Forest Fire Recurrence

Research Question: Does a region that experiences a forest fire in one year become less susceptible to a fire the next year?

To answer this question, every fire was mapped to any partitions that it overlapped to allow for a granular analysis. Then for each partition, the number of years between successive fires was computed. For example, if some partition had a fire in year *x*, we computed the number of years until this partition had another fire. The initial results are shown in Figure 5 below.

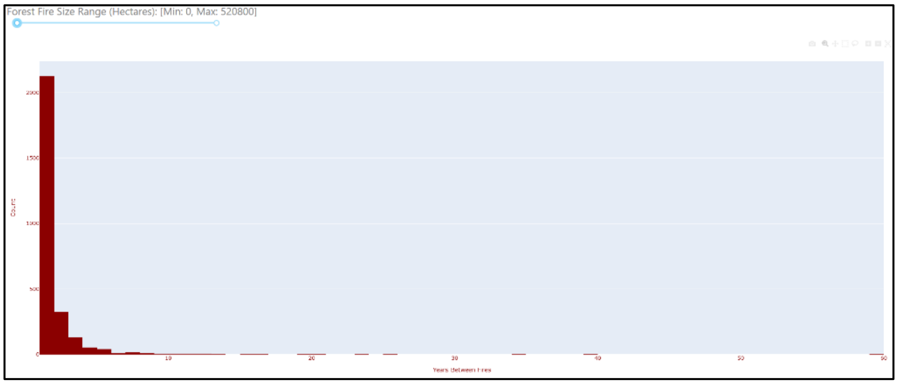


Figure 5: Number of Years Between Fires (No Size Limit)

Surprisingly, the resulting histogram is extremely right skewed. This indicated that a region that had a forest fire in one year was no less susceptible to having a forest fire in the next year. However, unsurprisingly if we increase the minimum fire size, the histogram begins to become less right skewed. An example of this is shown in Figure 6 below.

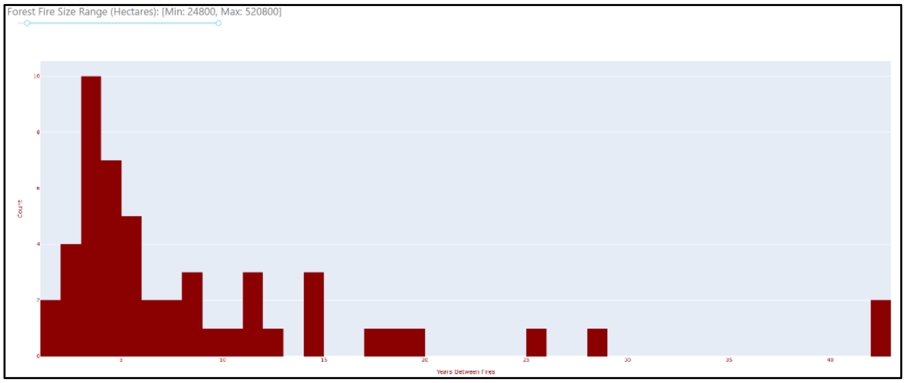


Figure 6: Number of Years Between Fires (Min Fire Size: 24 800 Hectares)

As we can see, the most common value in the histogram is trending towards the right which is likely due to larger forest fires being infrequent.

Weather

Research Question: Is there a temperature threshold where forest fires became more likely?

To answer this question, we used the daily maximum temperature (TMAX) value from the daily weather observations contained in the dataset from the "Global Historical Climatology Network"[11]. Then, we computed the approximate center of each forest fire and for the preceding 7 and 14 days for every forest fire, we mapped the center of each fire to closest weather station that had a TMAX value for the given day respectively. The 7 and 14-day TMAX averages were computed, and the 7-day average is visualized in Figure 7 below.

Surprisingly, the peak of the histogram is quite low at approximately 18 degrees Celsius. The 14-day TMAX average is similar. This indicated that the threshold when forest fires become more frequent is surprisingly low.

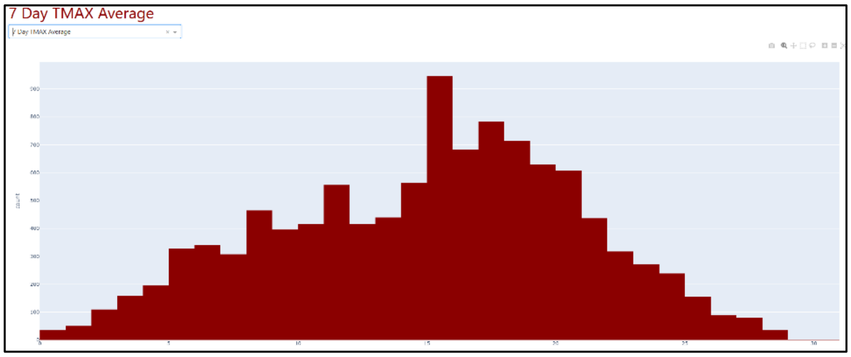


Figure 7: 7 Day TMAX Average

Forest Fire Risk Prediction

Research Question: Can we create a model that can assess the risk of a forest fire in each partition?

To answer this question, we leveraged the mappings discussed above between the partitions and the weather stations and forest fire perimeters respectively, and for every day and partition for approximately the last 100 years we computed the preceding 14-day TMAX and precipitation average for every forest fire. We added an indicator to indicate if a fire occurred on the given day and computed the day of the year as a feature as well.

[11] National Centers for Environmental Information. "Land-Based Station." Accessed Nov 24, 2021. https://catalogue.data.gov.bc.ca/dataset/fire-perimeters-current.

However, this created a significant problem; the dataset became heavily unbalanced as only 1% of the data contained forest fires. To combat this, we recognized that our model, a logistic regression model, does not require a large amount of data. So, we randomly sampled 5% of the data that did not contain forest fires and merged this subset with the data that contained forest fires. This resulted in a split of approximately 20/80% for data points with and without forest fires respectively. It should be noted that an imbalanced dataset was still desired so that an accurate representation of the truth was maintained. To be succinct, we desired an imbalanced dataset because forest fires are a relatively rare event given the context of an entire year.

An observant reader will suspect that given such an imbalanced dataset, the model would never classify an area as having a forest fire. This is certainly the case, however predicting a forest fire was never the intention rather predicting the risk of a forest fire was the end goal. We were able to extract this by examining the level of surety in the model's probabilistic output. That is, we defined logic such that if the model produced a probability of less than or equal to 0.75 the partition was high-risk, if the output was between 0.75 and 0.8 the partition was medium-risk and otherwise the partition was low risk. A sample result is shown below in figure 8.

Note, the available dates for selection in the dashboard has been constrained to Jan 1, 2020 to Oct 10, 2021 to improve speed.

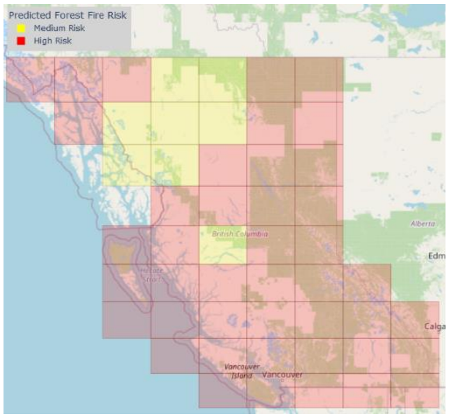


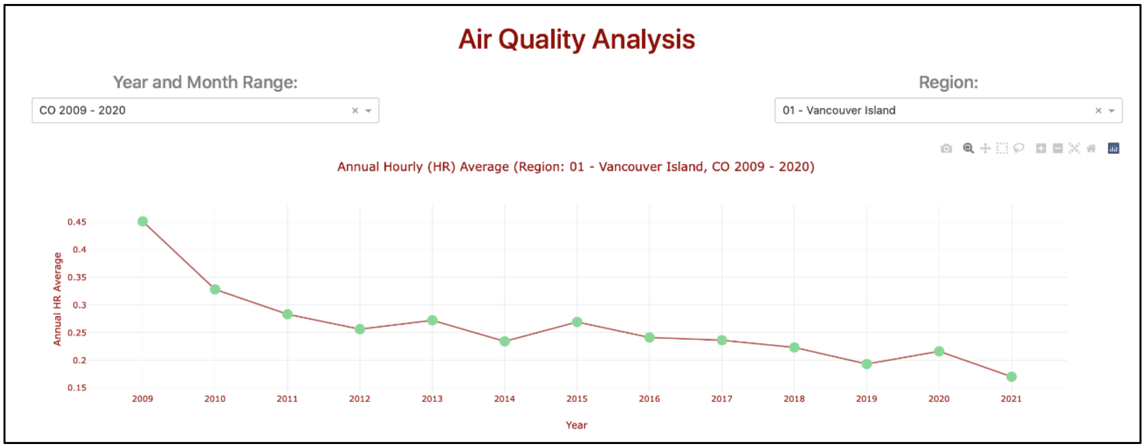
Figure 8: Forest Fire Risk Aug 8, 2020

BC Air Quality Data Analyses

Research Question: What affect do forest fires have on air quality over time?

To answer this question, we implemented interactive line graphs that show both the yearly (1-hr) averages and the monthly (1-hr) averages for CO, NO2, NO, O3, PM10, PM25, and SO2 for years 1980 – 2008 and 2009 – 2020. The user can also click on a point depicting an annual average on the annual averages plot (the top graph on the air analysis tab's page) and the plot below will show the monthly averages for that point (year).

These yearly averages can be seen for various regions of the BC by choosing the desired region from the "Region" dropdown. For example, looking at figure 8.1, Carbon Monoxide (CO) yearly averages for 2009 – 2020 did not increase as the forest fires increased over the years for the Vancouver Island Region.



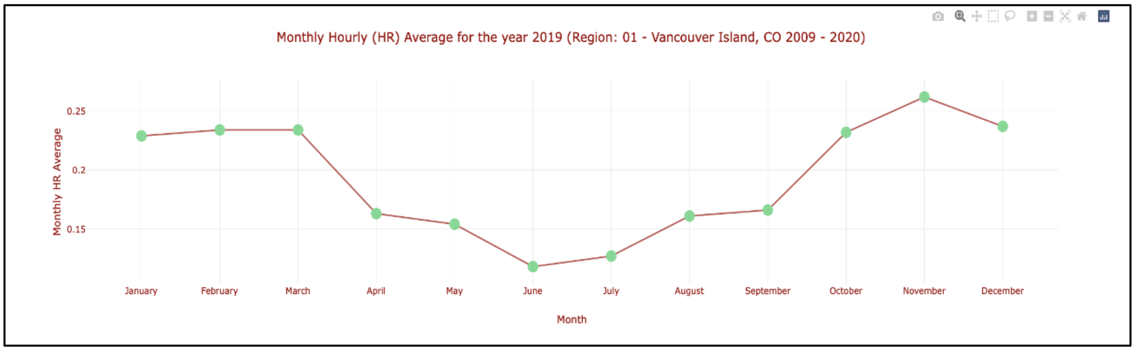
Air Quality Analysis

Year and Month Range: CO 2009 - 2020

Region: 01 - Vancouver Island

Annual Hourly (HR) Average (Region: 01 - Vancouver Island, CO 2009 - 2020)

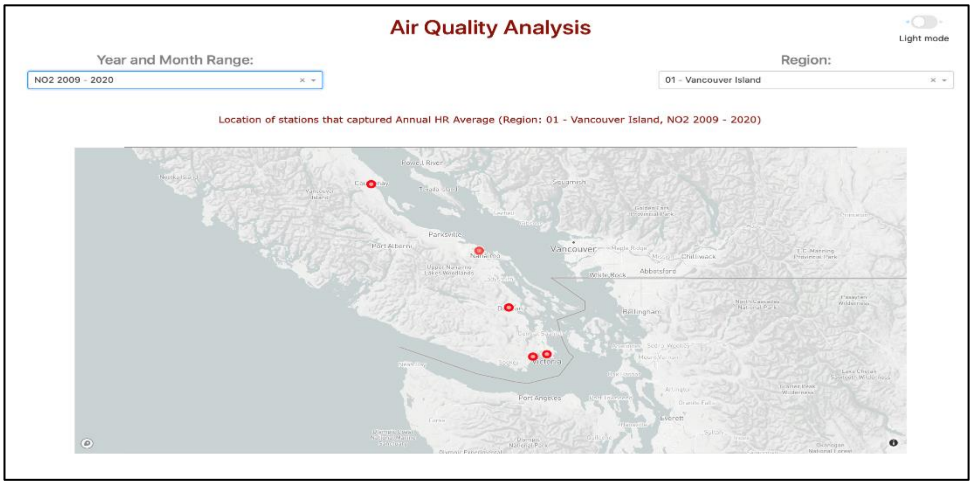
Figure 9.1: Line graph showing Carbon Monoxide's (CO) Yearly Averages for the Vancouver Island region from 2009 - 2020



Monthly Hourly (HR) Average for the year 2019 (Region: 01 - Vancouver Island, CO 2009 - 2020)

Figure 9.2: Line graph showing Carbon Monoxide's (CO) Monthly Averages for the Vancouver Island region for 2019

The second dashboard called the "Air Quality Station Mappings" shows the locations of all the air quality monitoring stations in a particular Region of BC. This allows the user to determine where an air quality measurement was taken. For instance, the figure 10 below shows all the air quality monitoring stations that captured Sulphur Dioxide (SO2) from 2009 – 2020 in the Vancouver Island Region.



Air Quality Analysis

Light mode

Year and Month Range: NO2 2009 - 2020

Region: 01 - Vancouver Island

Location of stations that captured Annual HR Average (Region: 01 - Vancouver Island, NO2 2009 - 2020)

Figure 10: Air Quality monitoring stations across Vancouver Island Region that captured Sulphur Dioxide (SO2) from 2009 – 2020.

Forest Fire Size Distribution

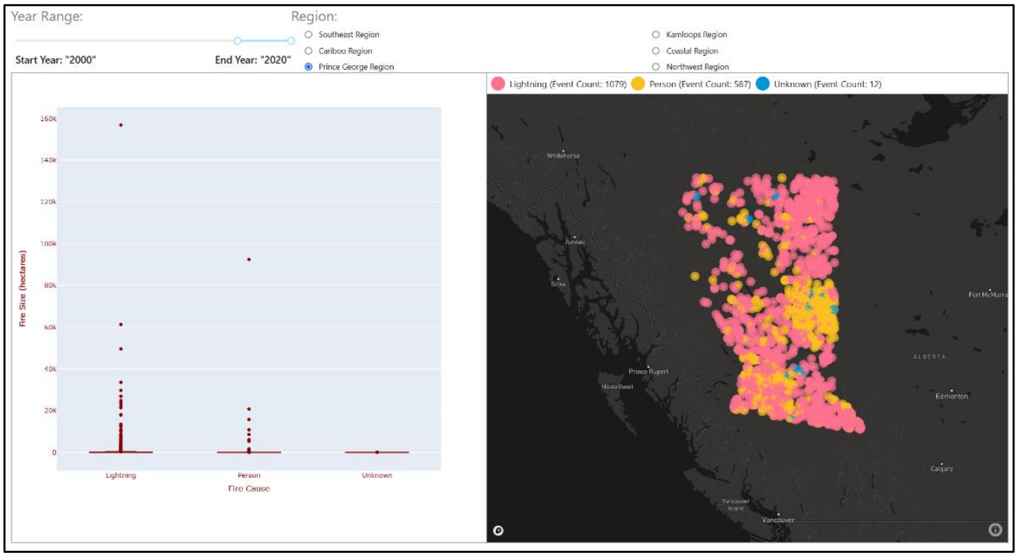
Research Question: What is the distribution of forest fires with respect to size?

To answer this question, we split the forest fires into three groups with respect to their causation. These groups are fires that were caused by lightning, person and unknown. Then, we visualized the results into a box plot that allowed us to understand the distribution of forest fires. By examining the boxplot, we found that most fires are relatively small, with several notable outliers that heavily skew the distribution. A sample of the results is shown in figure 11 below.

Forest Fire Cause

Research Question: What are the primary causes of forest fires?

To answer this question, we again split the forest fires into three groups with respect to their causation and then further split the fires by region. To be succinct, we leveraged the reduced regional polygons that are discussed under the Partitioning section under Problems and mapped each fire to a region. We did this by computing the centroid of a polygon that defines a fire and used the Ray Casting Algorithm[12] to map each centroid to a corresponding region. This then allowed us to view a geographic scatterplot of the causes of forest fires on a regional level. We found that overwhelmingly, lightning strikes are the primary cause of forest fires. A sample of the results is shown in figure 10 below.



Year Range:

Start Year: "2000" End Year: "2020"

Region:

○ Southeast Region
○ Cariboo Region
● Prince George Region
○ Kamloops Region
○ Coastal Region
○ Northwest Region

● Lightning (Event Count: 1079) ● Person (Event Count: 587) ● Unknown (Event Count: 12)

Figure 11: Sample Results of Forest Fire Size Distribution & Causation

Forest Fire Cycles

Research Question: Does the number of forest fires each year increase and decrease in a cyclical nature?

To answer this question, we plotted the number of fires vs. year in a bar chart ordered by year. The results clearly show a cyclical nature of fires where the number of fires will peak in a particular year, then trend downwards for a few years and then peak again beginning the cycle once again. A sample result of this that clearly shows the cyclical nature is shown in Figure 12 below.

[12] Wikipedia, "Point in Polygon."

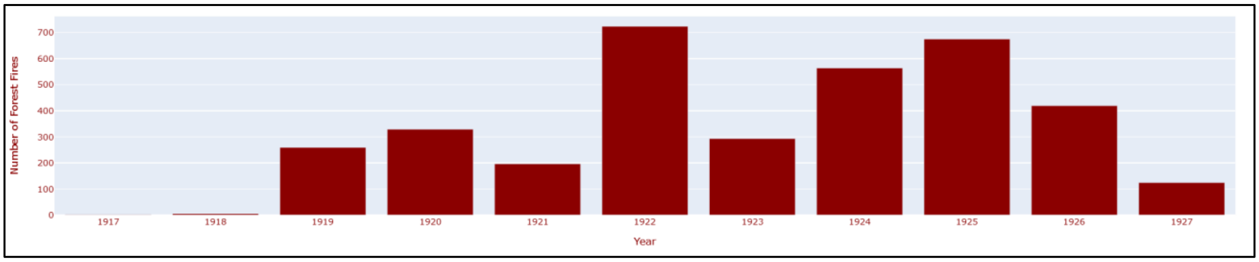


Figure 12: Number of Forest Fires by Year

Project Summary (Points 20)

Category	Marks	Description
Getting the data: Acquiring/gathering/downloading	1	• Consisted of downloading various large csv and geojson files.
ETL: Extract-Transform-Load work and cleaning the data set.	4	• Performed ETL on several large datasets in a variety of forms and used a variety of tools (Spark, AWS S3, AWS Redshift, AWS PostgreSQL).
Problem: Work on defining problem itself and motivation for the analysis.	2	• Defined the problem and emphasized its importance due to global warming and increasing number of forest fires as a result.
Algorithmic work: Work on the algorithms needed to work with the data, including integrating data mining and machine learning techniques.	4	• Developed two solutions/techniques to allow for a scalable way to examine different regions with respect to forest fires. This included an analysis explained why our approaches work using big-O notation. • Created novel approach to assess fire risk using machine learning that to the best of our knowledge has not been done before at least in BC. • Integrated a variety of different data sets and types together.
Bigness/parallelization: Efficiency of the analysis on a cluster, and scalability to larger data sets.	1	• Ensured all tools maintained a high degree of scalability and ensured all techniques/algorithms are scalable as well. • Analyzed several large datasets such as the “Forest Cover Inventory” and the air quality datasets.
UI: User interface to the results, possibly including web or data exploration frontends.	3	• Developed and deployed a comprehensive and interactive dashboard in Dash to explore data
Visualization: Visualization of analysis results.	3	• Developed and deployed a comprehensive and interactive dashboard in Dash to visualize results
Technologies: New technologies learned as part of doing the project.	2	• Learned Dash, Plotly, AWS Redshift, and AWS PostgreSQL