Rodney Cordova

Term Project

COMP 4449

**Mystery at the Wildlife Preserve:**

**Plume Analysis**

Throughout postindustrial history, air pollution has been a major reason for biological species loss. By its very nature, air pollution is often very hard to detect. Many forms of air pollution require specialized monitoring equipment and expert knowledge to identify harmful emissions. In this project, I was presented with a hypothetical scenario in which a native species of birds is experiencing population loss. It is suspected the population loss may in part be caused by chemical emissions from four local manufacturing plants. The objective of this analysis is to utilize the provided data to answer three questions.

1. Characterize the sensors’ performance and operation. Are they all working properly at all times? Can you detect any unexpected behaviors of the sensors through analyzing the readings they capture?
2. Which chemicals are being detected by the sensor group? What patterns of chemical releases do you see, as being reported in the data?
3. Which factories are responsible for which chemical releases?

**Background:**

The scenario is set in a mid-sized city called Mistford on a large nature preserve. A student from a local college, Mitch, is studying nesting pairs of the Rose-Crested Blue Pipit and has noticed signs of a decrease in the population. The Pangera Ornithology Conservation Society is sponsoring the student to undertake additional studies to identify the possible reasons for the population decline. Mitch suspects that the cause of population loss may be due to air born chemicals emitted by local manufacturing factories. There are four factories in the area that may be contributing to the chemical emissions. Mitch has been able to gather sensor data from the local government. Each of the government’s 9 sensors has readings for three months in 2016 of four chemicals present in the air of the region. The monitored chemicals are:

* Appluimonia – In general, most substances that cause odors in the outdoor air are not at levels that can cause serious injury, long-term health effects, or death to humans or animals. However, odors may affect quality of life and sense of wellbeing. Several odor-producing substances, including Appluimonia, are monitored.
* Chlorodinine – Corrosives are materials that can attack and chemically destroy exposed body tissues. They might be hazardous in other ways too, depending on the particular corrosive material. An example is the chemical Chlorodinine which is harmful if inhaled or swallowed.
* Methylosmolene – This is a trade name for a family of volatile organic solvents. This chemical was strictly regulated in the manufacturing sector. Liquid forms of Methylosmolene are required by law to be chemically neutralized before disposal.
* AGOC-3A – New environmental regulations and consumer demand have led to the development of low-VOC (Volatile Organic Compounds) and zero-VOC solvents. Most manufacturers now use one or more low-VOC substances and Mistford’s plants have wholeheartedly signed on. These new solvents, including AGOC-3A, are less harmful to human and environmental health.

Additionally, Mitch was able to acquire meteorological data for the region during the observation period.

**I – Data:**

Mitch provided two datasets, one containing the sensor readings and one containing the meteorological data. In the meteorological data each observation consists of the date and time the reading was taken, the wind direction (using a north-referenced azimuth bearing), windspeed in meters per second, and elevation. Each day in the observed timeframe consists of eight observations. The observation is taken at midnight (00:00:00) and then every three hours after that with the last reading being taken at 18:00:00. When the missing values are plotted (see Appendix I) it becomes evident that the elevation is missing from all except the first observation. We are told that the meteorological data is always taken from the same sensor location. This means the elevation is constant, thus it was dropped from the dataset along with a blank column (Unnamed: 3). The missing values plot also showed two observations missing data (443 and 458). Upon closer inspection observation 458 appears to be a blank row so it was dropped. Observation 443 has a date stamp but is missing the wind direction and windspeed. Rather than dropping this observation I decided to forward fill the windspeed and direction values based on the previous value. As seen in appendices II and III there does not appear to be any trends related to wind speed and wind direction relative to the date of the observations.

The sensor readings dataset is much larger than the meteorological datasets consisting of over 78,000 observations despite covering the same time frame. The reason this dataset is so much larger has to do with how the observations are collected. For each day in the three-month period, each of the 9 sensors makes a scan at the beginning of each hour. Each scan takes a reading of the parts per million level of each of the four observed chemicals. As a result, the number of observations is much higher. Each observation consists of the chemical being detected, the monitor / sensor number, the date and time, and the reading level (parts per million). Based on the missing value plot (see Appendix IV) none of the observations are missing any data.

**II – Sensor Performance:**

The first question being asked is to evaluate the sensor performance. This is a logical starting point, as if the sensor readings are erroneous then the conclusions drawn from them may also be faulty. To start I decided to investigate the number of total readings. Well, the missing value plot showed all the observations were complete the total number of observations in the sensor data appears low. Based on how the observations are made I would expect there to be 79,488 total observations (9 sensors x 4 chemicals x 24 hours x 92 days). However, the dataset only consists of 79,243 observations. When the count of observations by day is plotted on a heat map (see appendix V) it becomes apparent that there are missing observations during the first week of each month. Additionally, when the total reading is plotted by the hour of day they occurred (see appendix VI) all the missing observations occurred at midnight. This leads me to believe that the missing observations are due to some type of sensor reset that is occurring at midnight in the beginning of each month rather than an error with the sensors. Furthermore, the 245 total missing readings would account for 0.30% of the dataset. Such a low missing value rate should not impact any trends in the data.

When evaluating unexpected behaviors in a dataset the main technique that comes to mind is anomaly detection. Anomaly detection is the process of identifying data points, events, or observations in a dataset that deviate from the norm. I expect there to be some anomalies in all sensors but if those anomalies are more prevalent in readings from a specific sensor, then that would indicate that sensor may not be functioning properly. When implementing unsupervised anomaly detection algorithms on time series data like our sensor log it is important that the data is stationary and have no auto correlation (Jyenis, 2020).

Shape

Description automatically generated Stationarity is a measure of how much the mean and standard deviation of a data set change over time. If the mean and standard deviation change over time the data is not stationary. Autocorrelation occurs when the data is correlated with itself in a different time period. To check for stationarity, I performed a Dickey-Fuller test. This test checks the null hypothesis that the time series has a unit root vs the alternative that the time series does not have a unit root, meaning it its stationary (Brownlee, 2020). I conducted the test with an alpha value of 0.05. With a p-value approximately equal to zero I reject the null hypothesis in favor of the alternative and conclude the data is stationary. For testing autocorrelation, I generated Auto Correlation Function (ACF) Plots. The ACF plot show if elements of the time series are positively correlated, negatively correlated, or independent. Points above zero will indicate a positive correlation and points below zero indicate a negative correlation. In this case (to the right) the points are centered around zero indicating no autocorrelation. Since both assumptions of stationarity and no autocorrelation are satisfied, I proceeded with modeling the analysis with two unsupervised machine learning algorithms.

K-Means Clustering:

The first anomaly detection method I used was k-means clustering. K-means clustering groups points that are similar based on Euclidean distance into K number of clusters. In this case K is equal to 2 where one cluster is normal points, and the other is anomalies. K-means works by randomly assigning each observation to one of the groups then iterates over that process until the clusters stop changing. Then for each cluster, the cluster centroid is calculated where the centroid is the vector of feature means for the observations in the cluster. Finally, each observation is assigned to a cluster with the center closest to the observation based on Euclidean distance (Kizil, 2020). A new feature vector was added to the data with the point predictions where 0 is a normal reading and 1 is an anomaly. When grouped by sensor I found most sensors have and anomaly percentage around 1% but sensor 1 has 11.79% anomalies and sensor 9 is 100% anomalies (see appendix VII).

Isolation Forest:

The second anomaly detection method I used was isolation forests. Like K-means isolation forest can be used for clustering. Isolation forests work based off the decision tree algorithm. The algorithm works by building a collection of isolation trees from random subsets of the data, then aggregating the anomaly score from each tree to come up with the final anomaly score for each point. This is done by first randomly subsetting the data until every point in the data set is isolated. Each point is ran through the prediction process. For each isolation tree in the forest, a binary search across the isolation tree is preformed until the leaf is searched and an anomaly score is assigned based on the depth of the path to the leaf (verma, 2021). The formula for the anomaly score is , where h(x) is the path length of observation x and c(n) is the average path length of unsuccessful searches and n is the number of external nodes. Then the aggregate anomaly score is calculated based on the anomaly scores of the individual isolation trees (Lewinson, 2019). The sklearn implementation of isolation forests was used to predict anomalies on the sensor data. A new feature vector was added to the dataset with the point predictions where 1 is considered normal and -1 is considered an anomaly. When grouped by sensor I found similar results to the K-means cluster. Most sensors have and anomaly percentage around 1% but sensor 1 has 11.79% anomalies and sensor 9 dropped to 81.79% anomalies (see appendix VII). Both methods found that sensor 9 is mostly anomalies so that sensor will be excluded from future analysis. The other anomalies will be retained as they could represent patterns in the data.

**III – Chemical Patterns:**

Chart, bar chart

Description automatically generated To begin I decided to look at the total parts per million of each chemical detected by the sensor group. As seen in the table to the right, AGIC-3A had the highest part per million chemical levels detected followed by Methylosmolene, Chlorodinine, and Appluimonia. The fact that AGIC-3A is the most detected matches with the information we were provided that most manufacturers have switched to this type of chemical that is less harmful. More concerning is the high levels of Methylosmolene. The information that was provided indicates that Methylosmolene is required by law to be neutralized so the fact that it is being detected at the second highest rate is troubling.

I next decided to check to see if differences in the groups of chemical readings are significantly different from each other. I first generated a box plot (see below). Visually, the means of each group along with their interquartile ranges appear very similar. However, these visual tests are not the most reliable, so I also decided to conduct an ANOVA to test the null hypothesis that the group mean readings between the chemicals are equal verses the alternative that at least one group mean is different.

Chart, box and whisker chart

Description automatically generated

With a p-value less than 0.01 I reject the null hypothesis that the group mean readings between the chemicals are equal in favor of the alternative.

With that in mind, I decided to move on to check for trends in the chemical emissions over the span of the time series. When each reading is plotted individually over the time frame, the readings tend to group around the mean and it is difficult to identify any trends as seen below.

A picture containing line chart

Description automatically generated

In order to limit the noise and enhance trends in the readings I decided to aggregate the readings of each chemical by summing all the readings for each day. The resulting plot shows the daily change in total chemical readings throughout each month (see appendix VIII). Like the initial plot of total chemical readings, AGOC-3A has the highest reading across all months followed by Methylosmolene. AGOC-3A and Appluimonia both peak in the middle of each month whereas Methylosmolene and Chlorodinie have a constant level with an additive seasonal trend. Finally, AGOC-3A and Methylosmolene both reduce in the last few days of each month potentially indicating the end of a production cycle at that point in time.

Diving deeper, I decided to see if there are any trends in the hour of the day that chemicals are emitted. This was done by aggregating the sum of each chemical during each hour of the day (see appendix IX). Chloridine and Appluimonia did not appear to have any trend related to the time of the readings. On the other hand, AGOC-3A and Methylosmolene appear to have a clear emissions schedule. For Methylosmolene, the emission levels peaked between 10:00 pm and 5:00 am then decreased by about 50% during the day. AGOC-3A emission levels were highest between 6:00 am and 9:00 pm then decreased during the night. The emissions of AGOC-3A and Methylosmolene appear to have an inverse relationship. When Methylosmolene is high, AGOC-3A is low and vice versa. To further verify these emission trends based on time, I further divided the readings by the month they occurred (see appendix X). For each month, the emissions schedule for AGOC-3A and Methylosmolene remained approximately the same and the other two chemicals still did not appear to have any significant patterns in emission timing.

**IV - Factories Responsible:**

Identifying the originating factory of the chemical emissions is a difficult problem. The sensors themselves do not store any information that directly links back to the factory. With that said, I do have the meteorological data for the area. Since the chemical emissions are transmitted in the air the direction the emission is carried will be dependent on the direction the wind is blowing. This means it should be possible to backtrack from the sensor taking the reading along the bearing of the wind at the time of the reading to find the closest factory to that bearing.

The first step in that process is determining the bearing from each factory to each sensor. The bearing is the horizontal angle between the direction of one point and another point. Bearing is measured in degrees between 360/0 and 359.9 where 360/0 is true north. The bearing between twos point a,b is determined as follows:

where

and

and is the difference in the Longitudinal value of the two points (MLInterview, 2020).

The plot in appendix XI shows the general layout of the region with the location of each sensor and factory and the table at the bottom shows the bearings between each factory and sensor.

A potential issue with this approach has to do with the meteorological data. As previously mentioned, the meteorological readings are taken every three hours starting at midnight. Whereas the sensor readings are taken every hour. If we look at the Wind Direction chart in appendix II, we can see the wind direction is constantly changing meaning the wind direction reading taken every three hours is not a good indication of the direction in the intermediate period and may introduce bias into the results if we use a nearest match approach for determining the wind direction for each sensor reading. Therefore, only sensor readings that fell on an hour with a corresponding meteorological reading were used in this part of the analysis.

Chart, bar chart

Description automatically generated The factory for each observation was then determined by selecting the factory with a bearing closest to the actual wind direction at the time of the observation. The plot to the right shows the total parts per million chemicals emitted from each factory. Radiance ColourTek is credited with emitting almost three times more chemical emissions volume than any other factory followed by Indigo Sol Boards, Kasios Office Furniture, and Roadrunner Fitness Electronics. As seen in appendix XII, the emission proportions are approximately the same for each chemical type. Although the most emitted chemical across all the factories was AGOC-3A.

Similar to how I checked that the chemical groups are significantly different I wanted to verify that the factory groups emissions are significantly different. Once again, this was done by conducting an ANOVA with the null hypothesis that the group mean between the factories is equal against the alternative that at least one group mean is different. With a p-value less than 0.01 I rejected the null hypothesis in favor of the alternative.

Next, I plotted the total emissions by factory over the course of the three-month period (see appendix XIII). As expected, the mean level of emissions from Radiance were the highest across the time range. However, there was a significant decrease in emission levels in December compared to the prior two months. The emissions from the other three factories remained consistent over the course of the three-month period.

Finally, I wanted to look at the time of day of emissions from each factory to see if there were any trends (see appendix XIV). The emissions from the Radiance factory peaked in the night between 10:00 pm and 3:00 am before decreasing in the daytime. This is a similar trend to the pattern observed for the level of Methylosmolene by hour. As previously noted, Radiance had the highest levels of Methylosmolene emissions. Therefore, it makes sense that the timing of Methylosmolene follows the peaks in the Radiance factory’s output. The Indigo and Kaisos emission peaked during the middle of the day like the AGOC-3A emission cycle, and the Roadrunner factory peaked at 12:00 pm but was mostly constant the remainder of the day.

**V - Results and Conclusion:**

The objective of this analysis was to help Mitch identify the chemicals present in the nature preserve that may be impacting the nesting population of birds as well as identify the responsible parties for those chemicals. Using anomaly detection, I determined one of the sensors being used by the government was not working properly and may need to be replaced. In terms of chemicals being released AGOC-3A was the most followed by Methylosmolene. Based on the provided information AGOC-3A is environmentally safe but Methylosmolene is harmful and may be contributing to the bird population loss. Finally, I used the wind bearing to identify the most likely sources of chemical emissions. Radiance ColourTek was identified as producing the most emissions for all types of chemicals.

With that said I believe that conclusion needs additional verification. One issue is that the sensor grid is only located on the western edge of the factories. Meaning if the wind is blowing in an easterly direction the emissions would not be counted by the grid. I would recommend that the government add air sensors to the east side to increase accuracy of the readings. Once that issue is resolved, I would suggest a potential next step could be implementing a regression-based model to predict chemical emissions levels. This would allow the local authorities to predict when each factory is going to exceed the regulatory levels of emissions.

Appendix I

Chart, treemap chart

Description automatically generated

Appendix II

Chart

Description automatically generated

Chart

Description automatically generated

Chart

Description automatically generated

Appendix III

Chart

Description automatically generated

Chart

Description automatically generated with low confidence

A picture containing chart

Description automatically generated

Appendix IV

Shape, square

Description automatically generated

Appendix V

A picture containing chart

Description automatically generated

A picture containing chart

Description automatically generated

A picture containing chart

Description automatically generated

A picture containing chart

Description automatically generated

A picture containing chart

Description automatically generated

A picture containing chart

Description automatically generated

Appendix V Continued

A picture containing chart

Description automatically generated

A picture containing chart

Description automatically generated

A picture containing chart

Description automatically generated

Appendix VI

Chart

Description automatically generated

Appendix VII

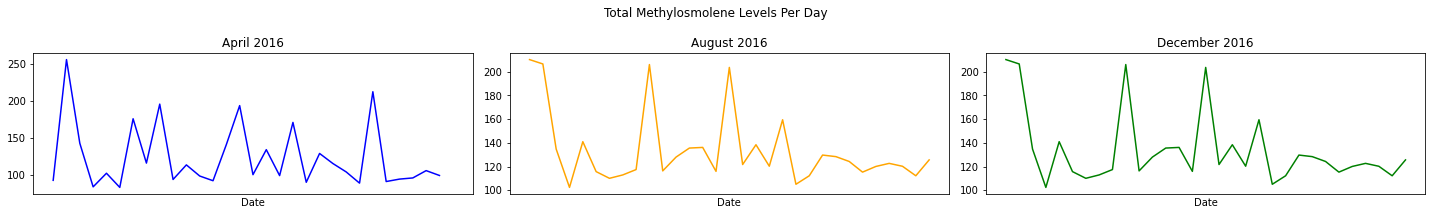
**K-means Isolation Forest**

Graphical user interface

Description automatically generatedGraphical user interface, application, Teams

Description automatically generated

Appendix VIII



Line chart

Description automatically generated with medium confidence A close-up of a logo

Description automatically generated with low confidence

A picture containing icon

Description automatically generated

Appendix VIIII

Chart, bar chart, histogram

Description automatically generated

Chart, bar chart

Description automatically generated

Appendix IX Continued

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

Appendix X

A picture containing icon

Description automatically generated

Icon

Description automatically generated

Chart

Description automatically generated with medium confidence

Icon

Description automatically generated

Appendix XI

Chart

Description automatically generated

**Bearings**

Table

Description automatically generated with medium confidence

Appendix XII

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

Appendix XII continued

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

Appendix XIII

Chart, line chart

Description automatically generated

A picture containing text

Description automatically generated

Chart, line chart

Description automatically generated

A picture containing shape

Description automatically generated

Appendix XIV

Icon

Description automatically generated with low confidence

Icon

Description automatically generated with medium confidence

Icon

Description automatically generated with low confidence

Chart, bar chart

Description automatically generated with medium confidence

References

Brownlee, J. (2020, August 14). How to Check if Time Series Data is Stationary with Python. Machine Learning Mastery. https://machinelearningmastery.com/time-series-data-stationary-python/.

Ellis, D. (2020, May 25). Calculating the Bearing between two geospatial coordinates. Medium. https://towardsdatascience.com/calculating-the-bearing-between-two-geospatial-coordinates-66203f57e4b4.

Jyenis, B. (2020, November 13). Anomaly Detection in Time Series Sensor Data. Medium. https://towardsdatascience.com/anomaly-detection-in-time-series-sensor-data-86fd52e62538.

Kizil, B. (2020, November 2). Introduction to Anomaly Detection in Time-Series Data and K-Means Clustering. Medium. https://medium.com/swlh/introduction-to-anomaly-detection-in-time-series-data-and-k-means-clustering-5832fb33d8cb.

Lewinson, E. (2019, September 26). Outlier Detection with Isolation Forest. Medium. https://towardsdatascience.com/outlier-detection-with-isolation-forest-3d190448d45e.

MLInterview. (2020, September 28). What are Isolation Forests? How to use them for Anomaly Detection? Ace the Data Science Interview! https://machinelearninginterview.com/topics/machine-learning/explain-isolation-forests-for-anomaly-detection/.

verma, P. (2021, April 12). Isolation Forest Algorithm for Anomaly Detection. Medium. https://heartbeat.fritz.ai/isolation-forest-algorithm-for-anomaly-detection-2a4abd347a5.