

COMPSCI 690V - Final Report

Mudit Bhargava(mbhargava@umass.edu)

Priyadarshi Rath(priyadarshir@umass.edu)

P.S: The report contains a detailed explanation about how we approached the problem, the challenges we faced, the solutions and a brief analysis of those solutions. A more detailed documentation of our solution and insights is presented alongside the plots within the python notebooks. Steps on how to generate the notebooks and any external packages required is attached as a PDF along with the code submission.

Demo Video: <https://youtu.be/CoMYrwgDE7c>

Overview: Mystery at the Wildlife Preserve

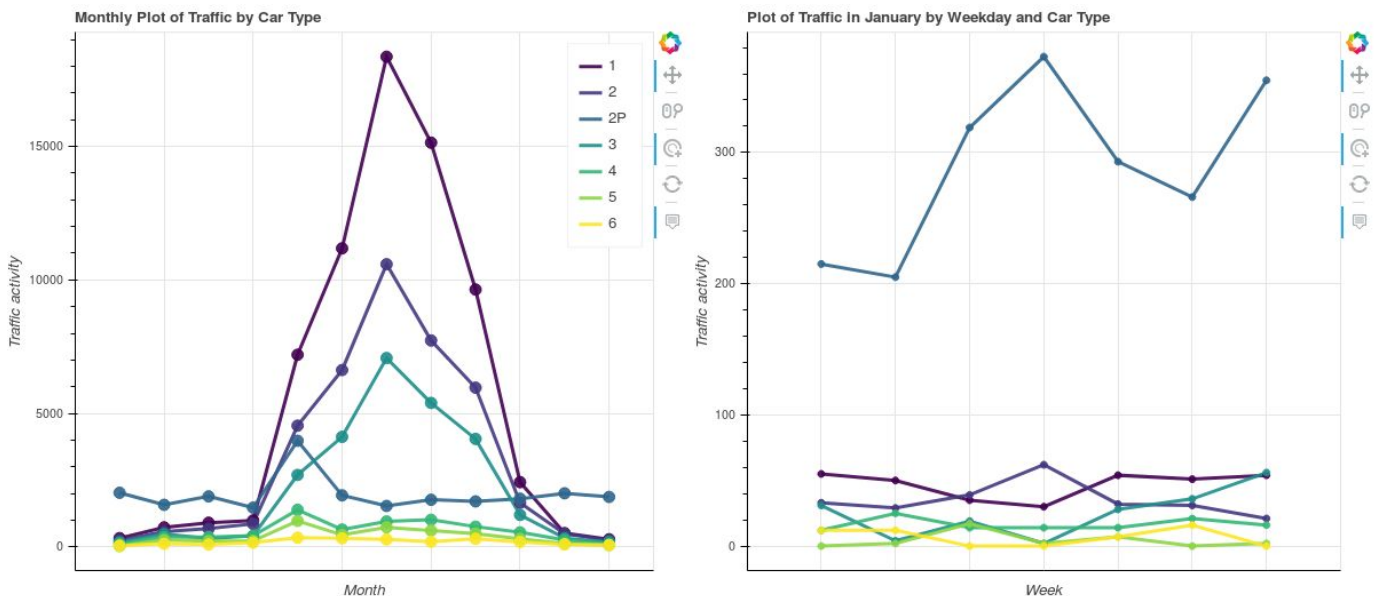
A mid-sized industrial town is located near a large nature preserve. Recently, the reserve has seen a significant decline of the population of the Rose-Crested Blue Pipit, a popular species of bird. Given several datasets of activity in and around the reserve, the task is to determine what factors there are that contribute to this decline.

Mini-Challenge 1: Traffic Activity in the Reserve

The mini-challenge looks at traffic RFID data from the reserve. The data was used to analyze daily patterns, patterns over a longer time and any anomalous patterns that may explain the declining bird population.

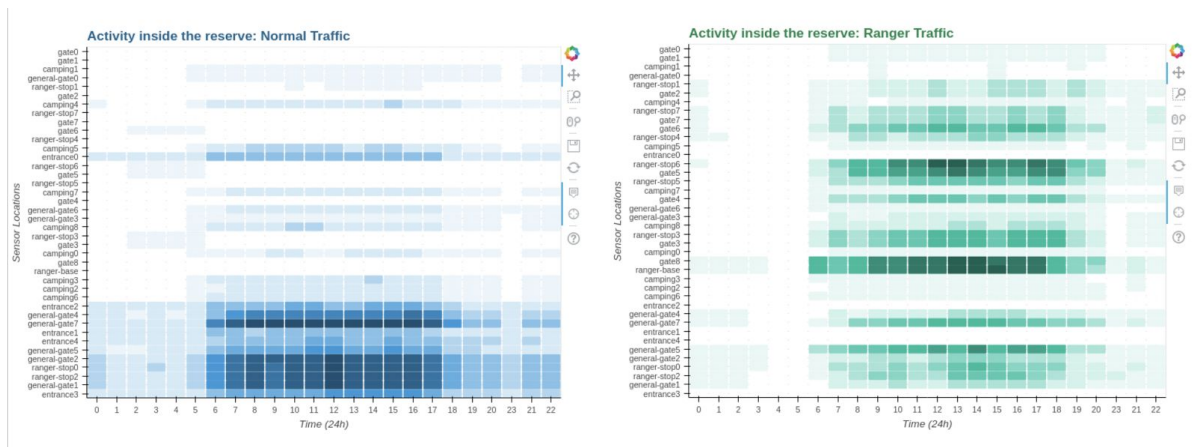
Traffic Activity over the year

The first idea was to get an overview of the density of traffic in the reserve. Since the traffic data spans one year, we decided to get an idea of monthly as well as weekly traffic patterns. This led to the plot shown below. From this visualisation, it is clear that the middle of the year(June-September) sees a lot more activity than the other months. Also, a general trend noticed was that there is typically more traffic during the weekend.

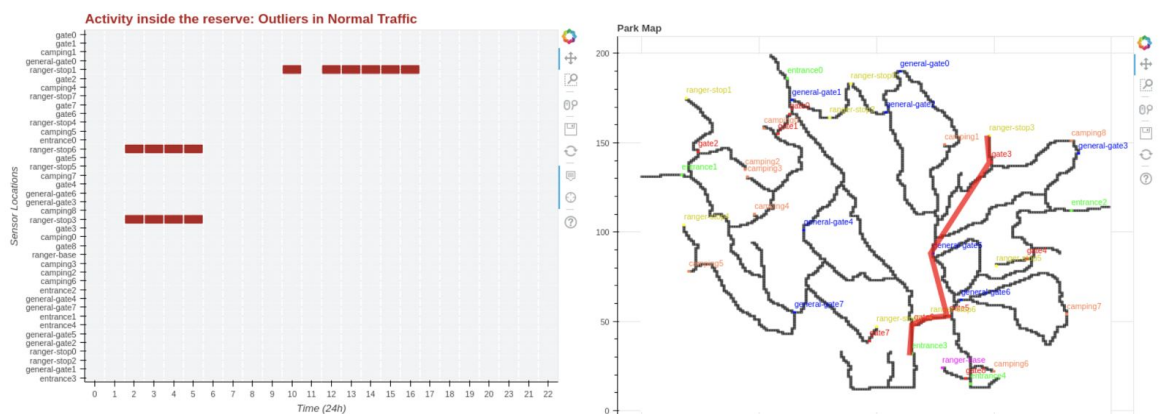


A Heatmap of Traffic Activity

We also wanted to understand the Traffic Activity inside the reserve over 24 hours in a day. This was done by plotting a heatmap of sensor locations v/s time of the day. The heatmap's intensity displays the average traffic near any sensor at any given hour of the day. The plot gave us a general idea about the traffic distribution for rangers and general public throughout the day.



In the normal traffic heatmap, we observed that there was some general traffic activity in areas that were restricted to public. This lead us to visualize a similar heatmap as above, but only for anomalous traffic activity (restricted zones).



The heatmap displayed 2 major anomalous activities, one between 10am to 5pm (possibly accidental) and another between 2am and 5am (strange). We further analyzed the anomalous midnight activities and found that all the vehicles were trucks that visited the preserve bi-weekly, took the same path (red line on map) and spent exactly an hour in the reserve.

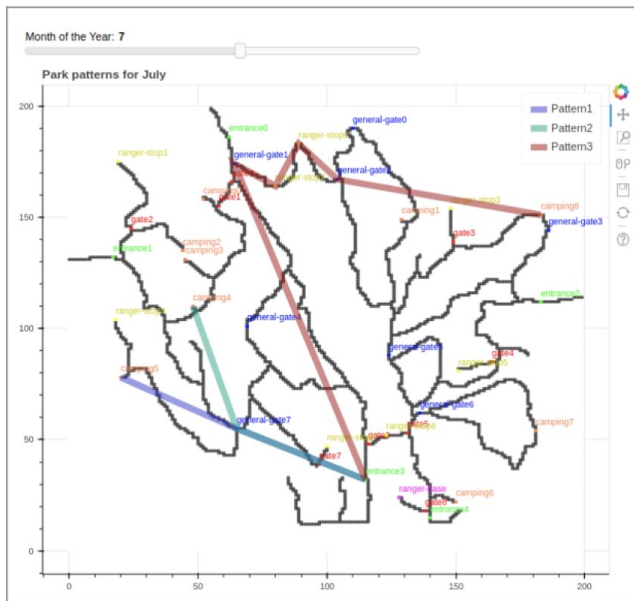
Challenges Encountered

Plotting a park map in bokeh was a slightly challenging task. Since we had to overlay other visualizations on top of the map, we had to draw out the map pixel by pixel using the Rect glyph.

Another challenge we faced was annotations of sensor locations on the map. The image file provided had color coded sensor locations e.g entrance (green), gates (red), campings (orange), but there was no information about the exact locations of the sensors. Hence we did not know which of the green dots corresponded to entrance1, entrance2 etc. After a little analysis, we found that there was a general trend in labelling of the sensors in the data. If you traverse the map along 0,0 -> 200,200 by rows and columns the labels appeared in descending order i.e ranger-stop7, ranger-stop6, ranger-stop5 ranger-stop0. Hence we were able to annotate all the sensor locations in a short amount of time

A Pattern of Life Analysis

To gain a better understanding of traffic activity within the reserve, we tried to understand people's movement patterns. For this we represented the movement of every car-id on each day as a separate document. The documents are clustered with DBSCAN using levenshtein distance (edit distance) as the metric. The edit distance tolerance is kept very low at 0.5 and the clustering is performed separately for each month. The intuition behind the low edit distance was that we want to find the most common movement patterns in the reserve. The clustering was done every month because based on the above analysis, we believe different months will give very different patterns.



The clustering and visualization of top 3 patterns gave us some interesting insights like in off-season months (Jan-May), the most common patterns are very short patterns from entrance to entrance signifying that people are mostly using the reserve as pass through. During the peak-season months (June-September), the trend is very different and the most common patterns are ones going from camping 4/5/8 to entrance gates. This also shows that people prefer to roam around the reserve before camping and tend to take the shortest exit path after camping is done

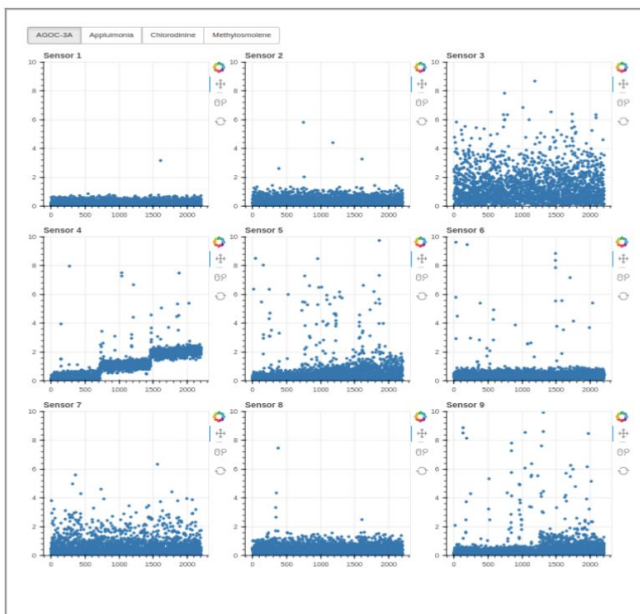
Challenges Encountered

One of the major challenges we faced in our clustering analysis was the number of points that needed to be clustered. Since every Vehicle-Id seen per day was considered as a separate document (~25,500) and a costly metric like edit distance was being used, clustering would have taken a very long time. Hence it made more sense, computationally and logically to break up the clustering analysis by month. Even then, peak season months had ~8000 documents and clustering could not be done on the fly during visualization. Hence edit distance metrics and top K movement patterns were pre-computed and stored as pickle files. The visualizations would simply load the patterns from the pickle and display it on the map.

Mini-Challenge 2: Industrialization

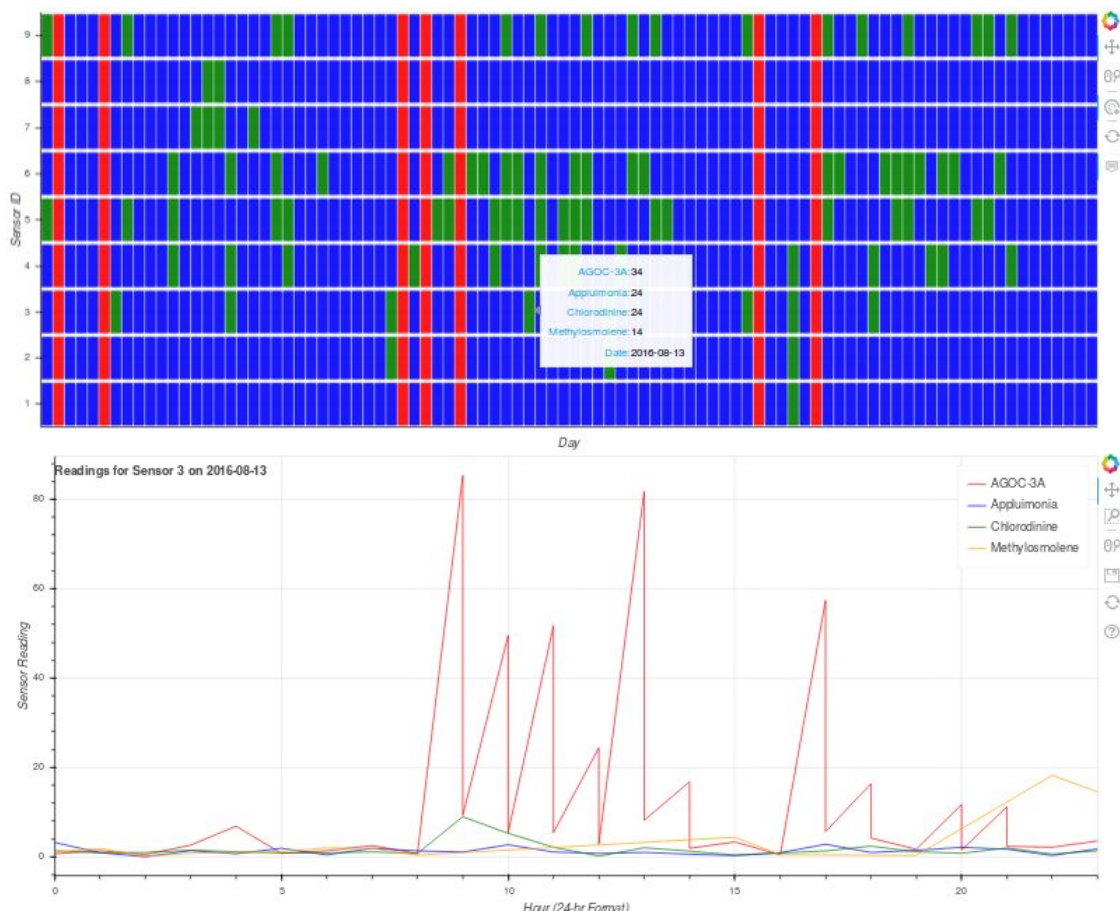
This mini-challenge looks at the effect increasing industrialization around the reserve. Mistford is home to 4 major companies that possibly release chemicals like **Methylosmolene**, **Chlorodinine**, **Appluimonia** and **AGOC-3A** (decreasing order of poisonous nature). We have been provided with chemical sensor readings from sensors located around the reserve and meteorological data on wind direction and wind speeds.

Data Exploration: The Sensors



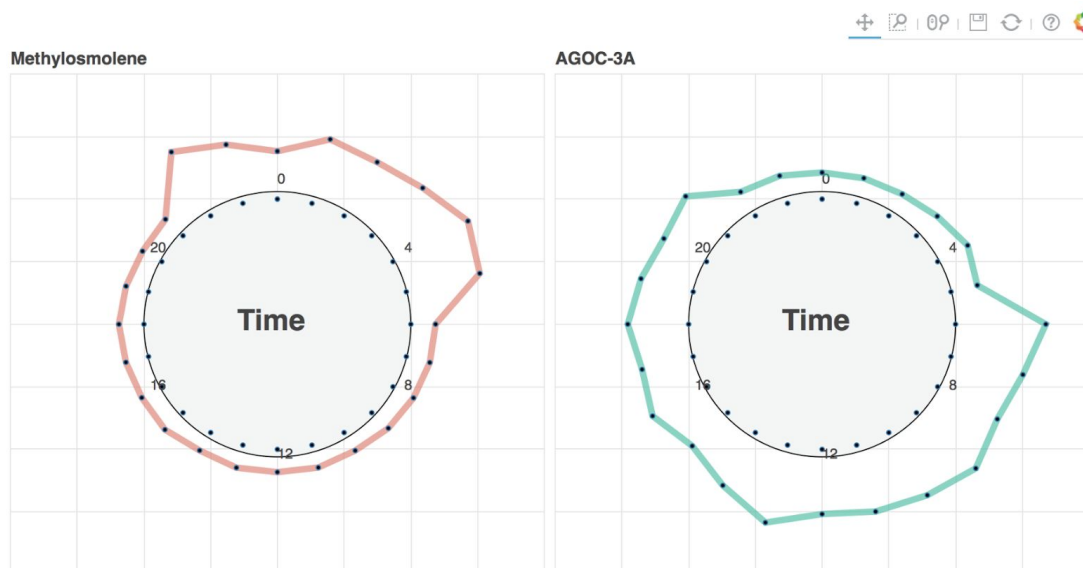
An important consideration in this mini-challenge was to identify whether the Sensors captured proper readings all the time. Looking at this grid layout of Sensor activity, a few interesting observations can be made. First, that Sensor 4 steadily "worsened" over time, as its readings show a constant shift in baseline readings. Sensors 5 and 9 also degrade over time, since the "noise" in their readings changes over time(not as sharply as Sensor 4, but still noticeable). Sensors 3 and 7 show a generally high amount of noise. Overall, the most reliable Sensors seem to be sensors 1,2,6, and 8.

The next step was to determine if there were any missing or anomalous readings captured by any sensor. So, we plotted the number of readings captured by each sensor over time in a heatmap. A snapshot of the plot is shown below. The color code is as follows: 24 readings per chemical per day(as is the normal case) is coded in blue. Less than this is coded red, and if there more than required, the activity is coded green. Also linked to the heatmap is a line plot, that shows the sensor activity for the day selected. The snapshot below shows that Sensor 3 captures conflicting readings for AGOC-3A at various times of the day.



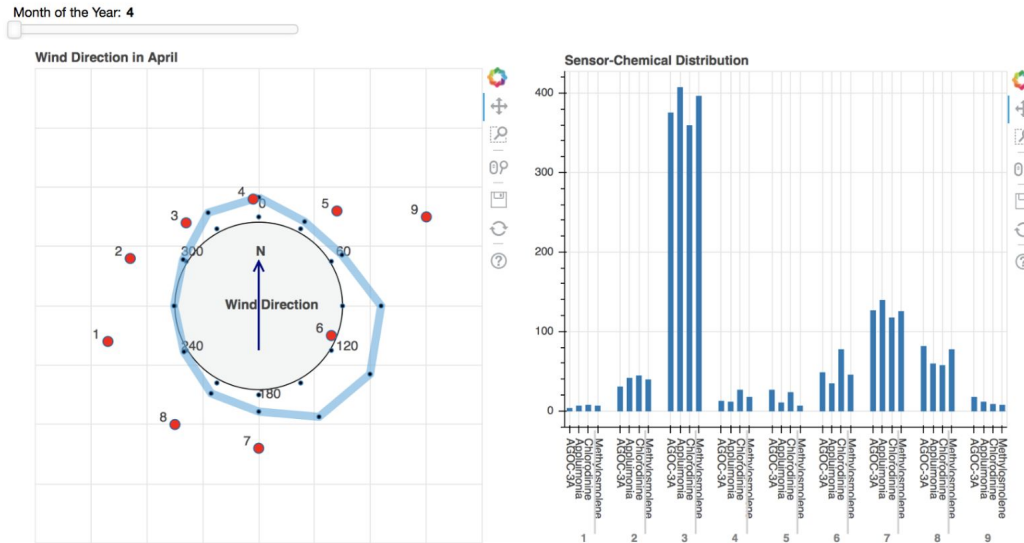
Data Exploration: The Chemicals

Now we turn our attention towards the chemical themselves and observe the general chemical release pattern throughout the day. For this we plotted a radial plot by time of the day. The distance of a point from the circle represents the magnitude of the average reading for the chemical at that time of the day.



It was surprising to see such a major difference between two VOC solvents. As per the problem descriptions AGOC-3A is a safer alternative to Methylosmolene. Such a major difference between the release patterns and major release of Methylosmolene in the wee hours of the night raises a lot of suspicions. It should be also noted that the release patterns of Methylosmolene coincide with the strange midnight truck activity found in MC1. Also as per the data description Methylosmolene is required to be neutralized and disposed off.

Next, we tried to understand the effect of wind on the sensor readings. For this we plotted a radial plot of wind direction distribution by month and distribution of chemical readings recorded by every sensor. Both the plots were linked to each other by slider (for different months). The plots showed that even though almost all chemicals are recorded by all sensors, wind direction plays an important role in sensor readings and their magnitude. For e.g sensor 1 has very few readings throughout the dataset, since the wind rarely blew in the south west direction



Challenges Encountered

One of the major challenges we faced in the above analysis was creation of Radial Plots. We couldn't find any out of the box radial libraries for Bokeh and hence had to create the plots from scratch. This involved working out the math to place all the points in right direction based on their magnitude and clockwise angle from vertical.

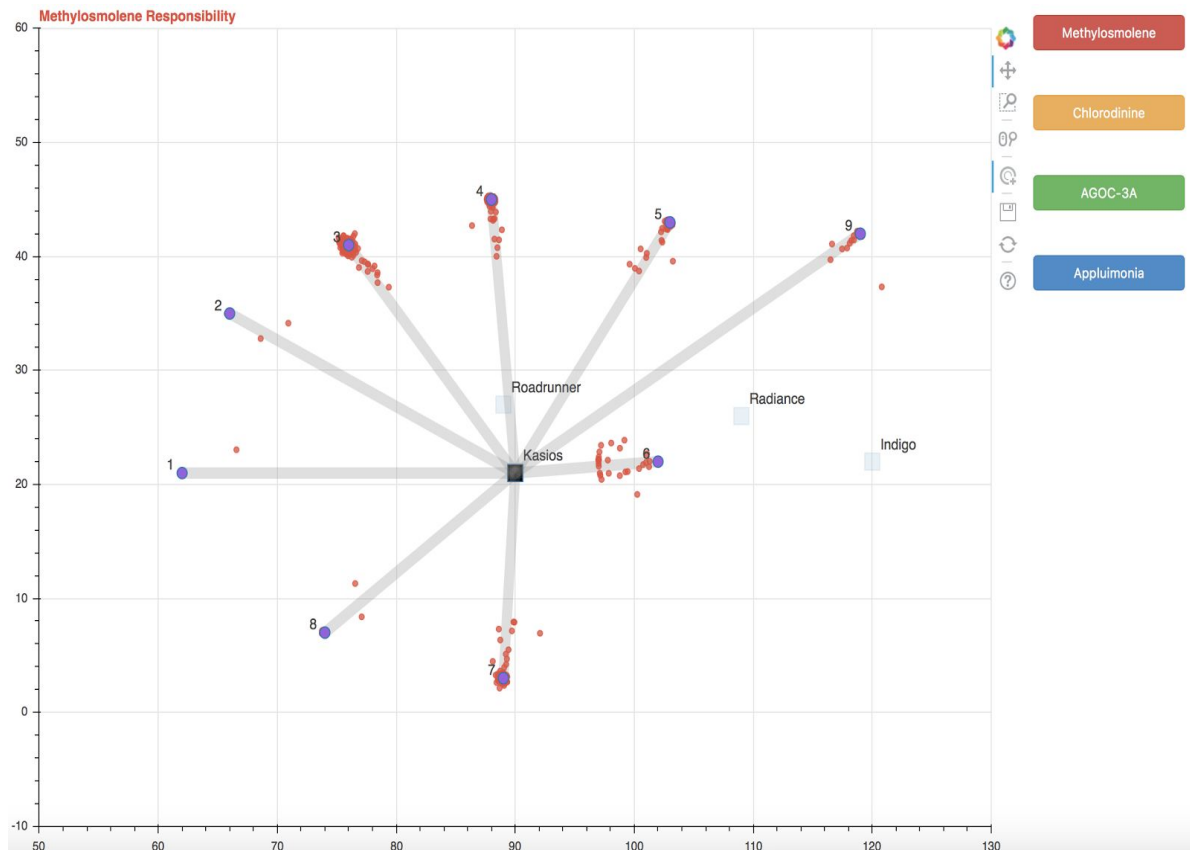
Another challenge we faced (and one we still haven't been able to resolve) is the distortion of radial plot when put alongside the bar chart. The problem seems to be with Bokeh forcing all objects to have the same sizing_mode within a row layout. We played around with different attributes of layout, but it did not solve the problem. Hence we were forced to display the 2 plots one below the other even though they are controlled by the slider.

Who's Responsible

Based on the above findings we now try and find which factories are responsible for which type of chemical release. For this, we plot the factories and sensors on a map and also plot the sensor readings on the map. The sensor readings are plot in the opposite direction of the wind flowing from the factories to sensors. The distance of the sensor reading point from the sensor represents the magnitude of the sensor reading. Overall this gives a trailing effect based on the wind direction. For instance in the below plot, it can be easily seen that almost all sensor readings for Methylosmolene are aligned in the direction of Kasios. Hence with a fair amount of confidence one can say that Kasios is responsible for the release of the most poisonous chemical, Methylosmolene. The plot was inspired by the one of the solutions to MC2 by Purdue University. The reason why we chose this plot is it because it gives a very clear and accurate picture of which factory may be responsible for the chemical releases.

Challenges Encountered

One of the challenges we faced while doing this analysis is joining of the sensor data with meteorological data. Both the datasets needed to be joined using time but the timestamps for both datasets were different. The chemical sensor data was recorded hourly, while the meteorological wind data was recorded every 3 hours. The two datasets were eventually joined by making an assumption that wind direction and speed remain constant between any two successive recordings of meteorological data.

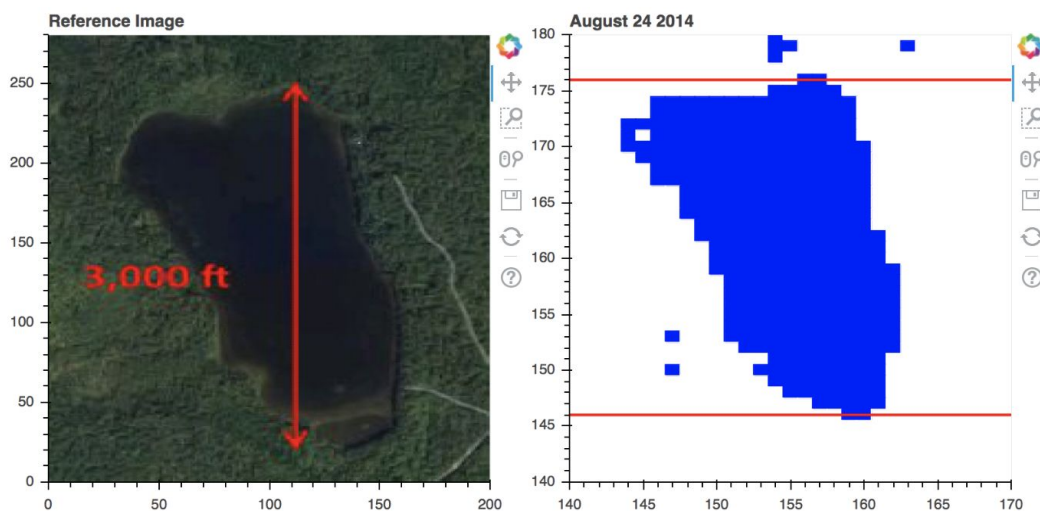


Mini-Challenge 3: Satellite Images

This mini challenge provides us with 3 years of multi-spectral images of the reserve. In the following analysis we explore these images to understand the changes in flora and fauna of the reserve and find possible anomalies or changes that may have lead to the declining population of the blue pipit bird.

Scale and Orientation

Our first task was to determine the scale and orientation of the satellite images. For reference we have been given image of Boonsong Lake (left plot) with its actual length. The plot on the right shows the satellite image of Boonsong lake. The two horizontal red lines mark the extreme ends of the lake. The red lines were found analytically using dynamic programming. Based on the pixel difference between the two red lines and the actual length of the lake, we found that 1 pixel corresponds to 30m in the real world. The location of Boonsong lake and orientation of satellite image was determined visually by comparing the left and the right plots



Challenges Encountered

We encountered a couple of major roadblocks while trying to solve the above problem. At first, we tried to approach this problem analytically. We looked into methods like Canny Edge detection and SIFT feature comparisons so that the reference image could automatically be matched with the satellite images and scale and orientation could be determined analytically. Unfortunately due to high learning curve for these methods and time constraints we weren't able to explore these methods further.

Another major challenge we encountered was the ability of Bokeh Server to handle large images. Each satellite image is a 650x650 image. When we plot more than 2 satellite images using bokeh server, the server was crashing with the following message "WebSocket connection closed: code=None, reason=None". We couldn't downsample the images, since that would affect the scale calculations. Hence we cropped the images as much as possible and displayed only the areas of interest.

Natural Features and their variations

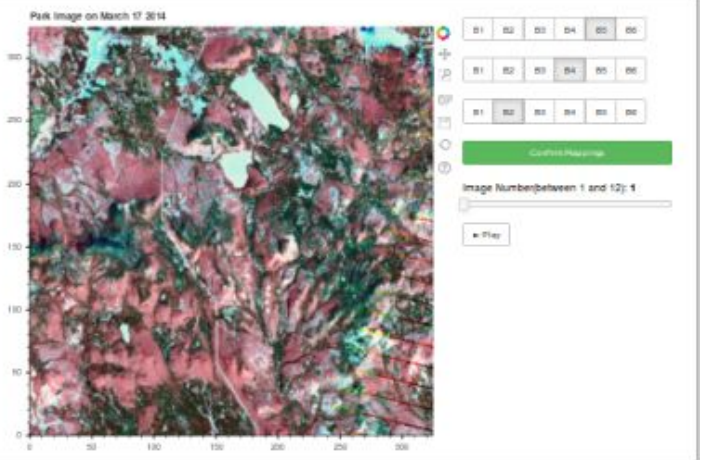
For this task, we created an interactive interface that allows the user to select color mappings as desired, and observe them. The interface is shown below:

The top row of buttons represents the color that is mapped to red, the middle represents a similar mapping for green and the bottom represents blue. So, if B5 is selected from the top row, B4 from the middle and B2 from the bottom, the mapping being plotted is (B5,B4,B2) -> (Red, Green, Blue).

There is also a play/pause button provided among the controls, which can help the user view changes across time.

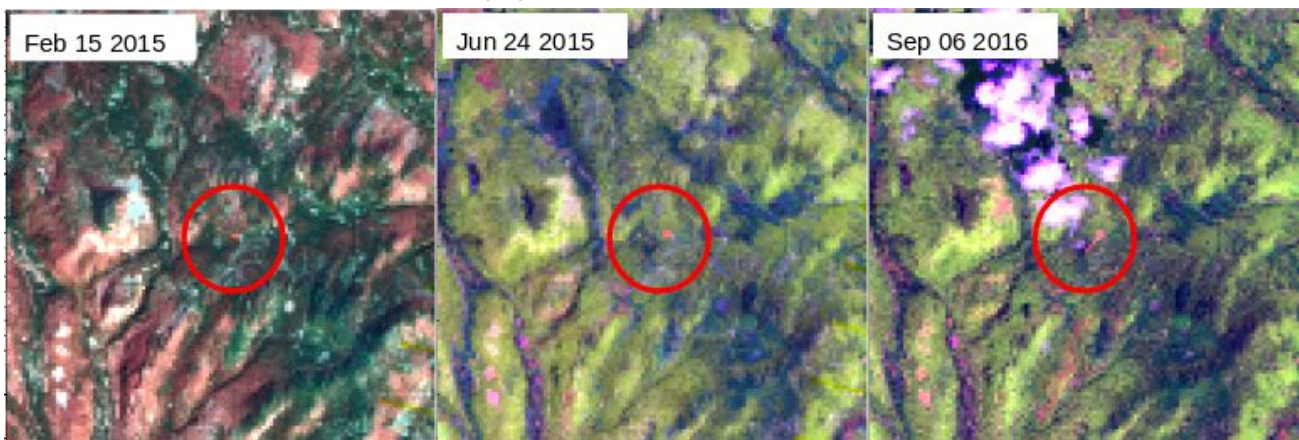
There are a number of natural features that can be seen in the reserve, like lakes, forests and townships.

Park image on March 17, 2014



Analysis:

There are some expected changes over time, like freezing of lakes over the winter, and an associated decrease of vegetation. One interesting thing to note is that the water bodies intermittently face some pollution. From MC1 and MC2, we know that an illegal and dangerous chemical is being produced and transported across the reserve. Since we are given the park map in MC1, we visually attempted to view the alleged dumping activity going on in the reserve. This showed a region that exhibited changes in features that were vastly different from expected seasonal variations. This region is shown over time, located at the center of each snapshot. Progressively, this small region in the center of the image grows, and is located near streams connected to the lakes. This is probably the cause of the lake pollution in mid-2015, and also the bird population decline.



Challenges Faced:

We were able to visually identify a suspicious region, but were unable to identify analytical approaches to obtain the same. We tried various methods like (a) analysing individual bands that were sensitive to soil mineral content; (b) analysing NDVI in order to obtain the effects of the region on local flora, but they weren't able to highlight any particular area that displayed strong presence of pollution.

Grand Challenge: Putting it all together

From Mini-Challenge 1, we concluded that there were some illegal vehicles in the reserve in the dead of the night, spending fixed amounts of time on a fixed route. Moreover, these vehicles are all of the same type, which is trucks. This pattern is too regular to be a set of isolated events, so this needs to be kept in mind.

From Mini-Challenge 2, we concluded that Kasios is regularly producing a highly dangerous chemical called Methylosmolene, in the nighttime. Also, given in the data for Mini-Challenge 2, Methylosmolene production releases toxic residue that needs to be dumped.

So, we know that Kasios produces illegal chemicals at night and that some unidentified trucks dump some chemicals in the night. Could Kasios be responsible for this?

Mini-Challenge 3 confirms that suspicion, since there is a region that can be seen in the multi-spectral images, which is not subject to usual seasonal variations. Also, this region lies within the reserve, possibly on the path the trucks took during their night trips. Since Kasios produces illegal Methylosmolene regularly and the dump also grows with time, our final conclusion is that Kasios is responsible for the decline of the Pipit. This claim is further enhanced when one tries to link the dump with the suspicious pollution that happens in the large lakes in the center of the image. The lakes are connected by streams, to the region in question, and the effect of said region on flora is clear. Thus, Kasios is the company that is causing the decline of the population of the bird, and action needs to be taken.