Suspense at the Wildlife Preserve: Like a Duck to Water

Kayla Strunk

Department of Computer Science, Denver University

COMP4449: Data Science Capstone

Dr. Claudio Delrieux

March 10, 2023

Introduction

Purpose and Significance of Project

The Ornithology Department at Mistford college has noticed that the mating pairs of the rose-crested blue pipit have decreased in the Boonsong Lekagul Wildlife preserve. There are four manufacturing firms near the wildlife preserve; however, the main suspect is a furniture manufacturing company, Kasios. Kasios is alleged to be dumping hazardous waste, specifically Methylosmoline, but they deny this allegation and advised that the area was inspected, and the quality of the soil does not differ from that of any other area within the preserve.

To gather evidence against Kasios the Ornithology Department went to the alleged dumping site to take soil samples, but the department found out that the site in question will be a new ranger station and new topsoil has been trucked in. Unable to get soil samples, the Ornithology Department was aided by the Hydrology Department; the Hydrology Department provided a dataset of water samples along with a map of the sampling locations. The water samples were analyzed to uncover any wrongdoing by Kasios or any other chemical contamination which may be affecting the wildlife in the preserve.

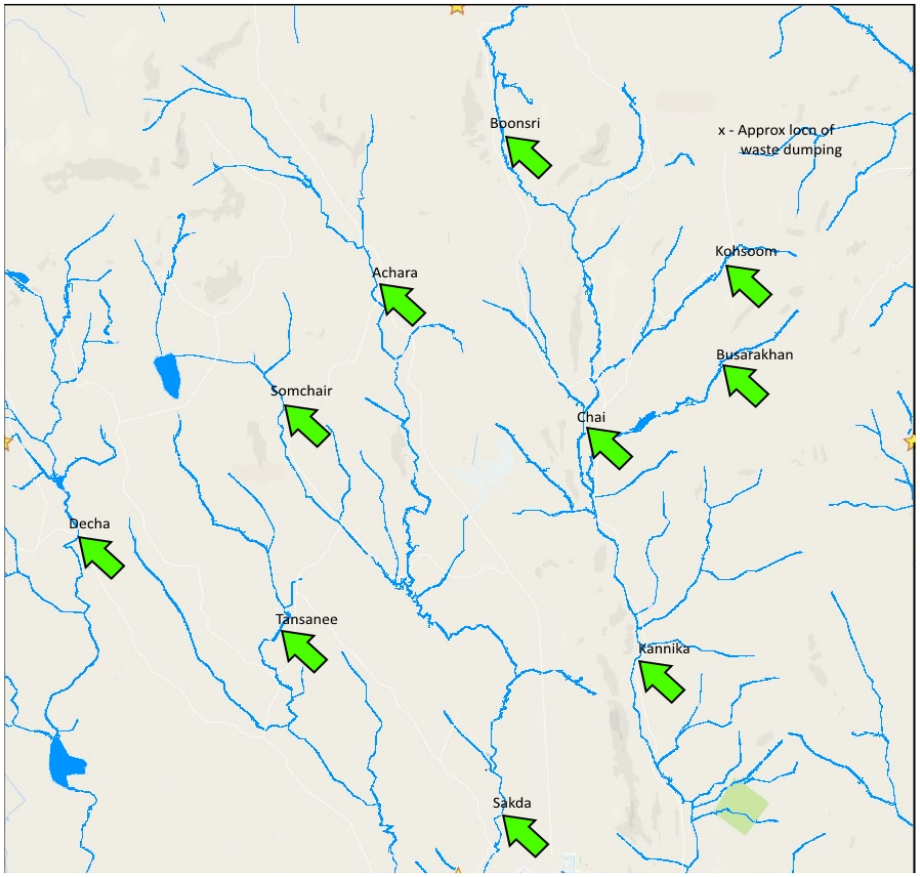
Goals

There were four main goals of this investigation. Characterize the past and most recent situation with respect to chemical contamination in the Boonsong Lekagul waterways—are there any trends of possible interest? What anomalies are found in the waterway samples, and how may they affect analysis of potential problems to the environment? Is the Hydrology Department collecting sufficient data to understand the situation across the preserve; what changes should be made to the sampling approach? After analysis, is there any concern for the rose-crested blue

pipit or other wildlife; what differences in sampling strategy may help in better understanding the situation at the wildlife preserve?

Description of Data set

The data provided by the Hydrology Department included one map.



As seen above, the map contains the 10 sampling locations as well as the alleged dumping site location; there are three main waterways in which the 10 sampling locations were taken from. Based on the map provided by the Hydrology Department it appears that Koshoom, Boonsri, Busarakhan, Chai, and Kannika sampling locations are all a part of the waterway which is closest to the alleged dumping site. If Kasios is indeed dumping toxic chemics at the site

indicated on the map, it should be expected to see spikes of these chemical levels mainly in Koshoom and/or Boonsri as well as smaller spikes in levels the further away sampling is done. The other five sampling sites, Achara, Somchair, Sakda, Tansanee, and Decha are located further away from the alleged dumping site, so increased levels of toxic chemicals may not be as apparent.

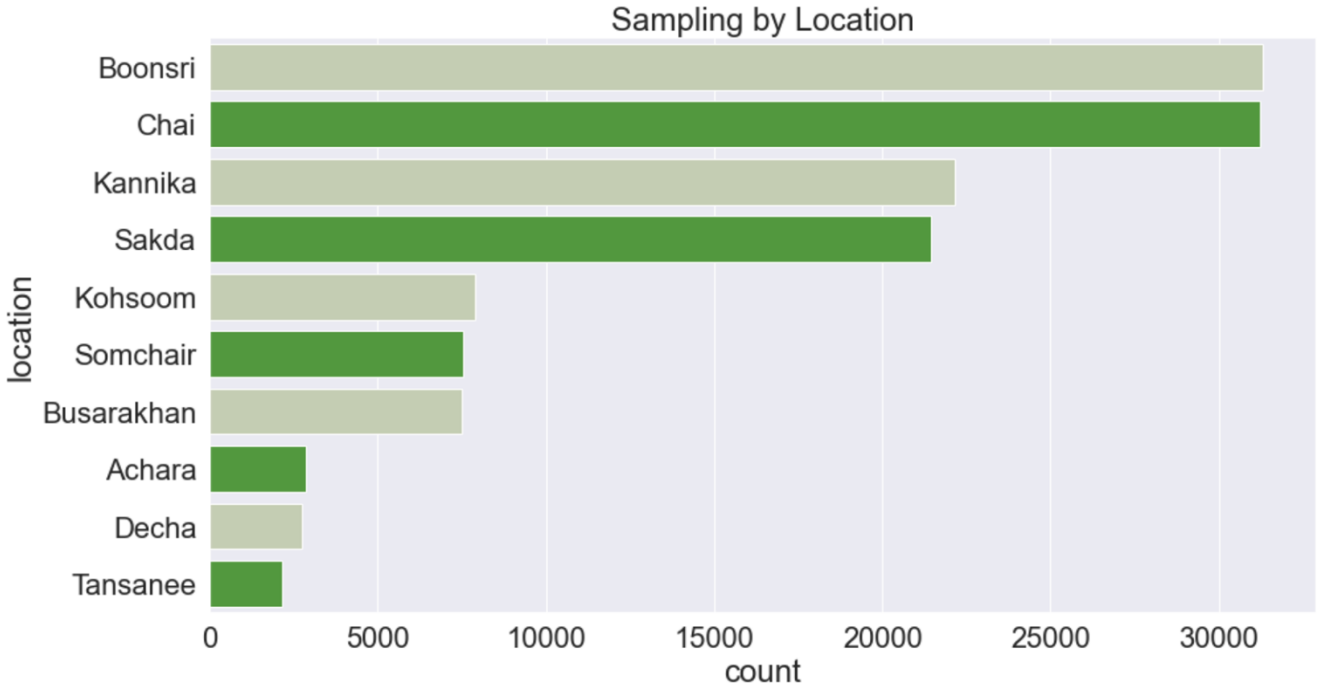
Along with the map of sampling sties, the Hydrology Department also provided two spreadsheets of data. The first spreadsheet contains the sampling data, and the second spreadsheet outlines the units in which each of the chemicals were measured. The samples dataset contains 136,824 samples taken between 1998 and 2016. The Hydrology Department sampled 106 unique chemicals from the 10 sampling sites during this time. All the samples were taken in milligrams per liter except for water temperature which was taken in degrees Celsius.

Data Preprocessing

Exploratory Data Analysis, and Visualizations

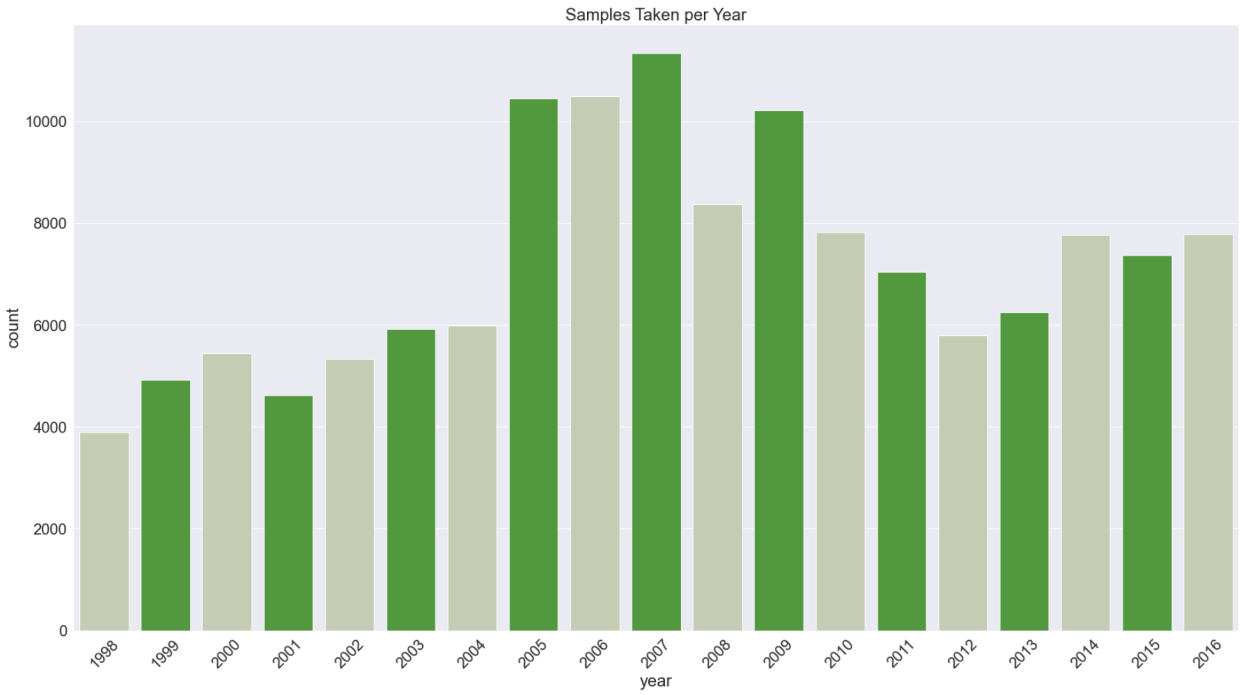
The first part of the overall analysis to be conducted is investigating the quality of the data received from the Hydrology Department; the quality will be determined through looking at the sampling strategy used by the department. Based on the count of samples for each chemical, it was readily apparent that each chemical was not equally sampled for. Water temperature was the most sampled with 5031 instances whereas PAHs were the least sampled with only 7 instances—the average number of samples taken was 1290. Because the chemicals were not sampled for equally, it may be difficult to determine the influence of the chemicals with lower sample sizes over time.

The sampling frequency at each location was also determined by obtaining a total count of samples taken at each location over the 19-year sampling time.



For this investigation, it again appears that the sampling strategy of the Hydrology Department is inconsistent; not all locations were sampled from equally. The Boonsri and Chai locations were sampled from the most with just over 30,000 samples whereas the Decha and Tansanee locations were sampled from less than 5,000 times. The locations that were sampled from more often may paint a better picture of any chemical contamination over time when compared to that of the locations sampled for less over time.

In addition to the amount each chemical was sampled with relation to the other chemicals and how often each location was sampled, the frequency in which the chemicals were sampled is also of importance.



Based on the graph above the number of samples taken over time varied as well. The number of samples taken was the lowest in 1998 and peaked in 2007. This may be due to the strategy of the Hydrology Department, or other outside factors such as funding available for the research purposes of obtaining these samples. To see if there were any discrepancies in sampling numbers month by month, this data was graphed as well (Appendix Figure A1 and A2). Overall, the samples taken per month did not differ greatly for each year; the total number of samples differed each year; however, the sampling numbers were consistent each month. The most noticeable differences were in October 2001, January 2002, and February 2014. For each of these months, the number of samples taken was much lower than samples taken for the rest of the year. This may be due to issues with equipment, or possibly not having enough people to take the samples.

Finally, because it is shown that each location and each chemical was not sampled for equally, a visual of the samples not taken was constructed. One heatmap was created for the

missing samples at each location, and another was created for the missing samples for each chemical. These visualizations make it very clear that there are a lot of 'missing' samples that could have been taken for each chemical or at each location. The missing samples are depicted by the lighter color in the heatmaps (Appendix Figure A3 and A4). This missingness could cloud the picture of trends overtime as we may see large increases or decreases that at first glance may appear anomalous, but with more data would not be out of the ordinary. The inconsistencies in sampling strategy were kept in mind when preparing the data to be run through models for outlier detection.

Data Preparation

The preparation of the dataset was minimal. The dataset did not contain any missing values and did not have any duplicated entries. Because the ID did not provide any useful information for analysis, this was removed. To better understand the timing of the samples, the year and month values were extracted from the sample date. In addition to extracting the year and month, the sample date was converted to a datetime object and set as the index. No features were removed; however, several instances were removed based on the number of samples taken per chemical. To ensure quality data for analysis and maintain an accurate picture of the chemical contamination over time, if the chemical was sampled less than 365 times it was removed from analysis. The threshold of 365 samples was chosen as this should allow for ample sampling during different conditions throughout the year. After these chemicals were removed, 65 of the original 106 chemicals remained for analysis.

Data Analysis

Model Building and Comparison

Once the preliminary analysis of the sampling strategies was completed two models were built to detect any anomalous samples during the 19-year sampling period. When building these models, it was kept in mind that not all the locations or chemicals were sampled for equally.

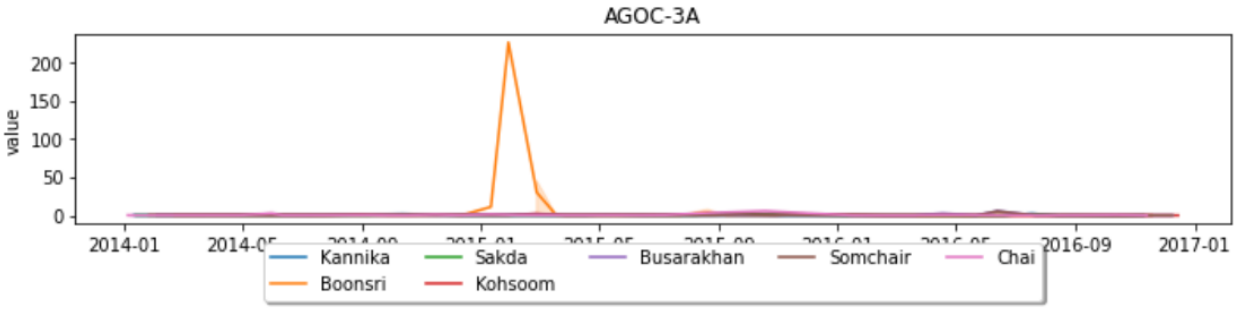
The first model constructed was a Local Outlier Factor (LOF) model. This model was chosen because it is built for detecting outliers in datasets with low dimensionality. The LOF model determines outliers in a dataset by comparing a data point to its neighbors, similar to a K-Means clustering approach. Because of the inconsistencies in sample collection, the contamination factor of the model was adjusted to 0.1 and the number of neighbors to compare to was set equal to one. These parameters were adjusted in attempt to account for the noise in the dataset and to reduce the number of instances marked as outliers. The LOF model detected 4,523 instances as being anomalous; these instances were spread across 48 chemicals. Due to the large number of chemicals detected as having anomalous instances, a second model was built for cross reference.

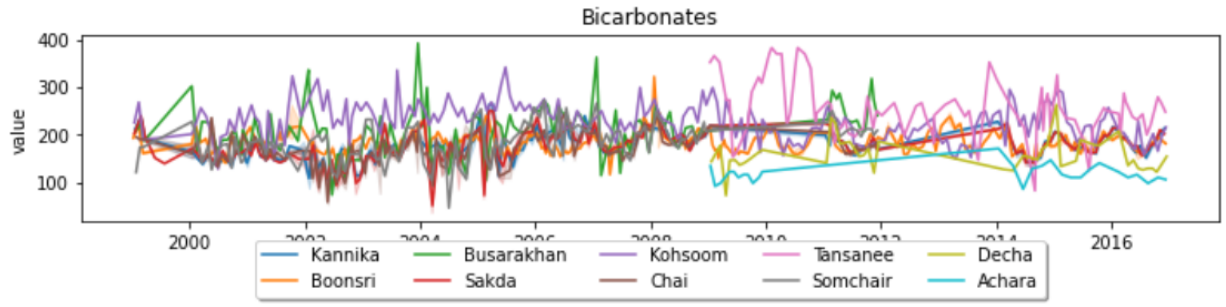
The second model to be built was the Isolation Forest model. The Isolation Forest Model works by scoring an instance by how many splits it takes to isolate the data point. The less splits it takes to isolate a point the higher anomaly score the data point receives. The contamination factor was also adjusted to 0.1 for this model. While the Isolation Forest model detected less chemicals as having anomalous instances (only 27) it did detect more instances as being anomalous (13,364). When cross referenced with the LOF model, it was found that the two models indicated 21 of the same chemicals as having anomalous instances. These 21 chemicals

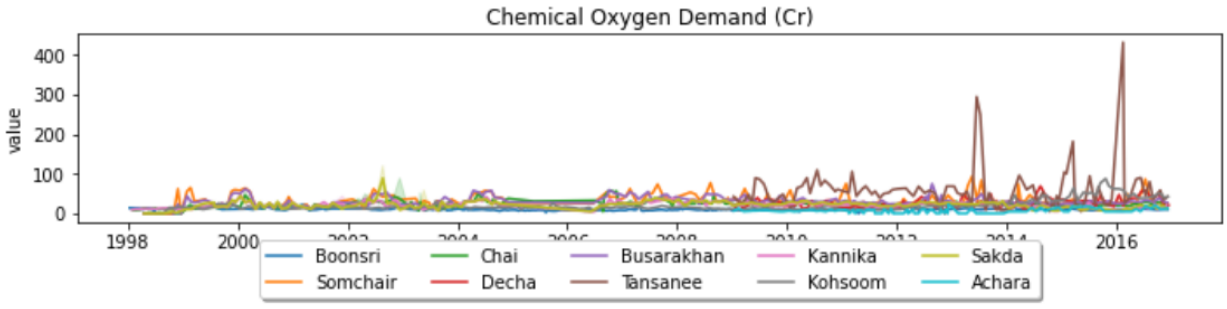
were further investigated for possible explanations to the issues being caused at the wildlife preserve.

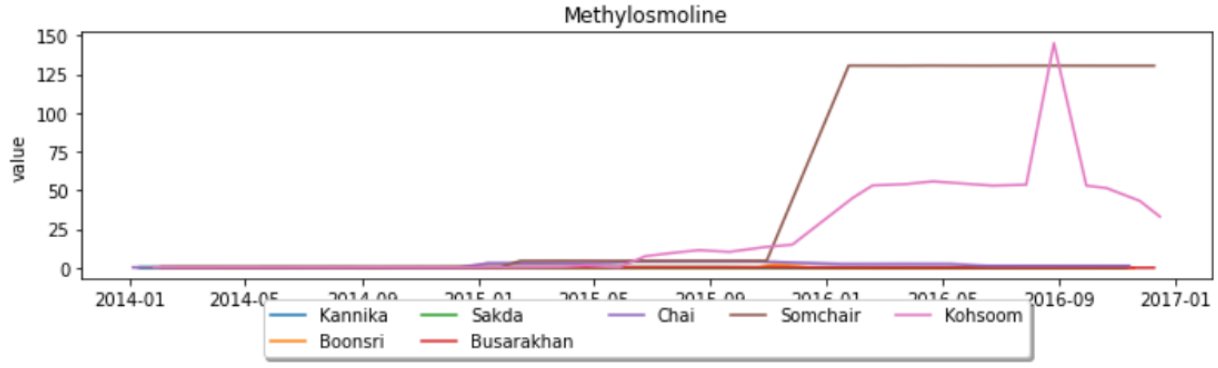
Although the timeframe of interest was over the past year (2016), the sampled values of the 21 chemicals were graphed over the 19-year timeframe for each location to provide a visualization as to when the spikes in measurement may have happened and what anomalies the models are picking up on.

The most notable spikes include AGOC-3 in Boonsri in 2015, Bicarbonates in Boonsri, Koshoom, and Tansanee in 2016, Chemical Oxygen demand in Tansanee and Koshoom in 2016 and Methylosmoline in Koshoom and Somchair in 2016.









The spikes in Bicarbonates, Chemical Oxygen demand and Methylosmoline- depicted above- all occurred within the timeframe of interest, so these were further investigated to see the affects elevated levels may have on wildlife. The elevated levels of Bicarbonates could result in an increase in the pH of water which is hazardous for soil and plants. Increases in Bicarbonates are the result of water running through limestone or dolomite rock. This increase in Bicarbonates could be an alternate explanation to issues with wildlife in the preserve and could be a result of the construction on the new ranger station. Increases in Chemical Oxygen Demand could be an indicator of decaying plant matter, human waste, or industrial waste—this increase supports the claim that Kasios could be illegally dumping their waste in the wildlife preserve. Finally, Methylosmoline, a toxic manufacturing chemical, and is the chemical Kasios is being suspected of dumping. Methylosmoline levels in all other locations are close to zero, so the increases in levels in Koshoom and Somchair is strong evidence that Kasios is dumping waste in the nature preserve.

The spike in AGOC-3A did not occur during the timeframe of interest but shows a similar pattern to that of the Methylosmoline spike in terms of samples at all other locations reading at or near zero. It may be of interest to the Ornithology Department to further investigate the affect that this chemical is having on the wildlife preserve as well.

Conclusion

Lessons Learned and Recommendations

Through the evaluation of the data provided by the Hydrology Department, there is strong evidence that indicates Kasios is dumping hazardous waste, and that it is affecting the wildlife in the preserve. Most notably there was a spike in levels of Methylosmoline found in the waters near Koshoom and Somchair. Because Koshoom is so close to the alleged dumping site, it provides strong evidence that the new ranger station is the location that Kasios was dumping its waste. There was also a small spike in Methylosmoline levels at the Chai sampling location which could indicate that the chemical is traveling through the waterways. The reason for the spike in Methylosmoline levels in Somchair is less clear as it is not connected to the same waterway as the Koshoom sampling location. This may be an indicator that Kasios has more than one dumping location within the wildlife preserve; it may be of interest of the Ornithology Department to gather soil samples from near Somchair to analyze further for additional contamination and support for claims of illegal dumping by Kasios.

Although the data provided by the Hydrology Department was able to provide insight into the disposal habits of Kasios, the inconsistencies in sampling did provide a lot of noise when it came to analysis. As not all the chemicals were sampled equally over time across locations, a slight increase or decrease in measurement may be picked up as an anomaly when it is an average measurement for the area. By sampling equally across locations for each chemical over time a clearer picture of the normal measurements for each chemical in that area and any anomalies would be more apparent. The analysis of the sampling strategies did not provide a clear picture as to why the Hydrology department is taking samples in the manner it is. It is possible that the sampling strategy may be the result of outside factors such as funding to the

Hydrology Department or even possible interference by Kasios. It may prove more useful for any future analysis to be done on water samples to target a smaller number of chemicals, or sampling locations. This would reduce the amount of effort needed to ensure equality across all measures.

Appendix A

Appendix A contains additional figures depicting the data sampling approach by the Mistford Hydrology Department.

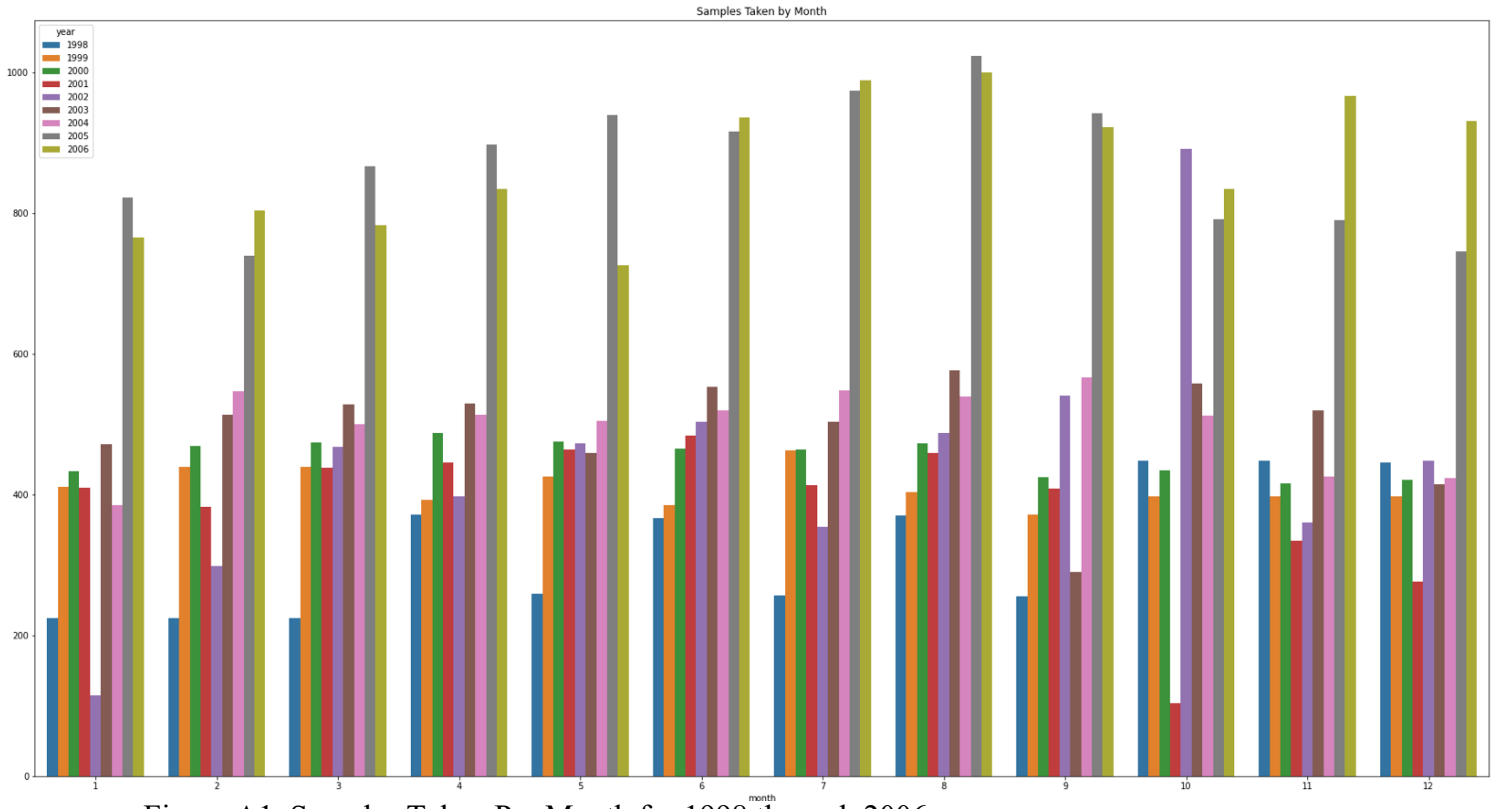


Figure A1. Samples Taken Per Month for 1998 through 2006

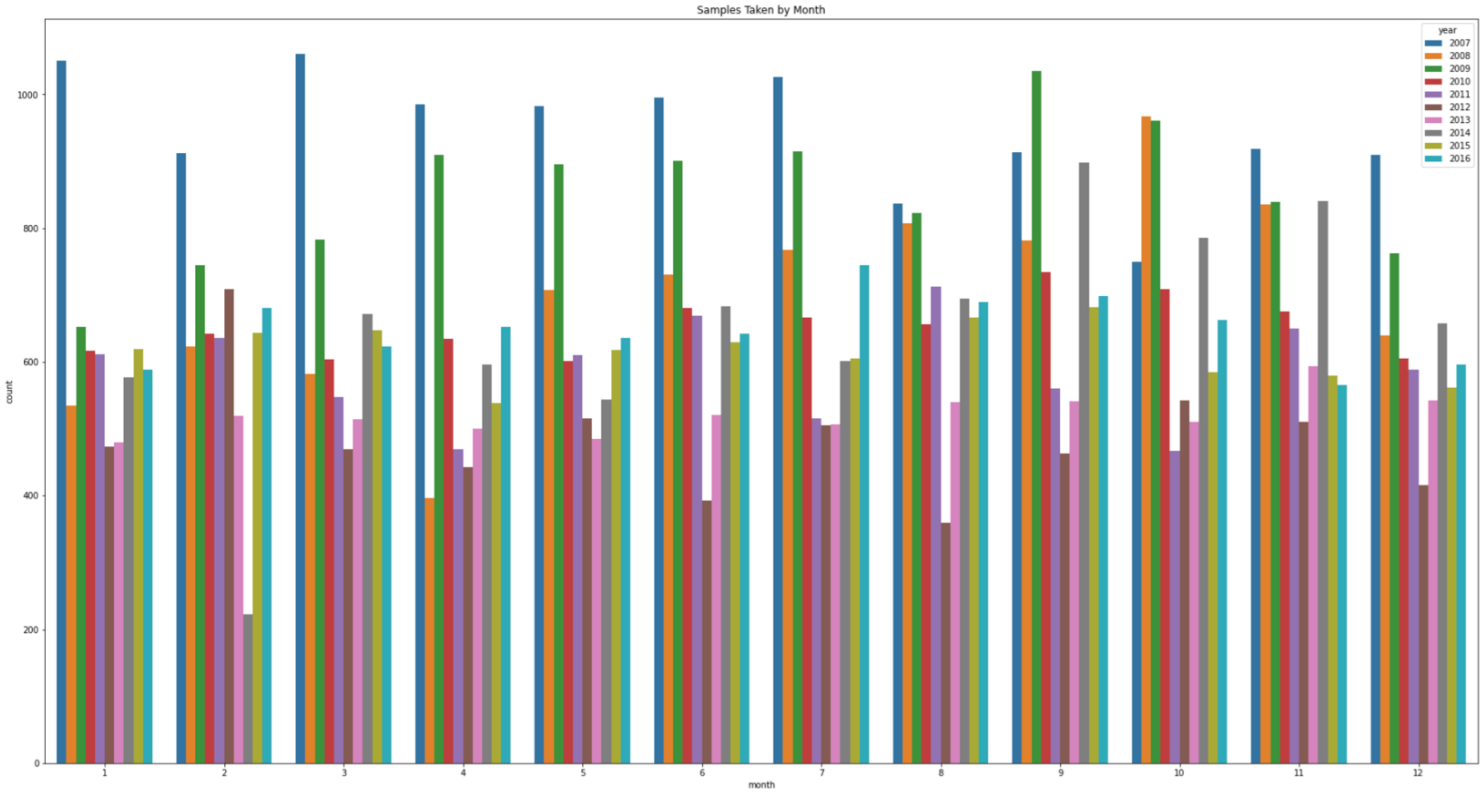


Figure A2. Samples Taken Per Month for 2007 through 2016

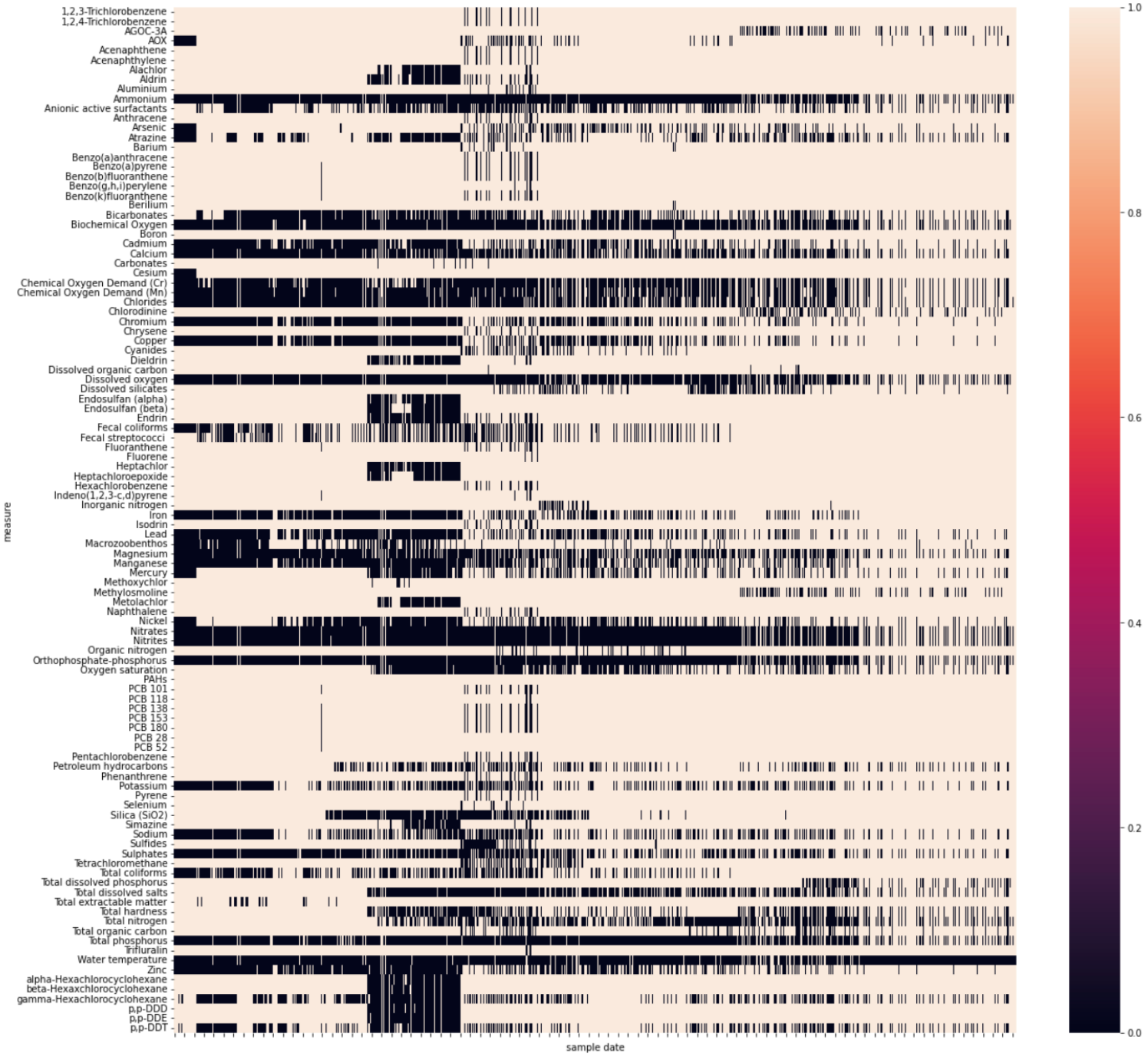


Figure A3. Missingness of Chemical Samples Between 1998 and 2016

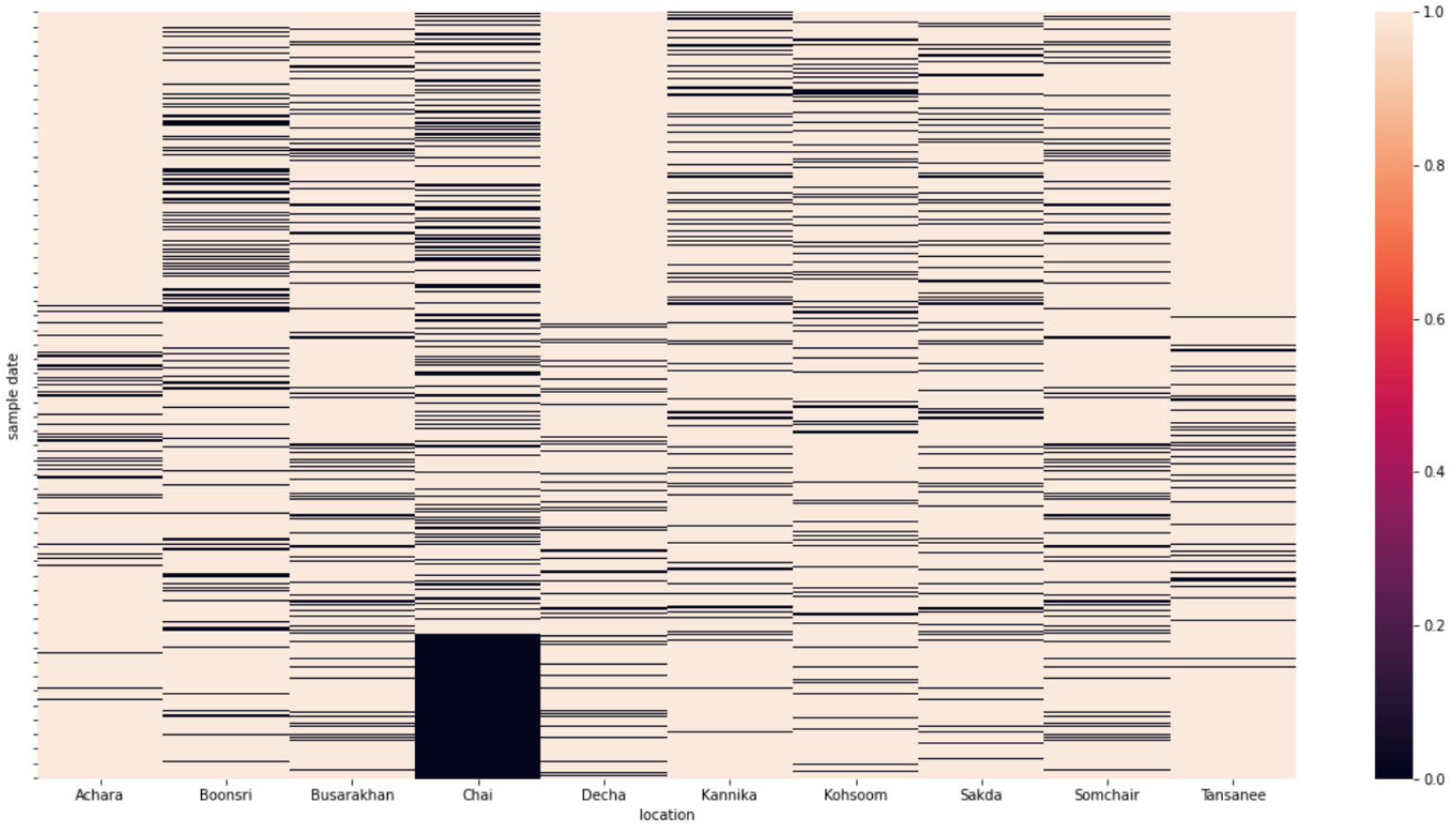


Figure A4. Missing Samples by Location from 1998 to 2016