What’s Your Water Quality? An Analysis of Water Quality in California

Team

For this project, the team consisted of *Sarthak Kar, Julian Ornelas, Max Ramirez, Kristina Stodder. Though this project was very involved and every member played a role in every step of the process, more specific breakdowns of group composition are as follows. Sarthak Kar did data cleaning and time series analysis.* Julian Ornelas *did time series analysis and visualization. Max Ramirez did data preprocessing and feature engineering. Kristina Stodder did data analysis and data visualization and worked on the website. The entire team worked on the report.*

Links

The link to the teams website is as follows:
https://kristinastodder.github.io/GEOG594/

The link to the video is as follows:
https://www.youtube.com/watch?v=FPJYCnqKMs4

Abstract

Bacterial contamination of water is of major concern in the state of California due to the abundance of bodies of water that are used for recreational purposes in the state. This paper will outline the process of analyzing water sampling data from water collection sites all across the states. Though data is collected in terms of specific bacterial levels within the water, the focus of this analysis is on the effect of bacterial contamination on recreational water safety. Accordingly, levels of bacteria measured at sites were compared to associated statistical threshold values to determine a binary classification of whether or not the water at that testing site was considered safe for recreational use or not. The aim of this study was to explore and model spatial and temporal relationships in recreational water quality. This analysis finds that higher levels of contamination and higher rates of water being classified as not safe occur around December and January, likely due to seasonal rain washing contaminants into the ocean. Additionally, the team has created a graduated symbol map which uses color to geographically visualize the current state of water quality in California. Finally, a time series forecasting model was fitted to the data to make predictions on the future state of water quality in California.

Introduction

From the beaches in San Diego to the crystalline waters of Lake Tahoe, California features some of the most magnificent bodies of water in the world. These are incredible places that make for great recreational water sites. However, when bacteria levels become too high, these bodies of water become closed for recreational use. The goal of this research is to identify spatial and temporal trends in water quality and beach closures across the state of California.

Various studies have identified a relationship between rainfall and poor water quality at beaches, lakes, and rivers (Lipp et al. 2001, Boehm et al. 2002, Dwight et al. 2002, Fukushima et al. 2021). This phenomenon is largely attributed to the transfer of contaminants (e.g. oil from roads, fertilizers from agriculture, pollution from urban development) from land into nearby waterways and downstream beaches (Dwight et al. 2002, U.S Geological Survey). In particular, the first heavy rain events of a season are thought to transport large quantities of nutrients and pollutants that contribute to the proliferation of unsafe levels of bacteria in the water (Orange County Coastkeeper 2016). As the state of California maintains a complex network of highways, vast agricultural lands, and a booming urban population, we expected to identify a correlation between rainfall and Poor water quality records throughout the state.

Data and preprocessing

The data for this project was collected from the California State Water Resources Control Board (CSWRCB) website. The 2020 to present dataset was used for this research for relevance of the data as well as for consistency in the water testing sites being utilized. The data contained over 300,000 observations with over 74 features included. Many of these features were deemed unnecessary, either due to irrelevance or incompleteness of the data. For the purposes of this project, the primary fields of the data that the analysis was focused on were as follows. Site Name and Site ID which identify the testing site of the observation. Sample Date established the date in which the observation occurred. Analyte identifies the bacteria being measured in the sample. Result is the actual measured value of the bacteria. Additionally, Latitude and Longitude values were given for each observation.

Literature Review and Research

Research was conducted to better understand the domain of aquatic bacteria. This was a necessary step in our process due to the nature of the data we used. The focus of our research was on better understanding what bacteria are relevant to water safety. From the data and from research, we found that E Coli, Enterococcus, Fecal Coliforms, and Total Coliforms are the concerning bacteria for recreational water safety. These bacteria were all measured in the data, however they were recorded with differing units of measurement, most probable number (mpn) /100 ml and coliform forming units (cfu)/100 ml. According to a table for Effluent Indicators for Bacteriological Indicators posted by the California State Water Resources Control Boards, these units of measurement are equivalent, thus no conversions needed to be accounted for. Additionally, research was done to determine the statistical thresholds for the measured bacteria levels. From the New Hampshire department of Environmental services and from the CSWRCB, the thresholds for E Coli, Enterococcus, and fecal coliforms are 88, 400, 104 mpn/100 ml. The threshold for total coliforms is dependent on the ratio of fecal coliforms to total coliforms. If the ratio is greater than 0.1, then the threshold is 1000 mpn/100 ml and if the ratio is less than 0.1, then the threshold is 10,000 mpn/100ml.

Methods

Feature Engineering

With our data collected and cleaned, the main step in feature engineering for the data was to insert the relevant threshold values for the bacteria and to then use those values to determine if an observation would be deemed unsafe for recreational water use. To do this, we created a feature of the ratio of fecal to total coliforms and using that created another column which contained the values for the thresholds related to the analyte present in the row. Another feature we added was the ratio of the measured value to the threshold value. This was done to quantify the severity of the bacteria levels. Then we created a “Safe” column to indicate whether or not the observed value of the bacteria was above the threshold. This was a binary value with 0 being where the sampling observed surpassed the threshold and 1 being where it did not. Additionally, the data was then grouped by water site and sample date to identify site samplings which included multiple observations. Then using the grouped data, if any of the “Safe” values for a site and date pair were deemed “unsafe”, the group was labeled as unsafe (0) or safe (1)

The resulting dataset was analyzed to identify any temporal patterns in water quality. “IsSafe” records with a value of 0 were considered “Poor” water quality, and “IsSafe” records with a value of 1 were considered “Fair” water quality. All descriptive statistics were carried out in R

Results

Temporal Analysis

After completion of preprocessing and feature engineering, the resulting dataset contained 359,722 records collected from 1,471 stations. The dataset consisted of 297,517 “Fair” water quality records (82.71%) and 62,205 “Poor” water quality records (17.29%).

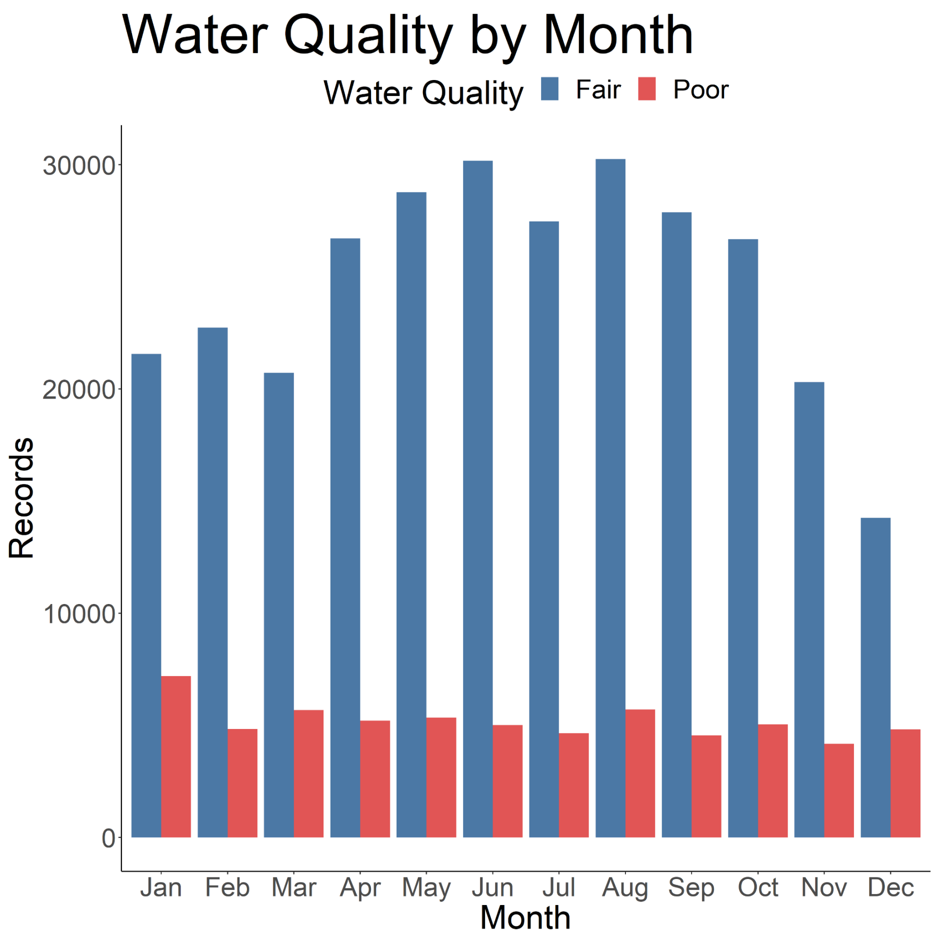
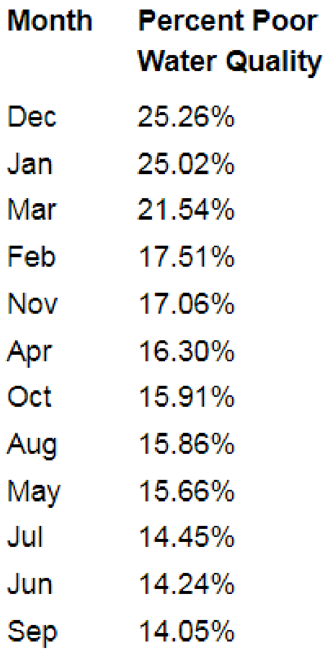


Figure 1: Counts of Fair and Poor water quality records by month

Further analysis showed that the highest number of records of Fair water quality occurred during warmer temperature months (i.e May - September) (see Figure 1). When examining Poor water quality, January (7,194) contained the highest number of records of Poor water quality, followed by August (5,703) and March (5,685) (see Figure 1).

Because each month was found to contain an unequal number of records, further analysis was performed to compare the percentage of Poor water quality recorded in a given month. December (25.26%) was determined to have the highest percentage of Poor water quality records, followed by January (25.02%), March (21.54%), February (17.51%), and November (17.06%) (see Table 1). Conversely, September (14.05%) was found to contain the highest percentage of Fair water quality, followed by June (14.24%), July (14.45%), and May (15.66%) (see Figure 2).



Month	Percent Poor Water Quality
Dec	25.26%
Jan	25.02%
Mar	21.54%
Feb	17.51%
Nov	17.06%
Apr	16.30%
Oct	15.91%
Aug	15.86%
May	15.66%
Jul	14.45%
Jun	14.24%
Sep	14.05%

Table 1: Ranking of month by percentage of poor water quality records

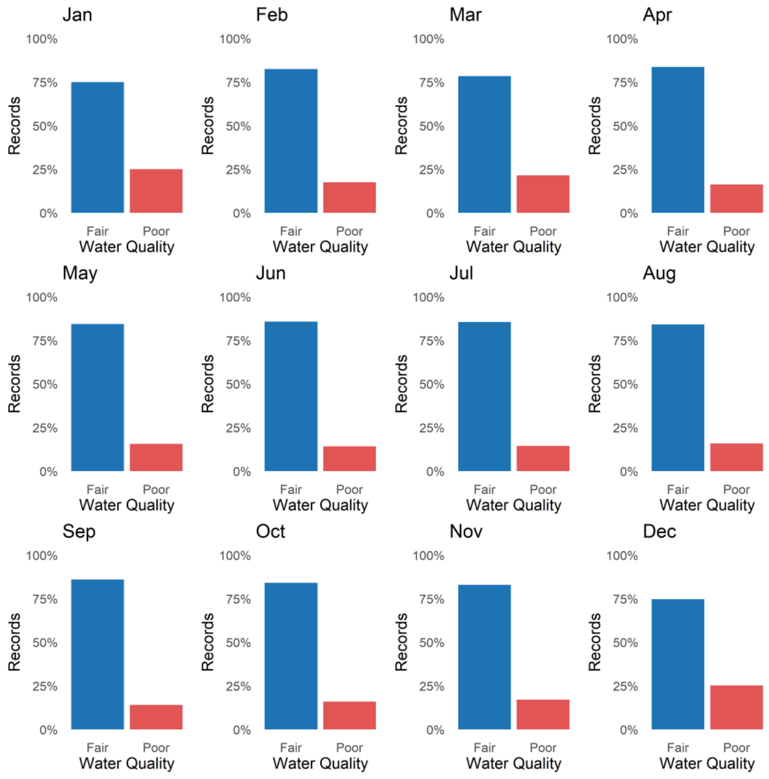


Figure 2: Percentage analysis of water quality by month

Spatial Analysis

In addition to temporal trends in the data, we wanted to look for any spatial tendencies factors that might drive bacteria levels and accordingly beach closures. To do this, we decided to visualize the data on a geographical level. Specifically, we filtered down the data into the most recent observation for each testing site and used that data to create a graduated circle map using tableau. This map is shown below in figure 3.

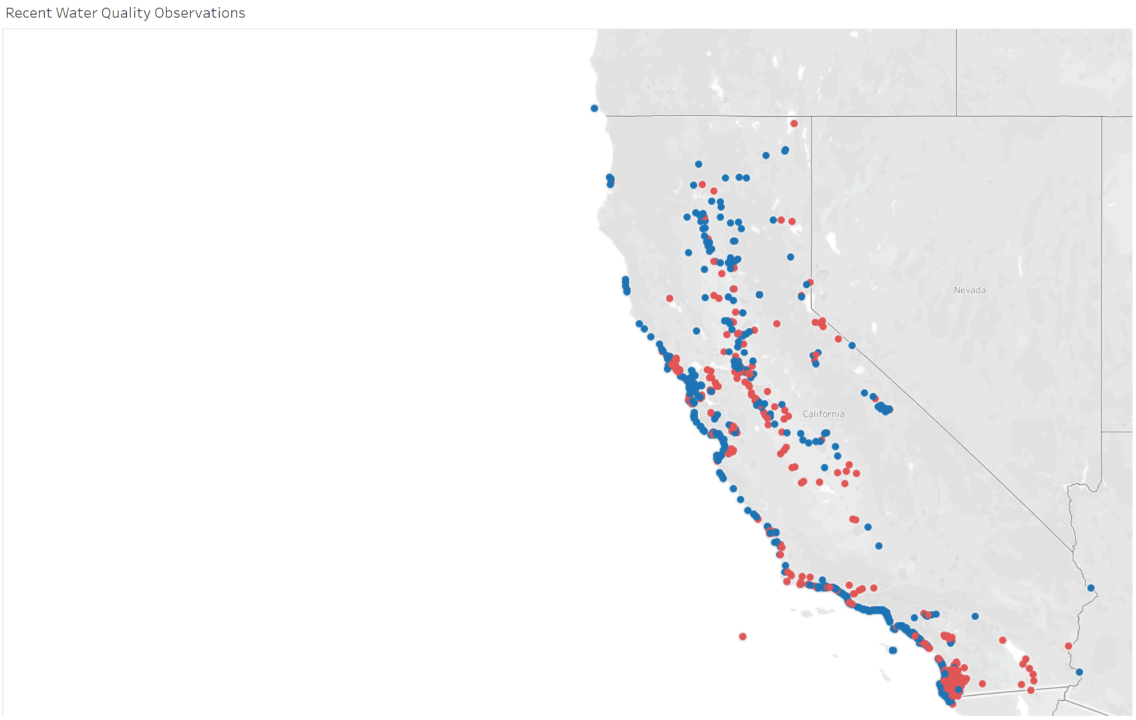


Figure 3. Most Recent Observation for each testing site

This map plots the water collection stations geographically using their latitude and longitude. Additionally, the color of the symbol is determined by the labeling of the water quality at the testing site. Here, the red circles indicate “poor” water quality while the blue circles represent a testing site with “fair” water quality. Using this map, we can visually see where clusters of unsafe water conditions are located. In this figure we can see that there is a dense cluster of “poor” water quality located to the south in the San Diego area.

Time Series Analysis

We wanted to create a forecast model based on our historical data to determine when the water quality at a particular location might return to a “fair” level. We chose to use logistic regression using the “IsSafe” attribute as our target variable since its values were binary, i.e. 0 and 1. This attribute was predicted using a one-lag shift on this same attribute. The historical data was split using a 70/30 train-test split and fed into the logistic regression model.

We then created a forecast for a period of 30 days using 14 lags (the last 14 IsSafe values) as input to predict each future value. An example graph for one of the stations is shown below in figure 4. The blue dots are the historical data and the red dots are the forecasted data. The stations we analyzed and forecasted all showed to be trending towards “poor” levels during the forecast period.

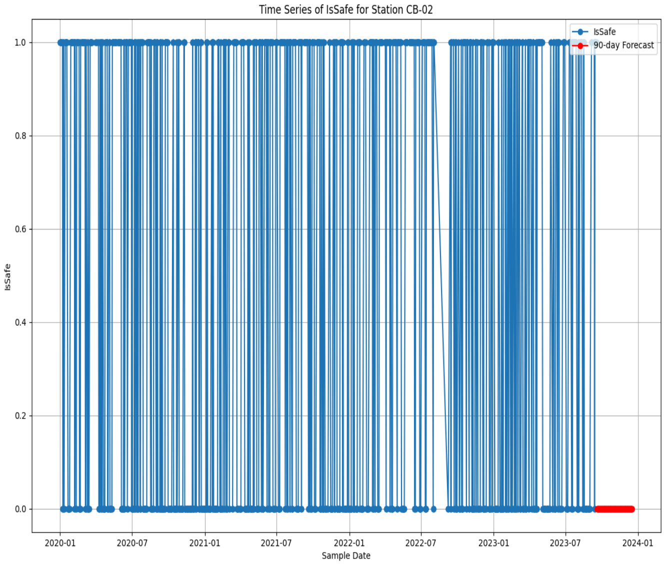


Figure 4. Logistic regression with historical values and 30 day forecast

Additionally, in this project, our approach also involved the utilization of the ARIMA (AutoRegressive Integrated Moving Average) model. Initially, our focus was on the Top 10 Stations, and we proceeded by comparing the behaviors of the average threshold ratio and the Issafe parameters.

To refine our analysis, we initiated the process by resampling the data on a monthly basis for both the Issafe and average threshold parameters. This strategic resampling not only facilitated a less noisy dataset but also condensed the original 1200 data points to a more concise set of 46 points as shown below in figure 5.

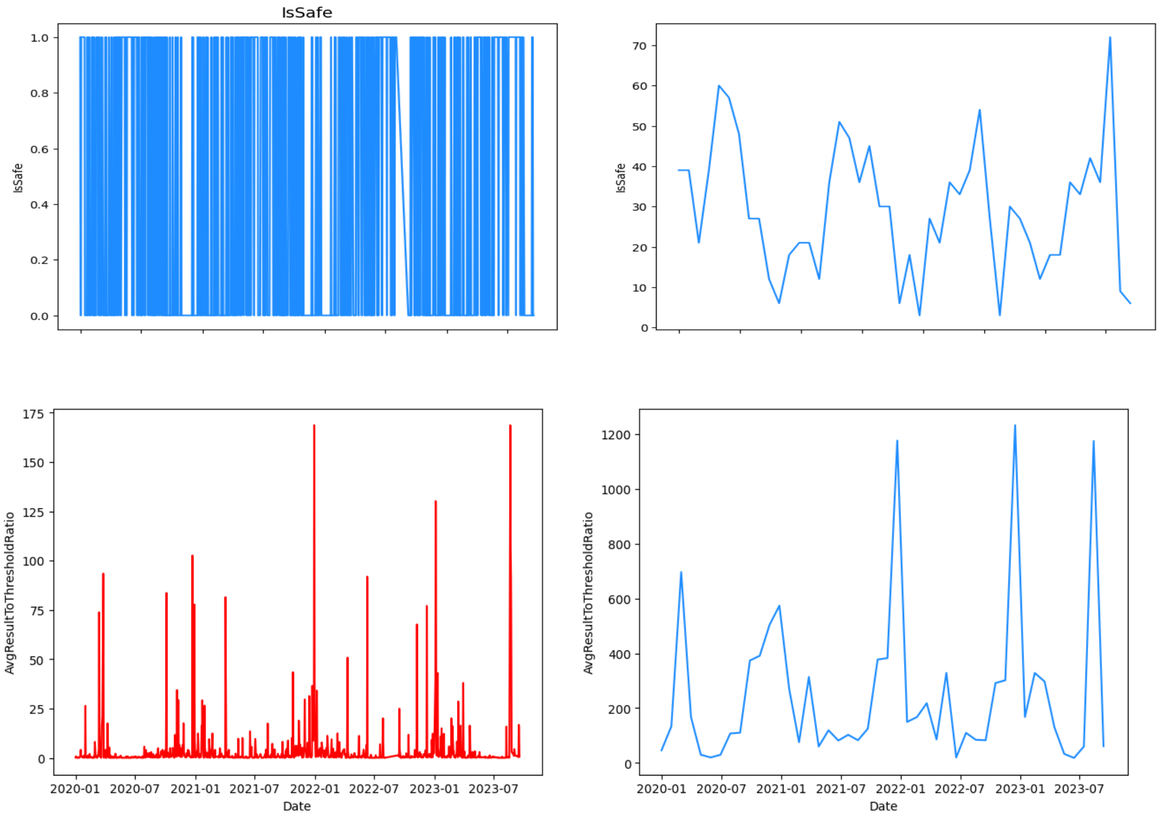


Figure 5. Resampled data for ARIMA model

Subsequently, to enhance the stationarity of the data, we undertook the deseasonalization of the dataset. Following this, our attention shifted towards addressing non-stationary components. To achieve this, we employed a Partial AutoCorrelation Test to identify and mitigate trends within the data. Also, an AutoCorrelation Test was conducted, allowing us to determine the presence of autocorrelation. By setting a lag of 1 and establishing a 95 percent confidence level, we successfully discerned patterns that significantly contributed to our predictive capabilities. These results can be seen in the following figure 6 and figure 7.

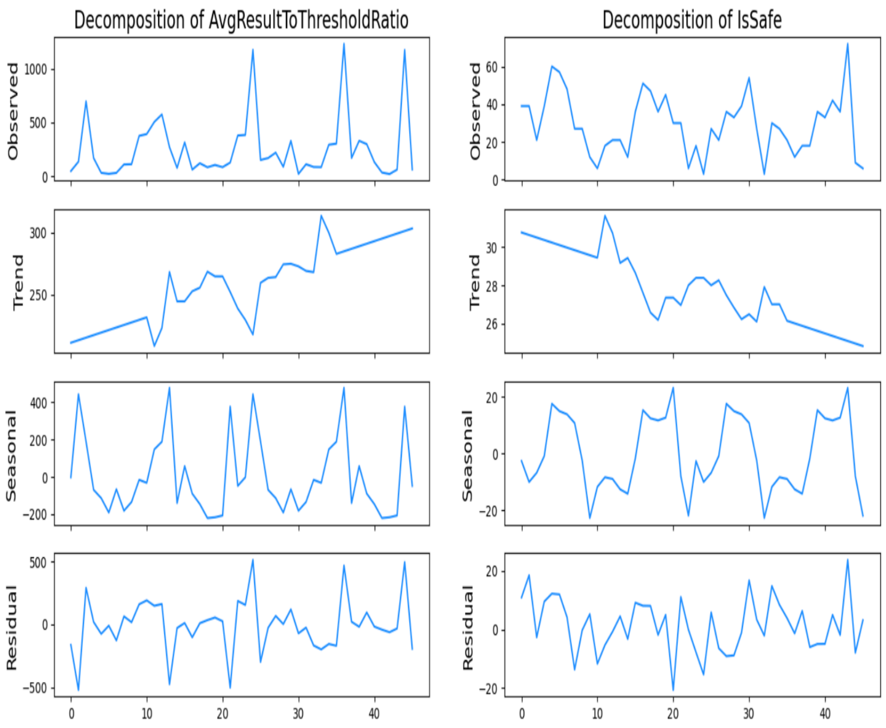


Figure 6. Time Series decomposition of resampled data

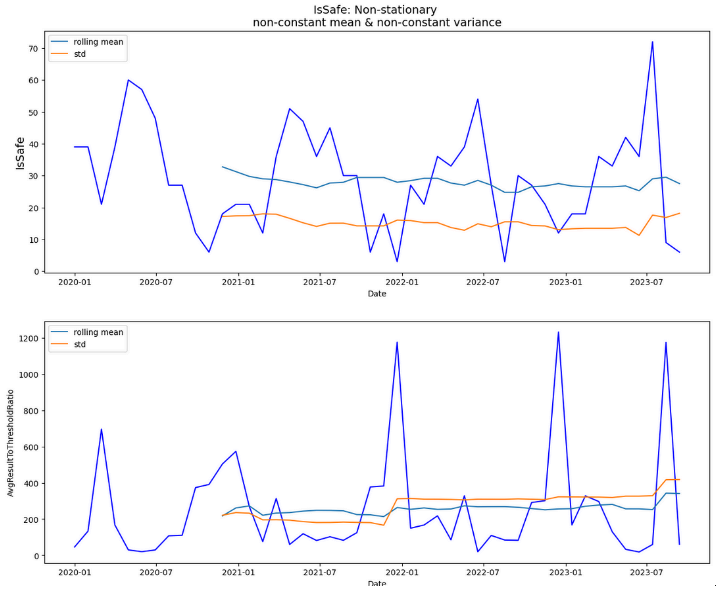


Figure 7. ARIMA forecast

Discussion

Temporal Analysis

Our results largely support the hypothesis that poor water quality is driven by rain events. We found that the highest percentage of Poor water quality records in California occurred in December and January, which is also when precipitation is highest throughout the state (Dettinger 2019). For example, in the Sacramento River Watershed region roughly 50% of the annual precipitation occurs from December to February, and 90% of the annual precipitation occurs from October to April (see Figure 8). Through our percentage analysis for water quality, we found that months from October to April were ranked as containing the highest percentages of poor water quality (see Table 1), and December, January, and February were ranked within the top 4 months with poor water quality (though March was ranked 3rd). These results suggest there may be a positive correlation between Poor water quality and rainfall in the Sacramento River Watershed region and throughout the state.

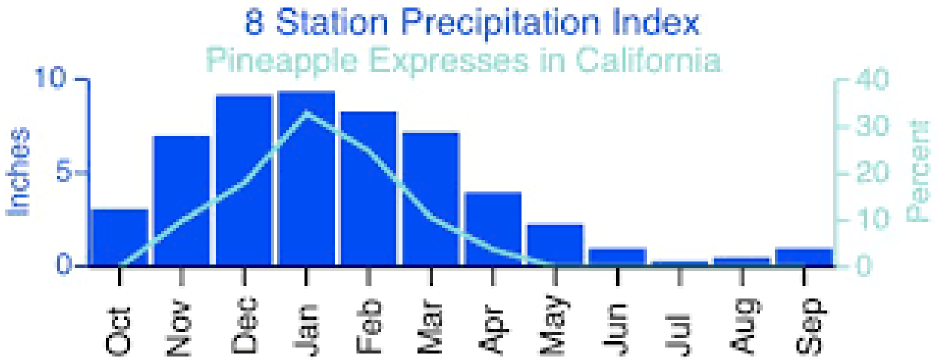


Figure 8: The monthly distribution of the Northern Sierra 8 Station Index and Pineapple Express (Dettinger 2019)

Additionally, we believe that there may be a higher percentage of Poor water quality records in December as opposed to February or March due to the accumulation of contaminants on land during the summer months. During the summer, there are few rain events to wash away contaminants and as a result they tend to accumulate on the land. However, during the first major rain event of a season, the high concentration of contaminants are transported from the land and can result in particularly polluted waterways. We believe that December may contain a particularly high percentage of poor water quality records because it occurs near the start of the rainy season, though further analysis should be conducted to substantiate this idea.

Spatial Analysis

As previously discussed in the results section, there is a clustering of “poor” water quality readings in Southern California, specifically centered in the San Diego area. There are a number of explanatory factors for this spatial trend, however we believe that one of the main factors driving this spatial distribution is the proximity to Mexico. Mexico notably has more lenient regulations regarding waste dumping than the United States. Due to less regulations on waste disposal, higher levels of contaminants enter the water in Mexico, leading to significantly higher levels of unsafe bacteria. This bacteria disperses in the water and as noted affects the water quality in the San Diego area.

Limitations

Statistical analysis was limited by available station data, as the number of water quality records varied each month (see Figure 9). In addition, station data for December 2023 was not incorporated in the dataset, further impacting our analysis. For this reason, we chose to perform statistical analysis that is largely descriptive in nature.

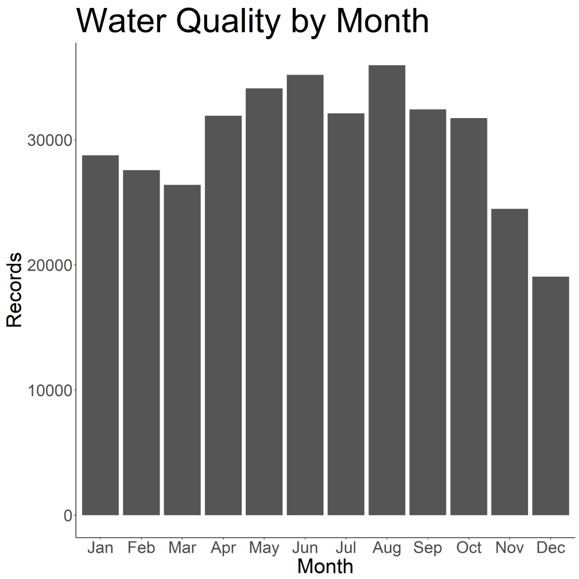


Figure 9: Total counts of water quality records by month

Future Directions

Future research should aim to perform temporal analysis across multiple years to determine if the observed trends are maintained in individual years. A stronger history of data would allow for the development of a more robust forecasting model with what we expect would be a strong seasonal component. Additionally, future research should try different levels of aggregation in the processing step. DIfferent levels of aggregation could lead to the potential for multiple seasonal components with different frequencies. Doing this would allow for a ensemble forecasting model where beach closures and water quality are forecasted at multiple time granularities and could allow for more precise forecasting.

Additionally, more data would mean more consistency across testing sites and would allow for a more robust spatial analysis. With the data present, it was difficult to truly quantify spatial factors. With more consistency across space, more complex spatial tendencies can be quantified. In addition, detailed weather data should be accessed for the location and date/time of each water quality record to determine if local weather conditions impact water quality.

Data Products

The following data products were used in this analysis.
Python was used in the process of data cleaning and preprocessing. Within python, the libraries that were used include *pandas*, *numpy*, and *scikit learn.*
R was used to perform statistical analysis and create visualizations. Within R, the libraries that were used include *dplyr*, *ggplot2*, *lubridate,* and *patchwork*. Tableau and ArcGIS Online were additionally used to create visualizations.

Additional data visualization tools can be accessed at: http://kristinastodder.github.io/GEOG594

Literature Cited/References

Boehm, A.B., S.B. Grant, J.H. Kim, S.L. Mowbray, C.D. McGee, C.D. Clark, D.M. Foley and D.E. Wellman. 2002. Decadal and shorter period variability and surf zone water quality at Huntington Beach, California. Environmental Science and Technology 36: 3885-3892.

Dettinger, M.D. 2019 California Precipitation. https://cwc.ca.gov/-/media/CWC-Website/Files/Documents/2019/08_August/Dettinger_CA_Precipitation.pdf

Dwight, Ryan & Semenza, Jan & Baker, Dean & Olson, Betty. (2002). Association of Urban Runoff with Coastal Water Quality in Orange County, California. Water environment research : a research publication of the Water Environment Federation. 74. 82-90. 10.2175/106143002X139776.

Fukushima, T., Kitamura, T. & Matsushita, B. Lake water quality observed after extreme rainfall events: implications for water quality affected by stormy runoff. *SN Appl. Sci.* **3**, 841 (2021). https://doi.org/10.1007/s42452-021-04823-x

Lipp, E.K., R. Kurz, R. Vincent, C. Rodriguez-Palacios, S. R. Farrah and J.B. Rose. 2001. The effects of seasonal variability and weather on microbial fecal pollution and enteric pathogens in a subtropical estuary. Estuaries 24: 266-276.

Orange County Coastkeeper. (2016). First Rain of the Season Means Orange County Water Quality. https://www.coastkeeper.org/first-rain-season-means-orange-county-water-quality/

U.S. Geological Survey. (n.d.). Nutrients and Eutrophication. https://www.usgs.gov/mission-areas/water-resources/science/nutrients-and-eutrophication

Surface water - indicator bacteria results - California open data. (n.d.). https://data.ca.gov/dataset/6723ab78-4530-4e97-ba5e-6ffd17a4c139

https://www.des.nh.gov/sites/g/files/ehbemt341/files/documents/2020-01/bb-14.pdf

https://www.waterboards.ca.gov/water_issues/programs/swamp/docs/cwt/guidance/3410.pdf