Bellingham Stormwater Data Analysis

Date: 2024-10-22  
Author: Gil Lund

Subject: Analysis of the Bellingham pollution monitoring data provided by Kirsten McDade at Re Sources on Oct 8th 2024.

# Background

This document focuses on analyzing e. coli bacteria data collected by Re Sources from the creeks and stormwater outfalls in Bellingham Bay and described in the following report (<https://www.re-sources.org/2024/06/three-years-of-bellingham-stormwater-monitoring-reveals-pollution-hotspots-including-taylor-dock/>) and mapped <https://maps.waterreporter.org/YWYyWhplWrDl/>.

This reports discusses three techniques for analyzing the data and determining how to perform further root cause analysis.

# Data Overview

The following summary describes the valid data points for e. coli bacteria measurements for the entire dataset (11 sites in Bellingham). The E-Coli column represents the bacteria levels in Colony Forming Units per 100 mL (CFU/100mL)

A screenshot of a black screen

Description automatically generated

## General Observations

The following observations document general characteristics of the dataset

Sample Date Range: 2020-12-03 to 2024-09-17 representing 45 sample month opportunities.

All sites were not sampled every month.

Sample sites: The following 15 sites were contained in the dataset. The site ‘Cedar’ was assumed to be ‘Cedar St.’ The rows ‘G06-1001’ , ‘G05-1002’, and ‘G11-1001’ were removed from the analysis.

'Bennett Ave', 'C St', 'Squalicum Creek', 'Cedar St', 'Olive St', 'Broadway St', 'Cornwall St', 'Whatcom Creek', 'Padden Creek', 'Little Squalicum Creek', 'Willow St', 'G06-1001', 'G05-1002', 'G11-1001', 'Cedar'

Total E.Coli measurements: 458

Duplicate Readings (Same Sample Date and Site): 66

TODO: Why were duplicate readings generated?

A screenshot of a computer

Description automatically generated

Analysis applied geometric mean to duplicate values

Min Value: 0  
Max Value: 20,480 CFU / 100ml

TODO: What is valid min/max values?

Bacteria threshold = 310 CFU / 100 mL

## Data Distribution Analysis

To get an impression for the distributions which fit the raw dataset a few visualizations (boxplot, complot, log-normal fit) were performed

Box and whisker and cum prob plots were generated to get an overall impression for the distribution of the dataset.

A graph with a box and a line

Description automatically generated with medium confidence

Figure - Overall Box and Whisker Plot

Long positive tail indicates significant outliers. Zooming in to see the box inter quartile range shows the following.

A graph with a blue square and black lines

Description automatically generated

Figure - Zoomed Box and Whisker Plot

Note: Q3 (75%) is above bacteria exceedance threshold of 310 CFU/100mL showing that there is a baseline exceedance issue with the data and not just excursions (aka. special cause outliers)

In summary, the outliers and wide IQR above exceedance level need further investigated.

Next is standard cum prob plot to test for normality. If the data is normal then the predicted line (red) will match the data (blue)

A graph with a line going up

Description automatically generated

Figure - Overall Cumprob Plot

The first look did not match well. A long trail of distribution > 2nd Quartile indication of outliers (aka. Point-source measurements) makes it hard to see how the core dataset matches

Filter data at 1000 cfu/100ml to see distribution shape

A graph with a red line

Description automatically generated

Figure - Cumprob Plot Zoomed

Normal distribution is not a great fit.

Try log-normal distribution.

A graph with a red line

Description automatically generated

Figure - Log Normal Prob Plot

Note that log-normal fit is better until about the 7.5 quantile level. Looks like there are multiple distributions blended into this dataset

Based on research on SW modelling defect data is expected to follow a log-normal distribution for bacterial counts in independent samples if variance is natural and only occasional spikes at higher concentrations. Below is an ideal model generated using the following code.

|  |
| --- |
| import pandas as pd  import numpy as np  # Set random seed for reproducibility  np.random.seed(42)  # Parameters for dataset  num\_samples = 500 # Total data points across all outfalls  outfall\_ids = ["Outfall\_1", "Outfall\_2", "Outfall\_3"] # Outfall locations  # Calculate samples per outfall to ensure uniform distribution  samples\_per\_outfall = num\_samples // len(outfall\_ids)  total\_samples = samples\_per\_outfall \* len(outfall\_ids) # Adjust to align  # Generate timestamps for each sample, spaced by the hour  times = pd.date\_range("2024-01-01", periods=samples\_per\_outfall, freq="H")  # Generate bacteria levels using a log-normal distribution  bacteria\_levels = np.random.lognormal(mean=3.5, sigma=1.0, size=total\_samples)  # Repeat outfall IDs and tile timestamps to align with the samples  outfalls = np.repeat(outfall\_ids, samples\_per\_outfall)  timestamps = np.tile(times, len(outfall\_ids))  # Create the DataFrame  data = pd.DataFrame({  "Timestamp": timestamps,  "Outfall": outfalls,  "Bacteria\_CFU\_100mL": bacteria\_levels  }) |

The output of this model is graphed below. We can know use this predicted model to match the actual data and see if they match. If there was not significant special cause exceedances then the sampled data should have a matching distribution to the model.

A graph of bacteria levels

Description automatically generated

Below is a histogram of the sampled data by site.

Check to see the data distribution per site models a log-normal distribution

A screenshot of a graph

Description automatically generated

Figure - Histogram Plot Per Site

The following visual observations were made.

### Three Year Distribution Observations

1. Bennett Ave – No distinct natural distribution. Special cause(s) likely or not enough data sampled. Needs further investigation (priority = high).
2. Broadway – Sufficient log-normal distribution
3. C St – Some log-normal shape with. Needs further investigation (priority = medium).
4. Cedar St - Sufficient log-normal distribution. No anomalies noted.
5. Cornwall St – Sufficient log-normal distribution. No anomalies noted.
6. Little Squalicum Creek – Sufficient log-normal. No anomalies noted.
7. Olive St – Minimal log-normal distribution. Needs further investigation (priority = medium).
8. Padden Creek – Needs further investigation (priority = medium)
9. Squalicum Creek – Sufficient log-normal distribution
10. Whatcom Creek – Sufficient log-normal distribution. No anomalies noted.
11. Willow Street – Not enough data to analyze

TODO: Use [StormReactor](https://github.com/kLabUM/StormReactor) to model pollution levels and ideal cases

How would I use python stormreactor https://github.com/kLabUM/StormReactor to model bacteria levels in a stormwater network?

## Time Series Analysis

The following charts show the time series analysis from Jan 2021-Sept 2024.

A graph of a graph of a graph

Description automatically generated with medium confidence

Figure - Median E.Coli by Month

Note that the summary data for all sites has two distinct periods (circled above) in which the median is well below the exceedance levels and steady (small variation).

Median from 2021 (Jan-Nov) and 2023 (Jan-August) is below the exceedance threshold. No data was collected for Oct 2023.

#### Time Series Trend by Site

Breaking the data down by site location yields the following graph.



Figure - E.Coli Readings by Month by Site

Zoom and clamp outliers provides better analysis resolution.

A graph of different colored lines

Description automatically generated

Figure - E.Coli Readings by Month by Site Zoomed

Note that some sites show the same low zones as the aggregate data but some do not. This points to evidence that specific outfalls/creeks have special cause variation and should be investigated further (Ex: Bennett Ave)

Because the readings vary so much and the dataset should follow a log-normal distribution the time series was also plotted with a log y-axis

A graph of different colored lines

Description automatically generated

Figure – Log (E.Coli Readings) by Month by Site

### 2024 Time Series Trend Observations

1. General: Readings with 0-1 CFU/100mL do not appear valid (TODO: What is discrimination of measurement?). Not enough resolution in measuring below 100 CFU/100mL (how to handle measurement deadband readings)
2. Bennett Ave: Baseline is consistently high. Concern
3. Broadway St: Recent exceedances (> May 2024) but trending down
4. C St: Recent exceedances (> May 2024) but trending down
5. Cedar St: Normal
6. Cornwall St: Two high readings in 2024. Recent trends normal
7. Little Squalicum: 2024 Normal, slight trend up
8. Olive St: Trending at threshold. Watch
9. Padden Creek: Trending above threshold in 2024.
10. Squalicum Creek: Trending above threshold since Aug 2024. Watch
11. Whatcom Creek: Normal, slight upward trend
12. Willow Creek: Normal. No data for 2024

### Yearly Exceedances Summary

Yearly E. Coli Exceedance Percentage Summary (> 310 cfu/100mL). This analyze matches the one used by Re Sources and outlined in their annual report. Exceedence Percentage threshold 20% (Verify)

A screenshot of a black screen

Description automatically generated

## Summary

Three analysis approaches were summarized

1. Three Year Distribution Modelling – Determining if measurement data showed a typical log-normal distribution indication of a majority of the measurements being due to natural variation.
2. Time Series Analysis (2024 focus) – Trend to determine if recent measurements were out of exceedance
3. Yearly Exceedance Percentage Summary – Overview of yearly exceedance levels

TODO: Table for each site with results from each of three analysis approaches above.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **#** | **Site** | **Dist Model** | **Time Series** | **2024 Exceedance %** |
| 1 | Bennett Ave | Fail | Fail | Fail |
| 2 | Broadway St | Pass | Pass | Fail |
| 3 | C St | Indeterminate | Pass | Fail |
| 4 | Cedar St | Pass | Pass | Pass |
| 5 | Cornwall St | Pass | Watch | Fail |
| 6 | Little Squalicum Creek | Pass | Watch | Pass |
| 7 | Olive St | Indeterminate | Watch | Fail |
| 8 | Padden Creek | Indeterminate | Fail | Fail |
| 9 | Squalicum Creek | Pass | Watch | Fail |
| 10 | Whatcom Creek | Pass | Watch | Pass |
| 11 | Willow St | Insuff Data | Insuff Data | Insuff Data |