Bellingham Stormwater Data Analysis

Date: 2024-10-22  
Author: Gil Lund

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

# Background

This document focuses on analyzing e. coli bacteria data collected by Re Sources from the creeks and 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 summary proposes 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)

A screenshot of a black screen

Description automatically generated

## Observations

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

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 exceedance threshold of 310 showing that there is a baseline exceedance issue with the data and not just excursions (aka. Special cause outliers)

In summary, the outliers and IQR above exceedance level should be further investigated.

A graph with a line going up

Description automatically generated

Figure - Overall Cumprob Plot

Long trail of distribution > 2nd Quartile indication of outliers (aka. Point-source measurements)

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.

A graph with a red line

Description automatically generated

Figure - Log Normal Prob Plot

Note that log-normal fits is much better until about 7.5 quantile

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.

A graph of bacteria levels

Description automatically generated

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

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?

### 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

## Time Series Analysis

The following charts show the time series analysis from Jan 2021-Sept 2024. Note that the median from 2021 (Jan-Nov) and 2023 (Jan-August) is below the exceedance threshold. No data was collected for Oct 2023.

A graph of a graph of a graph

Description automatically generated with medium confidence

Figure - Median E.Coli by Month

Time trend by Site



Figure - E.Coli Readings by Month by Site

Zoom and clamp outliers

A graph of different colored lines

Description automatically generated

Figure - E.Coli Readings by Month by Site Zoomed

Log Y-Axis

A graph of different colored lines

Description automatically generated

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

### 2024 Trend Observations

1. General: Readings with 0-1 cfu/100mL do not look valid (TODO: What is discrimination of measurement?)
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 E. Coli Exceedance Percentage Summary (> 310 cfu/100mL)

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** | **Exceedance %** |
| 1 | Bennett Ave | No |  |  |
| 2 | Broadway St | Sufficient |  |  |
| 3 | C St | Some |  |  |
| 4 | Cedar St | Sufficient |  |  |
| 5 | Cornwall St | Sufficient |  |  |
| 6 | Little Squalicum Creek | Sufficient |  |  |
| 7 | Olive St | Some |  |  |
| 8 | Padden Creek | Some |  |  |
| 9 | Squalicum Creek | Sufficient |  |  |
| 10 | Whatcom Creek | Sufficient |  |  |
| 11 | Willow St | Insuff Data |  |  |