E. coli Testing: Safeguarding Public Health*

Tianrui Fu

September 24, 2024

Swimming in beaches with high E. coli levels can lead to gastrointestinal issues, rashes, and infections, particularly affecting young children, the elderly, and those with weakened immune systems. In 1998, the implementation of the Ministry of Health Beach Management Protocol project led the Toronto department to start collecting beach water quality data in 2007, aiming to reduce the incidence of waterborne diseases in the population. Third sentence. Fourth sentence.

1 Introduction

The increase in E. coli concentration is a serious problem for humans. Elevated concentrations make beach water unsafe, hindering safe human use. Although some strains of E. coli are harmless, pathogenic microorganisms may still be present. In cases of increased concentration, healthy individuals among those infected can recover on their own; however, vulnerable populations such as children, the elderly, and those with compromised immune systems may suffer more severe illnesses, including general gastrointestinal discomfort, rashes, ear and eye infections, and in severe cases, bloody diarrhea, kidney failure, and death. The rise in E. coli concentration is caused by various factors, such as the runoff of animal feces carrying pathogens from land or sewage systems, aging sewage pipelines, and wastewater infrastructure issues. Considering the rapid growth of E. coli and the associated public health risks, real-time monitoring of beach water quality is essential.

According to the requirements of the Ministry of Health Beach Management Protocol (January 1, 1998), the Beach Water Sampling Program for the City of Toronto was implemented to reduce the incidence of waterborne diseases. Toronto's beaches have been certified under the Blue Flag program. From June to September each year, the relevant departments in Toronto collect water samples daily from all regulated beaches in the city to test for E. coli bacteria concentrations.

*Code and data are available at: LINK.

The number of E. coli in freshwater is determined by counting the number of yellow and yellow-brown colonies that grow after placing a 0.45-micron filter on m-TEC medium and incubating it at 35.0°C for 22-24 hours. The water quality standard for E. coli in Ontario and federally is 200 E. coli per 100 milliliters of water, while Toronto's beach water quality standard is 100 E. coli per 100 milliliters of water. This article utilizes 22,000 data points collected by government departments from June to September from 2007 to 2024 to assess...

2 Data

2.1 Raw Data

The data used in this paper is from Open Data Toronto and download by the opendatatoronto library. The dataset is used to analyze whether the E.coli value in the beach during June and September is satisfy for people to swim. All the data analysis is through the R Core Team (2023) (**tidyverse?**), (**tinytex?**), (**dplyr?**), (**janitor?**), (**ggplot2?**), (**here?**), (**kableExtra?**) and (**knitr?**). The dataset is published by Toronto Public Health and is part of The Beach Water Sampling Program, which collects E. coli values from two different beaches. The data is updated daily, and the data used for this analysis spans from June 3, 2007, to September 8, 2024. The raw dataset contains 21,882 water quality samples testing for E. coli concentrations. The dataset also includes beach itendifiers, names, sample site names, collection dates for each sample, and geographic coordinates (latitude and longitude).

2.2 Cleaned Data

In the raw data provided by Toronto Public Health, there were missing values (NA). During the data cleaning process, rows containing these NA values were completely removed to ensure that the NA values would not affect the analysis output and to simplify the analysis. Since the raw data also contained some particularly large outlier values, those were also removed. The cleaned data only includes the necessary columns for analysis, such as collection date, E. coli, site name, and beach name. Figure 1 shows the cleaned data samples, listing the results of tests conducted at different locations on the same day.

```
Attaching package: 'kableExtra'

The following object is masked from 'package:dplyr':

    group_rows
```

Beach_Name	Date	E.coli	Site_Name
Marie Curtis Park East Beach	2024-09-01	230	29W
Marie Curtis Park East Beach	2024-09-01	220	33W
Marie Curtis Park East Beach	2024-09-01	240	32W
Marie Curtis Park East Beach	2024-09-01	200	31W
Marie Curtis Park East Beach	2024-09-01	200	30W
Sunnyside Beach	2024-09-01	910	18W
Sunnyside Beach	2024-09-01	210	17W
Sunnyside Beach	2024-09-01	170	20W
Sunnyside Beach	2024-09-01	80	21W
Sunnyside Beach	2024-09-01	400	19W
Sunnyside Beach	2024-09-01	360	22W

Figure 1: Bills of penguins

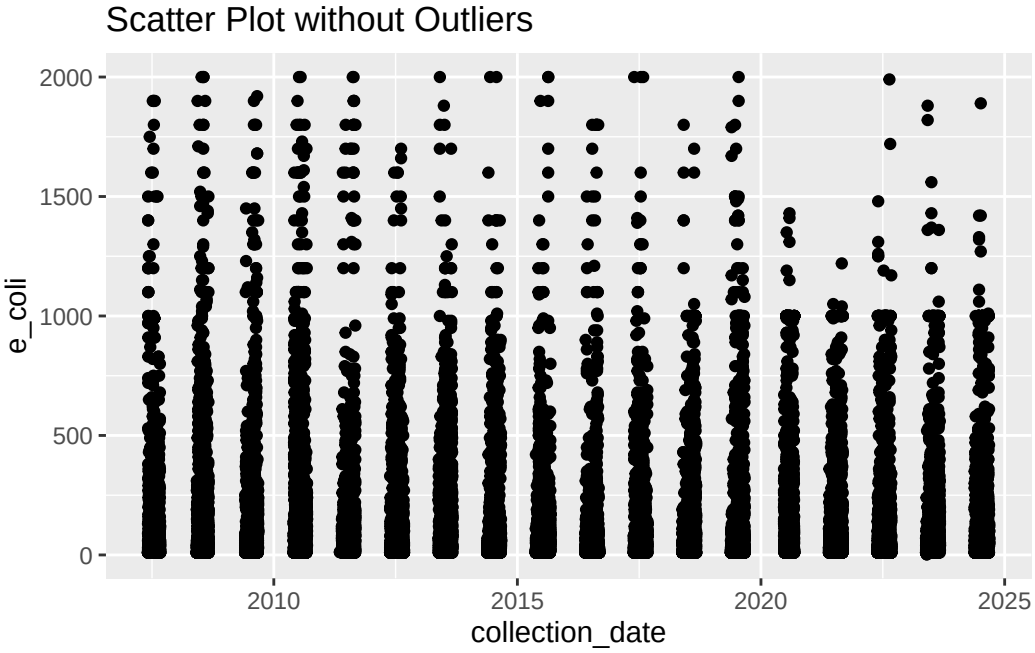

Figure 2: Relationship between wing length and width

2.3 Basic Summary of cleaned Dataset

Figure 1 shows the attributes of the data, with most E. coli values being below 500, meaning there are 500 E. coli per 100 ml of water. However, the scatter plot also reveals that the data from 2020 stands out compared to other years, but a more precise graphical output is needed for a better understanding of the dataset. Therefore, in Table 2, a bar chart was used to show the total amount of valid data collected each year. It is clear that due to the COVID-19 pandemic, the data collected by Toronto Public Health in 2020 is significantly less compared to other years. Figure 2 also provides a calculation of the distribution of collected E. coli data. Based on Canada's water quality standard of 200 E. coli per 100 ml, about 78% of the data falls within this standard, meaning that approximately 78% of the days from 2007 to the present were safe for beach activities. However, based on Toronto's stricter standard of 100 E. coli per 100 ml, only around 60% of the data meets this standard, meaning that about 60% of the days were safe for beachgoers. However, considering the additional variables of location and date, it is not yet possible to draw a reasonable conclusion about the overall water quality. A more in-depth analysis will be conducted in Section 3.

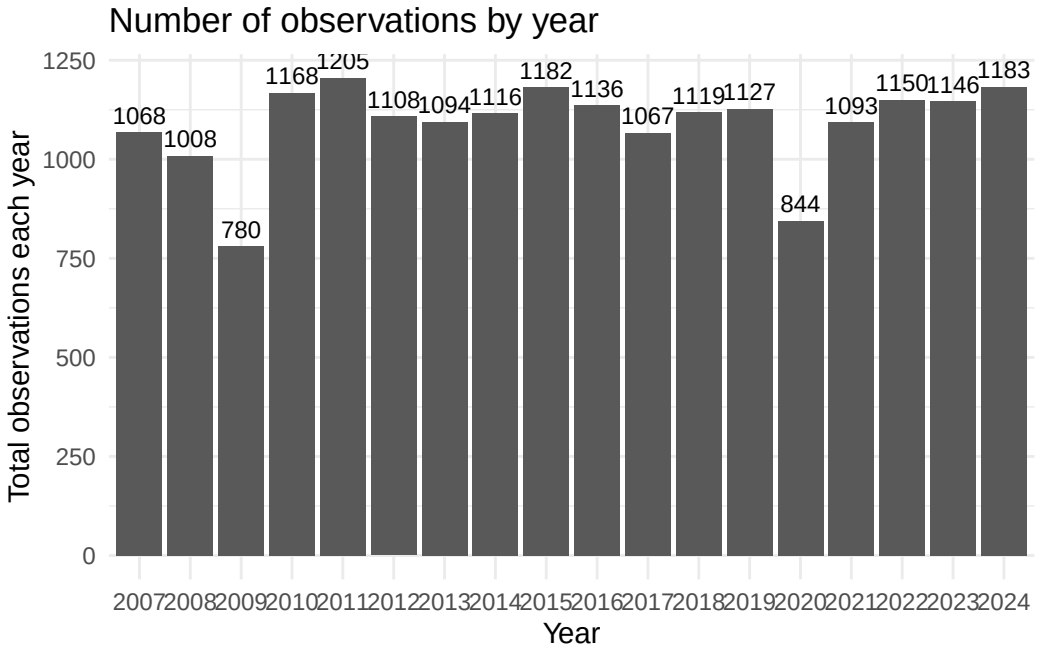

Figure 3: Bills of penguins

2.4 Dataset justification

The reason for choosing this dataset is that, during this summer, there were news reports about several incidents of people defecating on Toronto beaches, with even photos of the feces

Table 2: Portion of E.Coli levels in different ranges

E.Coli Levels	Sample Count	Portion
0-50	8588	43.82974
50-100	3609	18.41890
100-200	3009	15.35674
>200	4388	22.39461

Figure 4: Bills of penguins

circulating. This caused some panic among people in Toronto who were planning to visit the beach. To verify whether the beach closures were indeed caused by these incidents and to reduce potential bias, as well as out of personal interest, this paper uses data published by Toronto Public Health. The dataset spans 18 years, from 2007 to the present, containing E. coli test results for beach water quality. It effectively reveals the trends in water quality over the years and helps to project future trends, reducing doubts about the validity of the analysis due to a small data sample.

3 Result

3.1 Studying the relationship between time and E. coli concentration

3.2 Examing the portion of E.coli higher than the standard level by year

3.3 Investigating the relationship with E.coli and site location

4 Results

Our results are summarized in Table **??**.

5 Discussion

5.1 First discussion point

If my paper were 10 pages, then should be be at least 2.5 pages. The discussion is a chance to show off what you know and what you learnt from all this.

5.2 Second discussion point

5.3 Third discussion point

5.4 Weaknesses and next steps

Weaknesses and next steps should also be included.

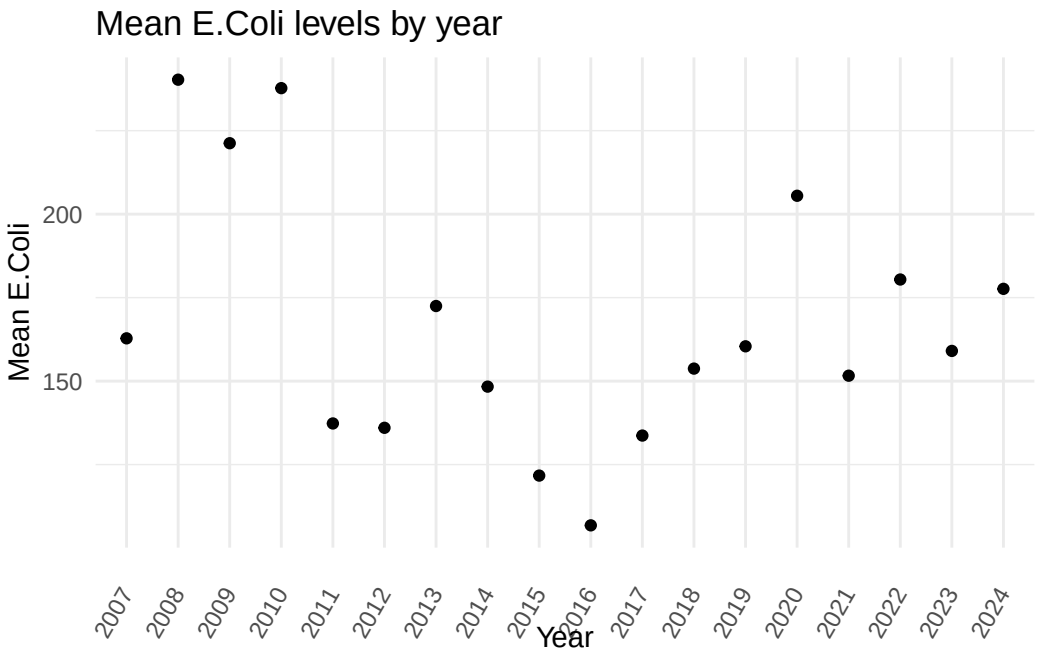

Figure 5: Bills of penguins

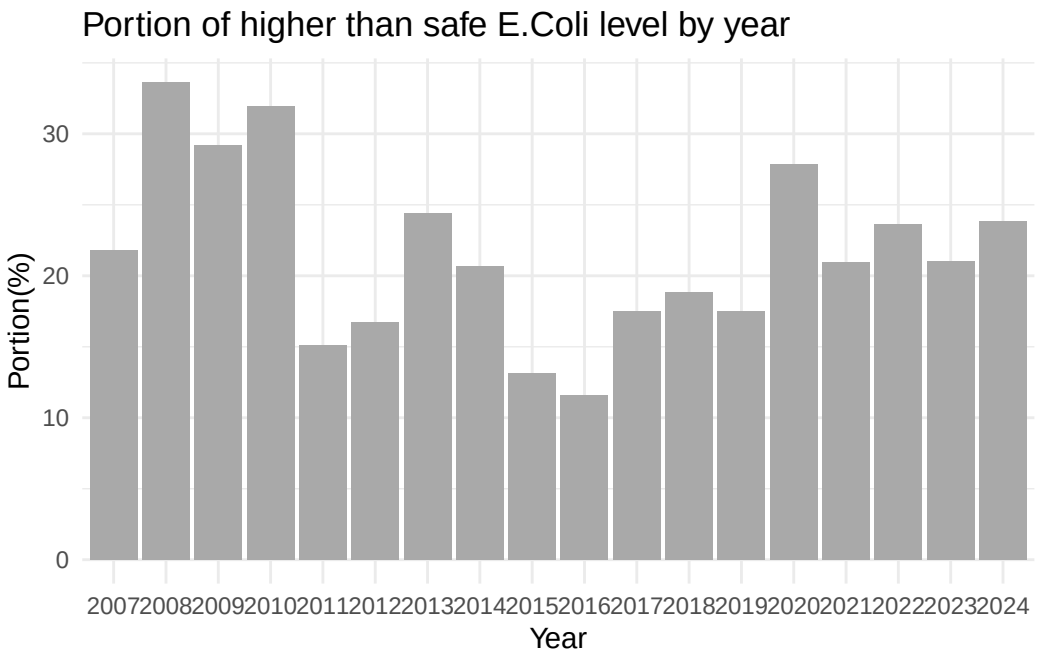

Figure 6: Bills of penguins

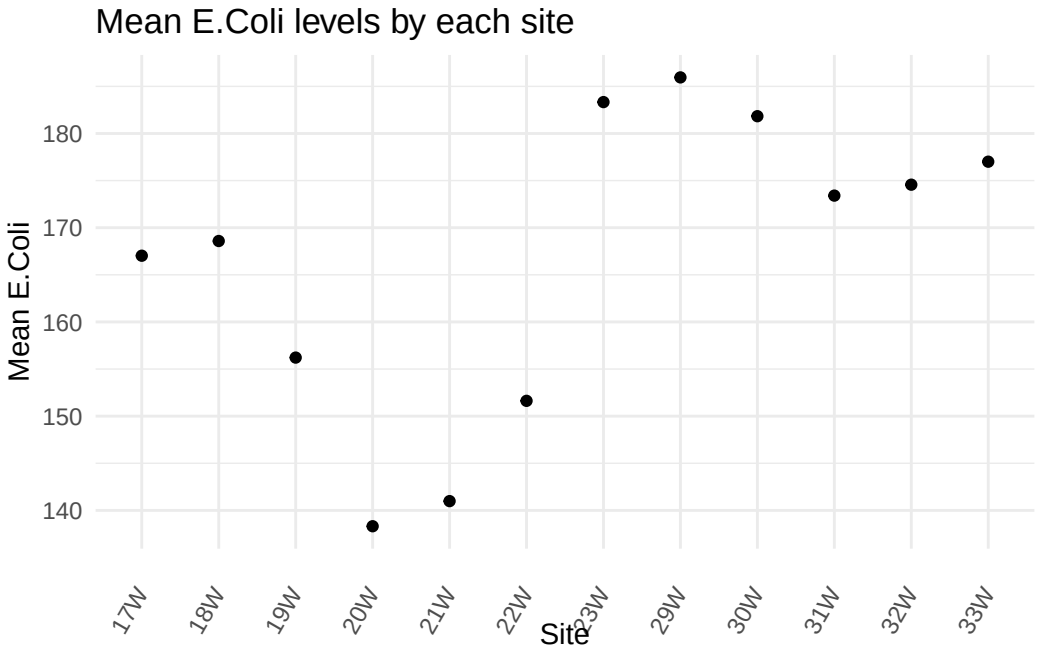

Figure 7: Bills of penguins

Table 3: Portion of E.Coli levels above safe level in different site

Site name	**Total Samples**	**Above 200 count**	**Portion(%)**
17W	1778	392	22.04724
18W	1774	395	22.26607
19W	1778	376	21.14736
20W	1782	308	17.28395
21W	1783	319	17.89119
22W	1770	352	19.88701
23W	460	127	27.60870
29W	1691	403	23.83205
30W	1693	392	23.15416
31W	1697	376	22.15675
32W	1696	367	21.63915
33W	1692	372	21.98582

Figure 8: Bills of penguins

Appendix

A Additional data details

B Model details

B.1 Posterior predictive check

In **?@fig-ppcheckandposteriorvsprior-1** we implement a posterior predictive check. This shows…

In **?@fig-ppcheckandposteriorvsprior-2** we compare the posterior with the prior. This shows…

Examining how the model fits, and is affected
by, the data

B.2 Diagnostics

?@fig-stanareyouokay-1 is a trace plot. It shows… This suggests…

?@fig-stanareyouokay-2 is a Rhat plot. It shows… This suggests…

Checking the convergence of the MCMC algo-
rithm

References

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.