Supplimentary Materials Accompanying Statistical Machine Learning Methods for Food Source Attribution

Prepared by Amos Okutse

2022-12-08

variable pre-processings

- Exploratory data analysis

Exploratory data analysis

Overall trends in the counts of *L. monocytogenes* through time

We used descriptive statistics to first examine the proportion represented by our main variable of interest, the isolation source which originally had 1401 unique values which were as a result of punctuation, case sensitivity as well as many variations of the naming conventions of a general source. For example, 'cheese', 'white cheese', 'ham cheese', and 'double cheeseburger'. For simplicity and for comparative purposes, we grouped the isolation sources into broader categories based on the patterns observed in this variable. Ultimately, the number of isolation sources was reduced to 38 broad categories including environmental, food, pork, chicken, beef,turkey, stool, water, other/unspecified. Environmental sources were highest at 54.34% followed by other/unspecified sources (9.65%). Water, dairy, and food sources represented 9.65%, 9.24% and 6.43%, respectively, while fish, beef, and pork represented 1.67%, 1.56%, and 1.47%, respectively.

We then used line plots to show an initial exploration of the trends in the number of *Listeria monocytogenes* over time. We filtered our data to work with a time frame from the year 2000 to 2022. The line plots in Figure @ref(fig:fig-three) show a non-linear trend over time. There was a moderate increase in samples collected from the year 2000 to 2008, which sharply increased until about the year 2018. From 2018 to 2020, there was variation in terms of steady decrease/increase that was later followed by another sharp decrease in the samples collected. However, we also observed a slight increase in the counts following the year 2020. Grouped by isolation types, we observed a higher count in the environmental/other source types compared to the clinical type which remained relatively lower throughout the entire period of sample collection.

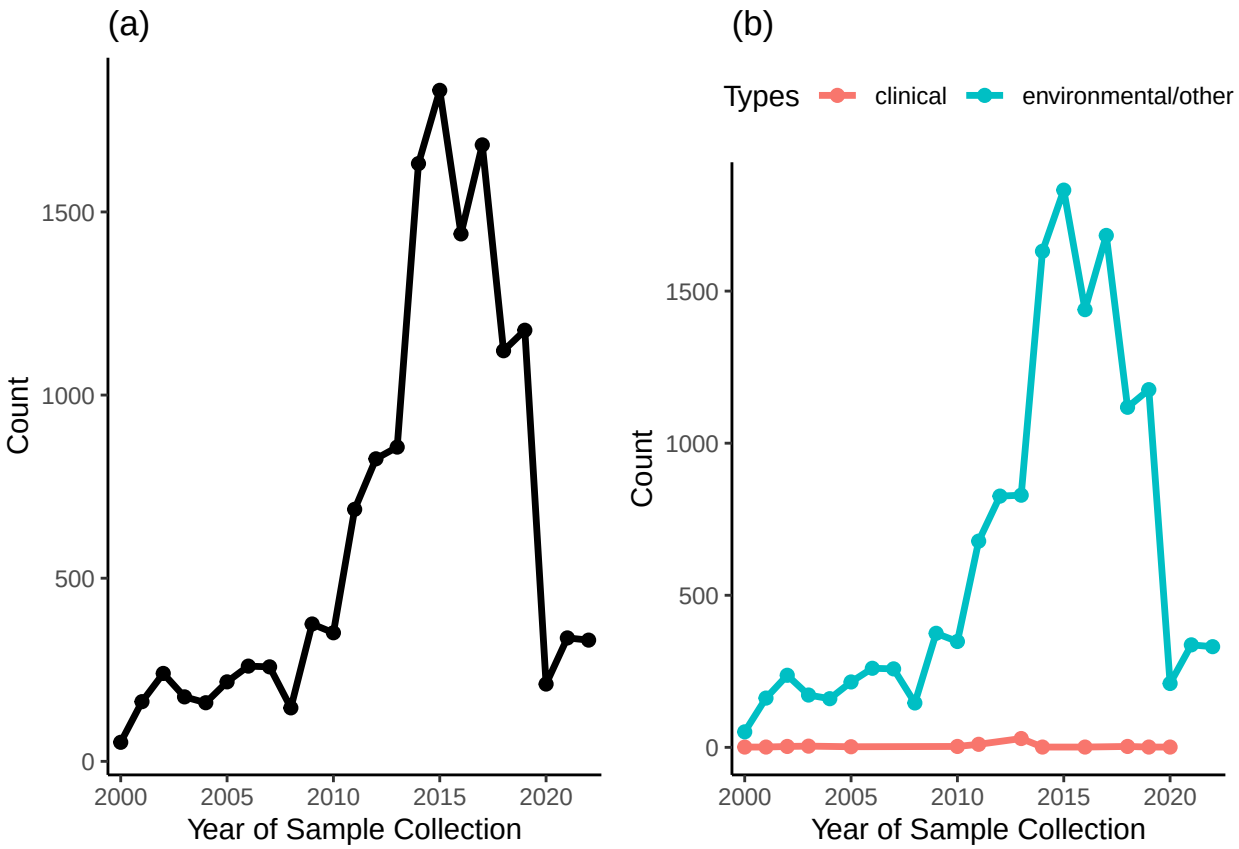

Figure 1: **(a)** Trends in the total counts of collected *L. monocytogene* pathogens. **(b)** Trends in the total counts of *L. monocytogenes* by isolation type.

Additionally, we explored summary frequencies of *Listeria monocytogenes* grouping by `Month` and `Isolation type`. Table @ref(tab:table-two) summarizes the counts of the *L. monocytogenes* where we observe that most cases of *Listeria monocytogenes* were observed in the early month of January; with 40% for clinical isolation type and 22.92% for environmental/other types. Additionally, for the clinical isolation types, frequent cases were observed in the months of September and October at 26.67% and 11.67% respectively. For the environmental/other isolation type, during the warmer months of April to August, we observed a moderate number of cases of *Listeria monocytogenes* ranging between 6.99% and 9.08%. There was missing clinical isolation types cases observed during the months of March, April and July. Looking at the trends by State, California had the largest number of *Listeria monocytogenes* cases throughout our study time frame, $N = 2672$ (18.38%), followed by New York and Washington DC at 12.78% and 8.67% respectively. Nevada, West Virginia and Puerto Rico had the least number of cases each at 0.02%.

Trends in the counts of *L. monocytogenes* over time by the top isolation sources

The `Isolation source` was re-categorized into 38 broad categories. Figure @ref(fig:fig-four) presents the trends over time in the counts of *Listeria monocytogenes* for the following sources: beef, chicken, dairy, pork, fish, food, potato, water. We observe that the most common isolate source in our data is dairy, water, followed by food and and pork.

Table 1: Counts of *Listeria monocytogenes* by month

Isolation	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
clinical	40%	3.33%	NA	NA	6.67%	3.33%	NA	5%	26.67%	11.67%	1.67%	1.67%
Environmental	22.92%	6.94%	7.16%	9.08%	7.28%	7.38%	7.52%	6.99%	6.71%	7.02%	5.31%	5.69%

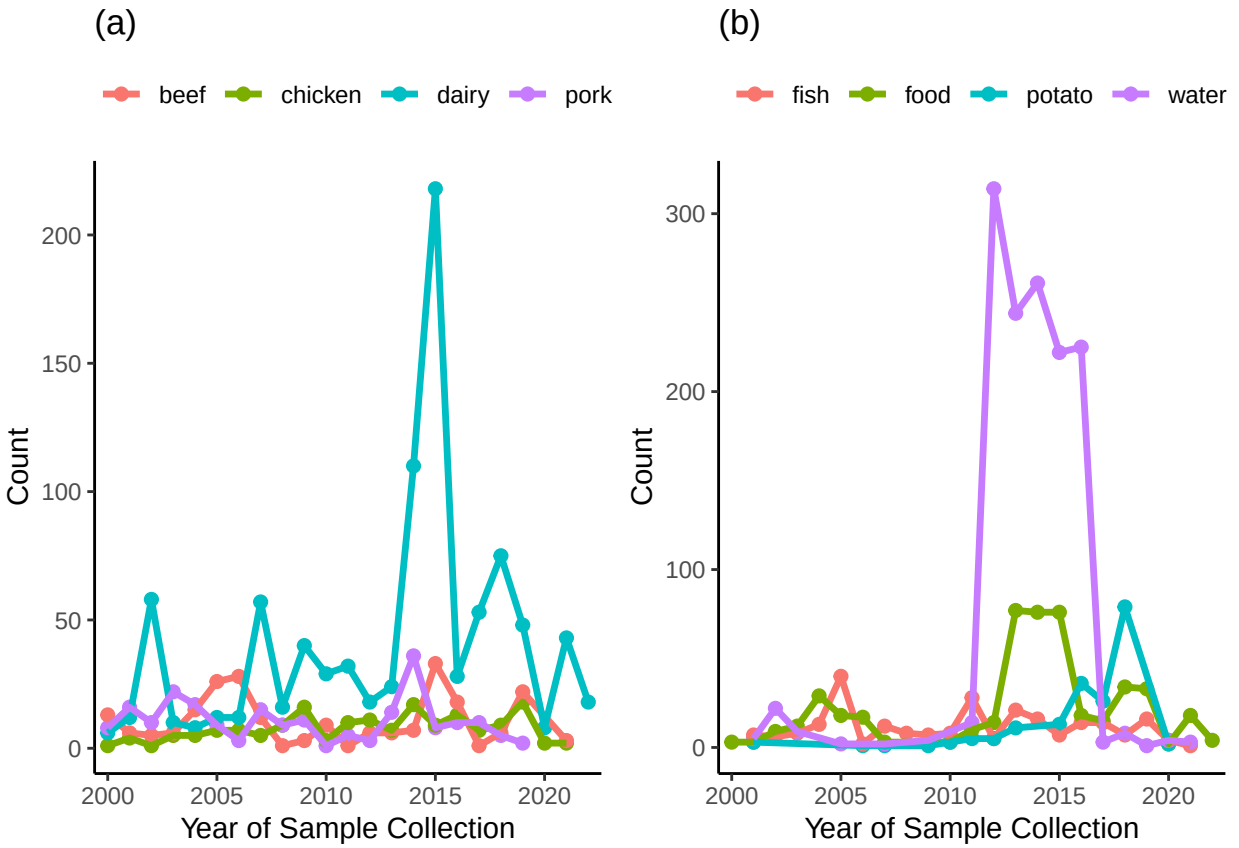

Figure 2: Line plots of top isolation surces for listeria monocytogenes counts over time. **(a)** Trends in the counts of pathogens from beef, chicken, dairy, and pork. **(b)** Trends in the pathogen counts from fish, food, potatoes, and water.

Serovar, AST phenotypes, AMR genotypes, and SNP Clusters

Our data had 14,534 unique isolates for *Listeria monocytogenes* with 14517 distinct `Biosamples` and no missing data on this information. Additionally, we looked at the distribution of `Serovar` and noted a relatively high percentage of missing data (n = 14275; 98.22%). `Serovar` information was entered using free text as there are many variations of names that could be representing similar information such as 1, 1a, 1/2a. On the other hand, the `AST phenotypes` variable, denoting the Antimicrobial Susceptibility Test was recorded in a raw string form. This variable represents the antibiotics that each isolate is either susceptible or resistant to. The `AMR genotypes` variable represents the Antimicrobial resistance (AMR) genes found in the isolate during analysis. We found 184 unique AMR genes in our data with no missing information. There were 1, 474 SNP clusters whose genome assemblies were closely related.

Distributions of `Min Same` and `Min Difference` variables.

Next we examined the distributions of `Min Same` and `Min Diff` variables. `Min-diff` is the minimum SNP distance to another isolate of a different isolation type (from an environmental isolate to a clinical isolate). Figure @ref(fig:fig-five) shows that `Min Diff` approximately follows a bi-modal distribution suggesting that a transformation of this variable may be useful prior to using it in further analysis. On the other hand, `Min-same` was the minimum SNP distance to another isolate of the same isolation type (clinical to clinical or environmental to environmental). Additionally, Figure @ref(fig:fig-five) shows that `Min Same` approximately follows an exponential distribution. A log transformation of this variable did not suggest a substantial deviation from the exponential distribution.

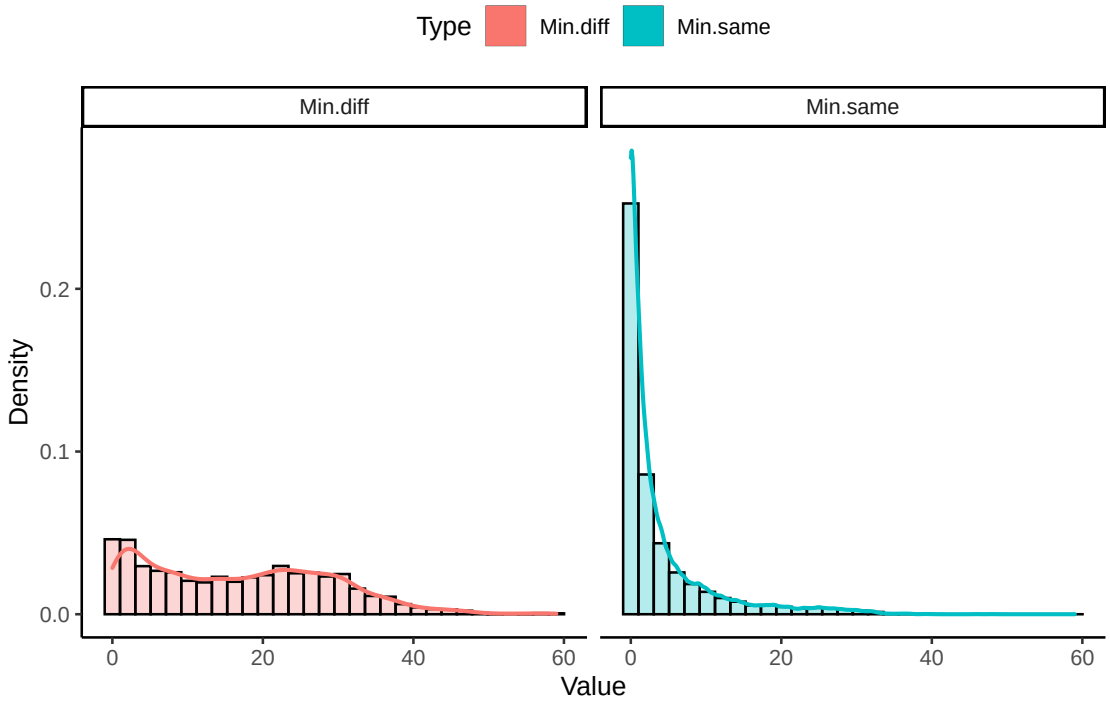

Figure 3: The distributions of the minimum SNP distance to another isolate of a different isolation type and the minimum SNP distance to another isolate of a similar isolation type.