Supplimentary Materials Accompanying Statistical Machine Learning Methods for Source Attribution

Prepared by Amos Okutse

2022-12-08

Variable pre-processing steps.

- There were 1401 unique isolation sources entered into the NCBI database from the date the data was generated, which comprised of 60 clinical types and 14,474 environmental/other types. To make meaningful comparisons in relation to our objective, for the initial work of this project, we aggregated the Isolation sourced category into 38 categories based on the top sources with relative frequencies ≥ 0.1. These categories were used in the exploratory analysis as shown in Figure 1 and 2.
- For the Methods section, we further aggregated the Isolation sourcing categories to be representative of the 7 broad sources contained in the IFSAC category scheme; Dairy, Poultry, Meat, Leafy greens, Fruits, Vegetables, Sea Food. Additional categories in the newly created source included environmental, human and other sources, bringing to a total of 10 categories.
- We also examined the collection date variable that was used to explore trends over time. Since the Collection date variable contained the date the sample were collected in the format the submitter supplied ranging from Month-Date-Year, Year-Month and Year only while the Create date was in the Year-Month-Date ISO format with time stamp the data was added into the Pathogen Detection Project, we first converted these into a standard form of Year-Month-Date. Then for Collection date variable with missing values, we chose to fill in these dates by using those from the Create date variable. Finally we created Year and Month variables and extracted the respective years and months from the Collection date variable to maintain consistency in terms of available year records.
- For the seasonality variable included in our model building process, we created 4 seasons (Winter, Spring, Summer, and Fall) based on the newly created Month variable.
- For the Location variable, we reduced this to only include 49 states in the United States and District of Puerto Rico, District of Columbia. We corrected for four wrongly abbreviated states (`FL`, `NC`, `CO`,`AZ`) and one state named `USA` was recoded to `Other`.
- Finally for the Single Nucleotide Polymorphism (SNP) variable, we reduced this variable to 19 categories based on SNP clusters with $N > 100$ observed frequencies.

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 1 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.

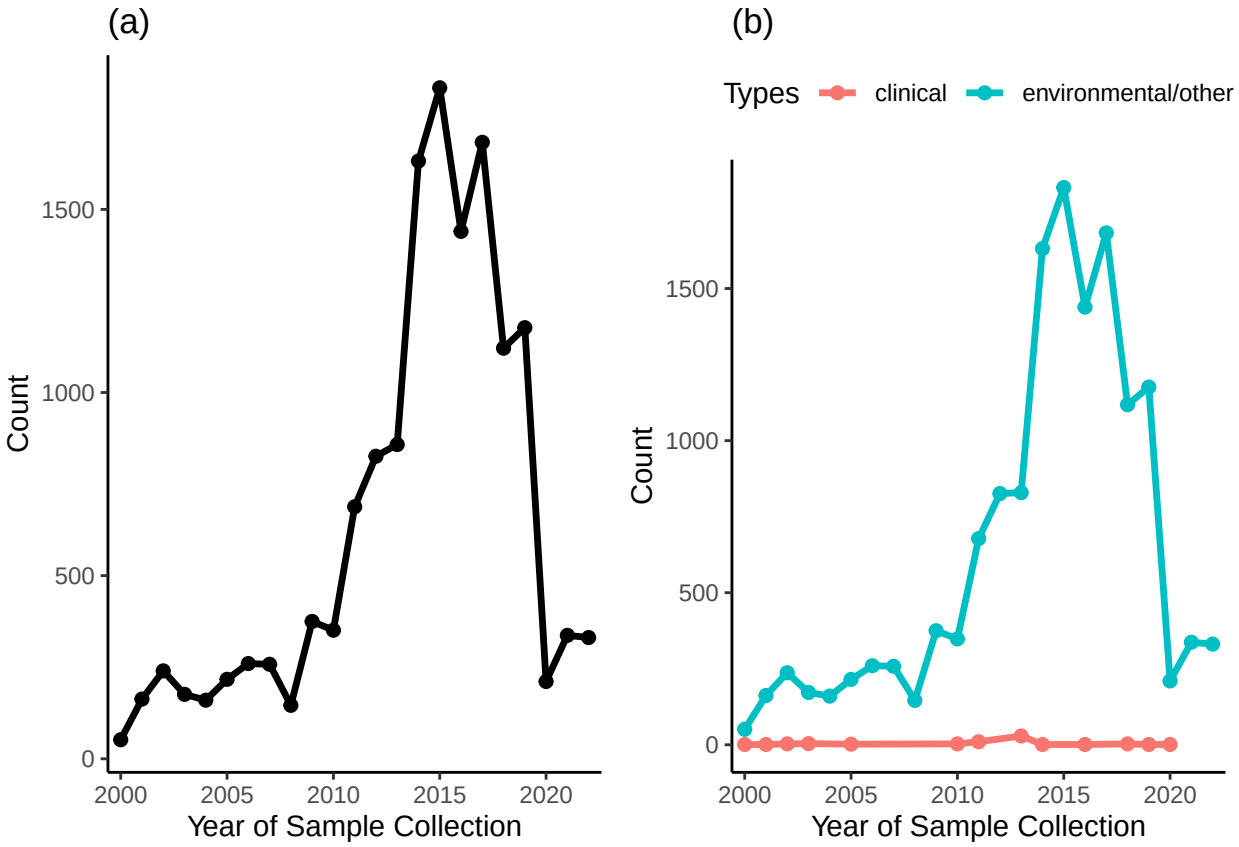

Figure 1: **(a)** Trends in the total counts of collected *L. monocytogene* pathogens. **(b)** Trends in the total counts of *L. monocytogenes* by isolation type.

Additionally, we explored summary frequencies of *Listeria monocytogenes* grouping by `Month` and `Isolation type`. Table 1 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%.

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%

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 2 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.

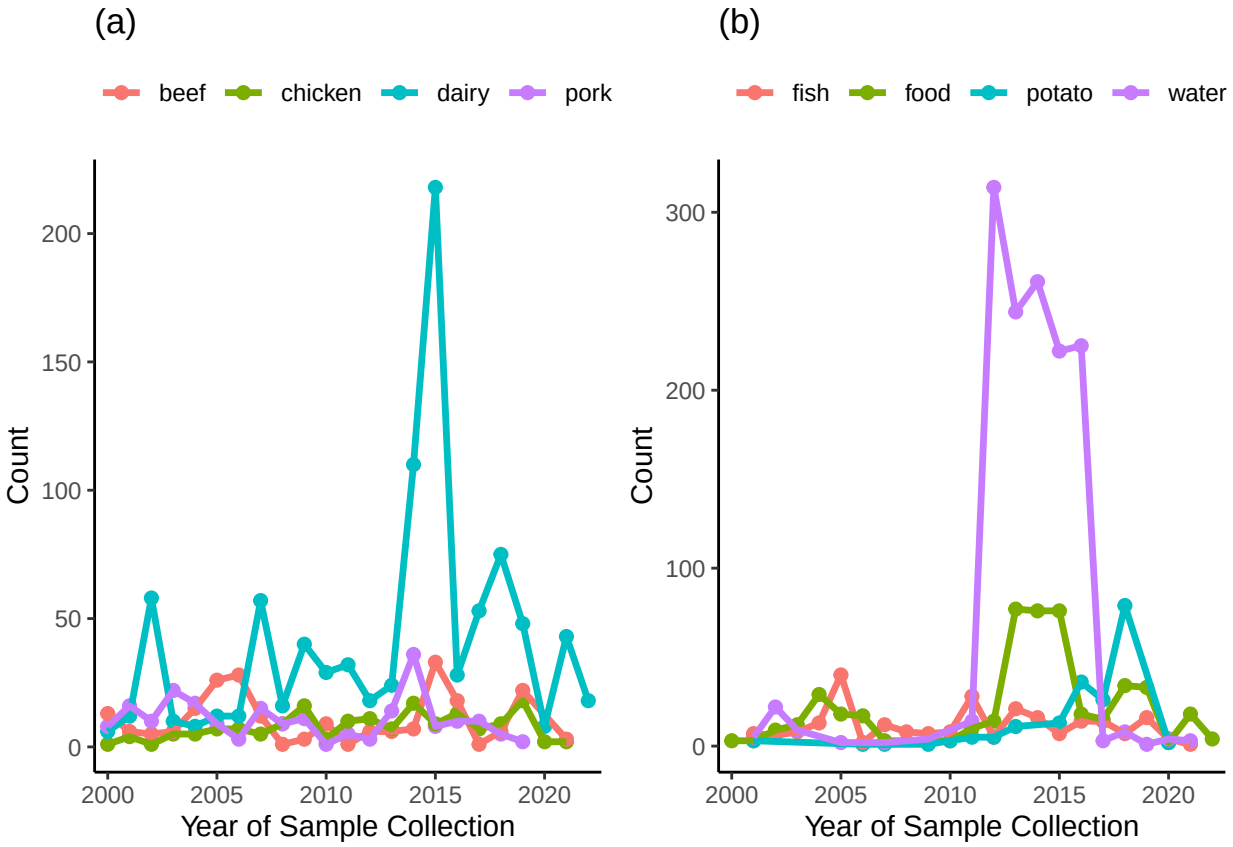

Figure 2: Line plots of top isolation surces for listeria monocytogenes counts over time. **(a)** Trends in the counts of pathogens from beef, chicken, dairy, and pork. **(b)** Trends in the pathogen counts from fish, food, potatoes, and water.

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 3 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 3 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.

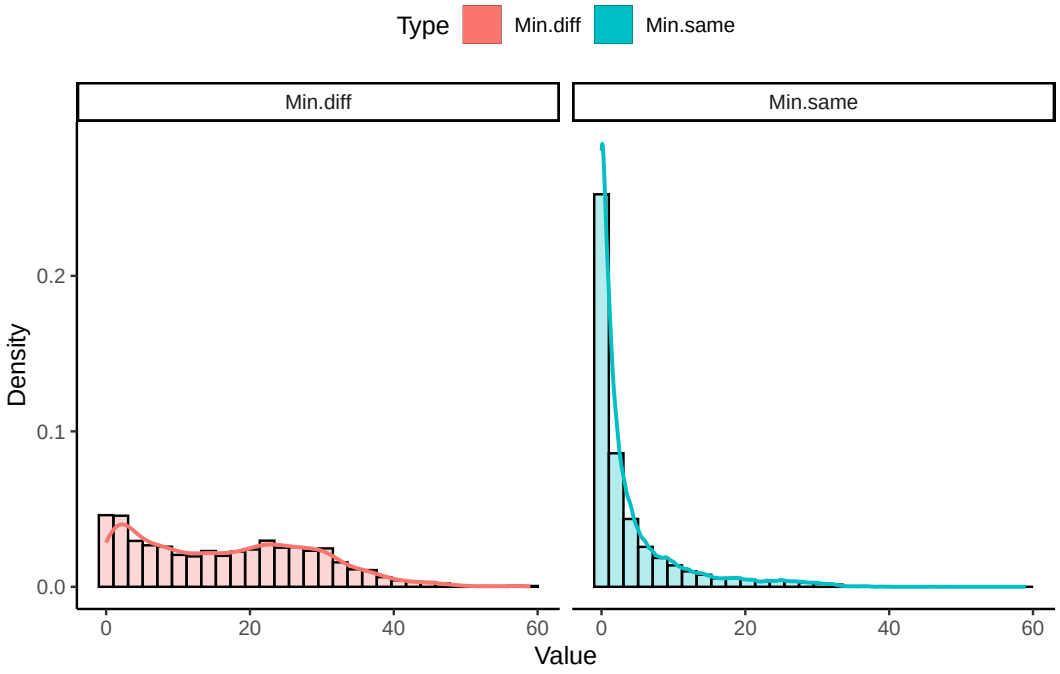

Figure 3: The distributions of the minimum SNP distance to another isolate of a different isolation type and the minimum SNP distance to another isolate of a similar isolation type.

Modeling

This section presents additional material and output from the statistical modeling including results based on additional data validation processes. Table 2 presents the proportions of each sample in the training and testing proportion of the data by the isolation source.

Table 2: Proportions of sample from each isolation source in the derivation and validation datasets

Food source	Train set sample size	Train set proportion	Test set sample size	Test set proportion
dairy	472	0.0754	152	0.0728
environment	4271	0.6823	1425	0.6828
fruits	189	0.0302	69	0.0331
human	49	0.0078	17	0.0081
leafy_greens	58	0.0093	12	0.0057
meat	191	0.0305	76	0.0364
other	468	0.0748	150	0.0719
poultry	189	0.0302	65	0.0311
sea_food	149	0.0238	49	0.0235
vegetables	224	0.0358	72	0.0345

Naive Bayes Gain and ROC Curves

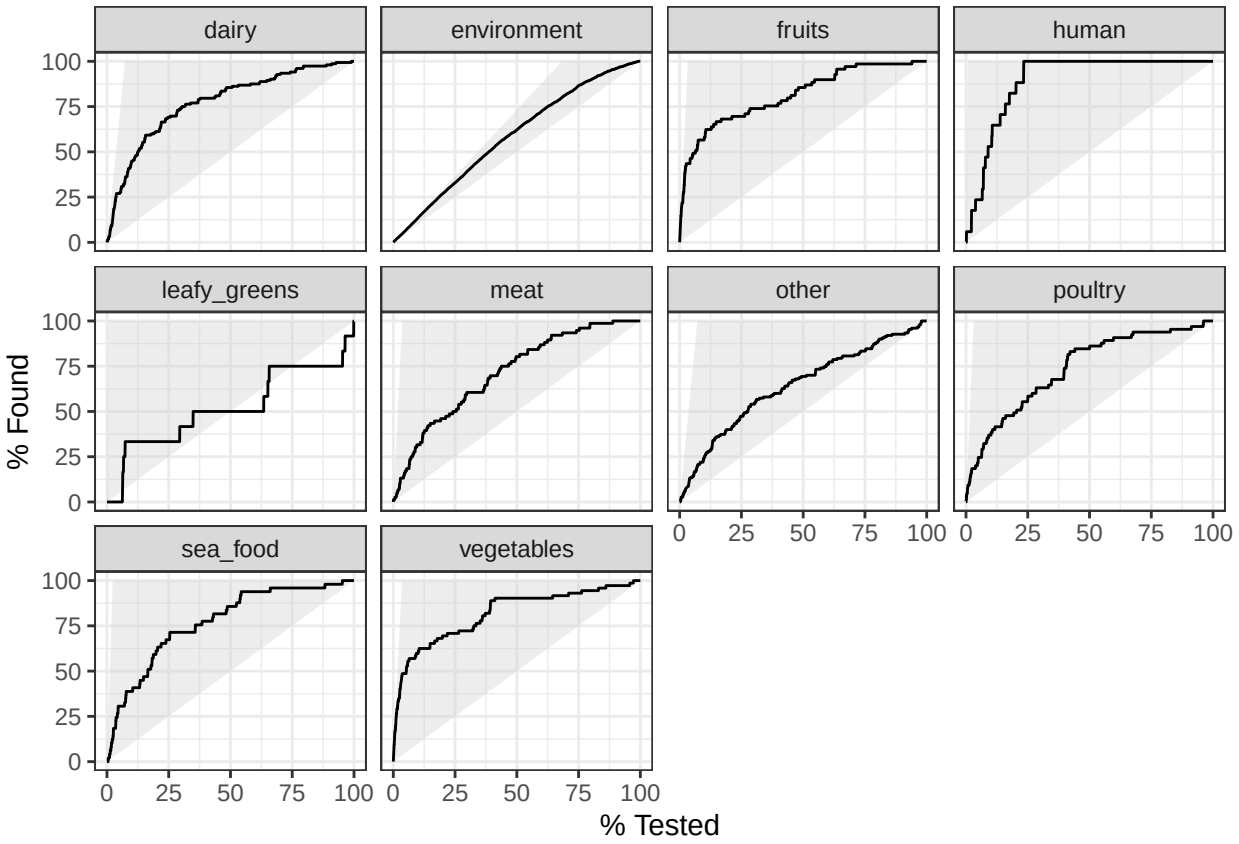

Figure 4: The Naive Bayes Gain Curve on the test data set

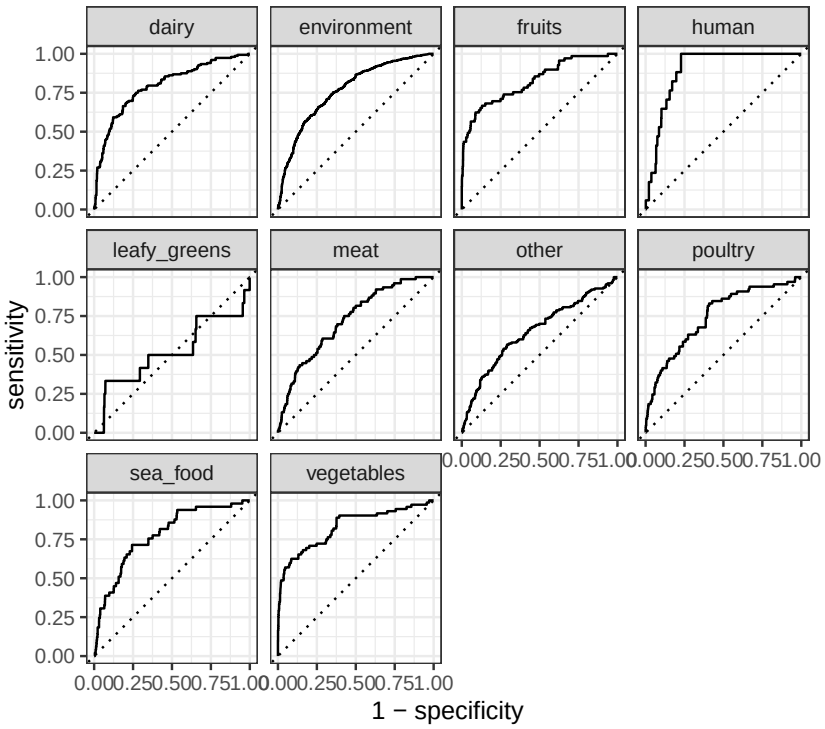

Figure 5: The naive Bayes AUC curve on the test dataset

Random forest gain and AUC curve

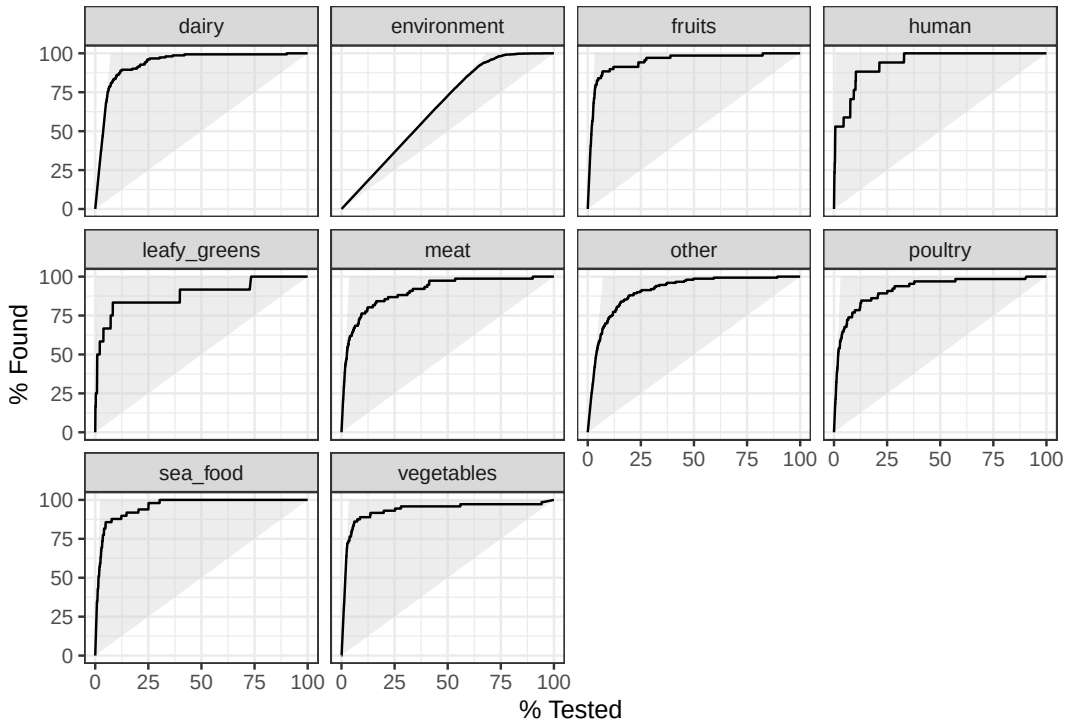

Figure 6: Random Forest gain Curve on the test data

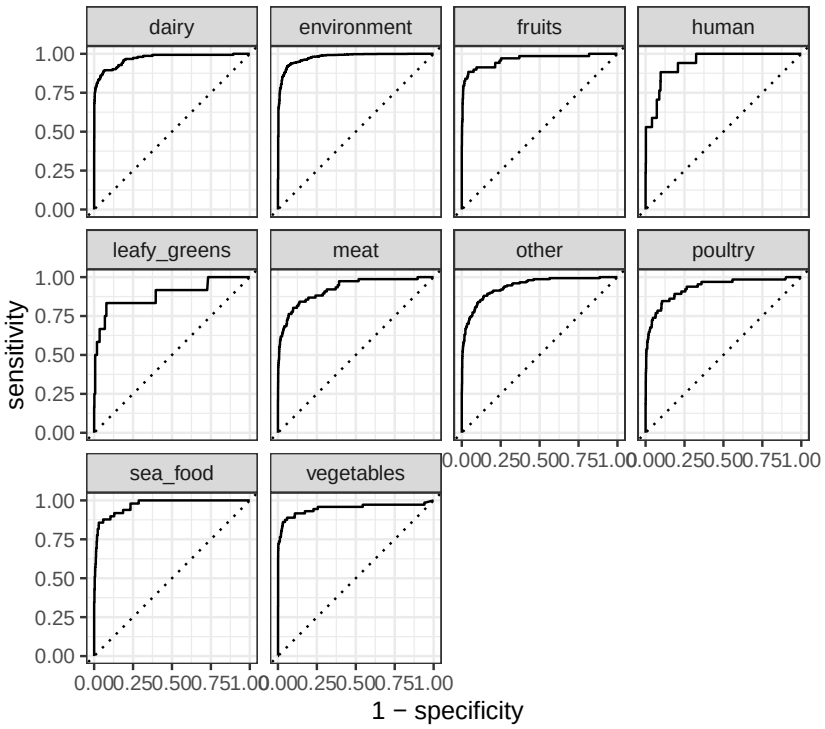

Figure 7: The AUC curve for the random forest regression model on the test dataset

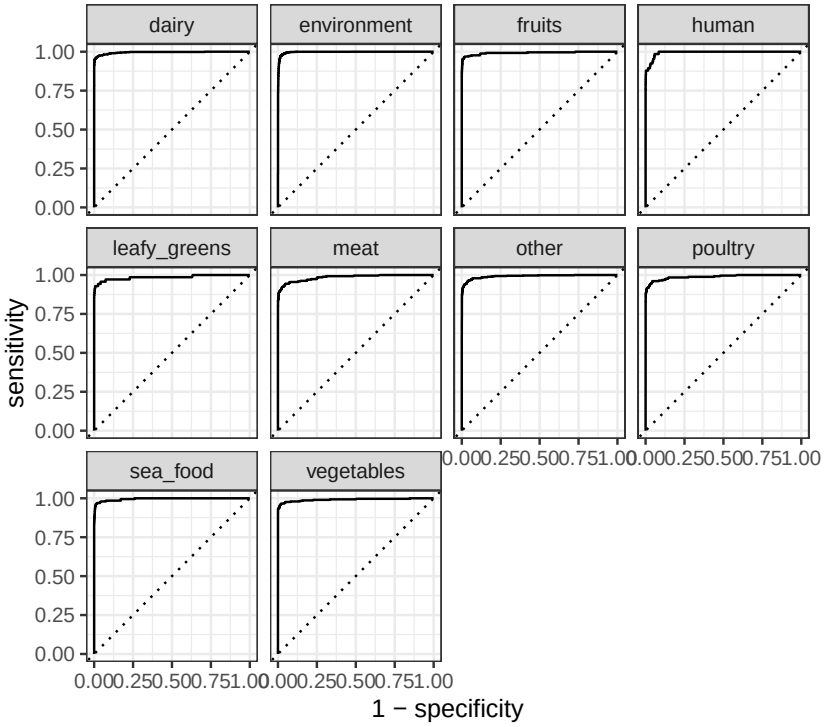

Figure 8: The AUC curve for the random forest model on the full data set

Testing the predictive ability of the random forest model

Table 3: Sample predictions from the random forest classification model

.pred_dairy	.pred_environment	.pred_fruits	.pred_human	.pred_leafy_greens	.pred_meat	.pred_other	.pre
0.0828587	0.8343876	0.0221598	0.0011165	0.0014453	0.0043944	0.0350184	
0.9103298	0.0121767	0.0422206	0.0026739	0.0016550	0.0064784	0.0125528	
0.0301276	0.9380841	0.0030585	0.0015042	0.0033271	0.0016892	0.0060091	
0.1028258	0.4723504	0.0103335	0.0040000	0.0880000	0.0383925	0.0795939	
0.0161343	0.8530499	0.0000000	0.0000000	0.0080000	0.0044531	0.0365290	
0.0028908	0.9604319	0.0006509	0.0000000	0.0004215	0.0133916	0.0039756	
0.0238354	0.9136383	0.0000000	0.0000000	0.0020000	0.0046718	0.0558545	
0.1525350	0.8151867	0.0057020	0.0009628	0.0031448	0.0103222	0.0054626	
0.0100000	0.8394850	0.0075500	0.0000000	0.0060000	0.0025274	0.0369823	
0.0267614	0.5360709	0.0132456	0.0020000	0.0070772	0.0094105	0.2504154	

Confusion matrix testing the performance of the random forest model on the full dataset

Here, we examine the performance of the random forest model. How many isolation sources were attributed to each isolation source by the random forest model? These have been presented as % in the main text.