High-Risk Serovar Prevalence in Sources of *Listeria monocytogenes*: A Clustering Analysis of Food-Borne Illness

Nathan Provost and Antonella Basso

Abstract

Relevance: *Listeria monocytogenes* (hereafter listeria) is a dangerous species of bacteria that poses a direct threat to public health, and the association between severe strains of listeria and their sources of origin has not been studied with sufficient precision and attention to detail.

Goals: We endeavor to create two clustering models for data pertaining to the sources, collection times, minimum self-same distances, and genetic information of various strains of listeria, so that we can assess the intensity of particularly problematic strains within clusters. We also aim to compare our new models to a pre-existing genetic cluster model included in the dataset in terms of similarity.

Methods and Setting: We employ traditional k-means clustering with optimal cluster selection using silhouette coefficients and network clustering with similarity weighting through the `linkcomm` package in R (Kalinka and Tomancak 2011). Data is assessed for the time period between 2017 and late 2022, with each year divided into four quarters. Sources of isolation are grouped coarsely and missing data is not used in our analysis, which allows for 2290 unique strains of listeria to be analyzed.

Outcomes of Interest: Serovar 4b prevalence by cluster, source prevalence by cluster, Rand index values, silhouette coefficient plots, community centrality, and community modularity are all outcomes of interest. Visualizations of cluster interactions and networks are also objects of interest.

Results: A strong sense of similarity exists between our sample networks and the SNP model, while a moderate sense of similarity exists between our k-means model and both our network model and the genetic model. Both genetic and non-genetic factors appear to influence serovar 4b prevalence, but specific recommendations cannot be made. Our methods provide framework for future research and reproductions.

Conclusion: It is likely that severe strains of listeria manifest more frequently in specific isolation sources, but our analysis is to limited to provide any further guidance on this matter. Future studies should aim to atomize isolation sources into more precise groups in order to better understand the relationship between strain severity and source. However, our methodology has laid the groundwork for more intensive research.

Introduction and Review of Literature

Listeria monocytogenes is a particularly problematic species of bacteria that causes numerous cases of food-borne illness each year both in the United States and across the world. The most vulnerable populations to its effects consist of infants, pregnant women, the elderly, and the immunocompromised, who frequently experience the common symptoms of infection that include flu-like symptoms (nausea, vomiting, fever, etc.) as well more severe symptomatic manifestations. (Rogalla and Bomar 2022) While it is often characterized medically as a food-borne illness, being found frequently in cheeses, cold meats, and unprocessed/improperly processed diary products, it exists environmentally in the soil and decaying organic matter, but does not surface as a point of infection nearly as frequently as it does in food-processing areas such as farms or improperly cleaned production facilities. (Rogalla and Bomar 2022) (Ward et al. 2004) Furthermore, the severity of strains extracted from differing environments, or more specifically, that have evolved to suit different environments, thus suggesting an increased presence, has been a subject of comparison and evaluation in recent literature (Ward et al. 2004), which is where the basis of our line of questioning lies.

It is essential to elaborate on the results of one recent article (Ward et al. 2004) that provides a rigorous overview of the differences in strains of listeria by serovars, small biological distinctions in the bacteria that separate one subgroup of the species from another. Developing robust methods of separating one strain of listeria from another is essential to providing comprehensive analysis of its pathology, since differences in serovars are often related to differing levels of disease manifestations in human beings as shown in this article. (Ward et al. 2004) Its results present a a phylogenetic tree with over 60 strains of listeria. These strains are grouped into three lineages (LI, LII, and LIII) which separate them by small genetic differences. These lineages are further organized by several different serovars, which distinguish the strains through ever more minute genetic and physiological differences. The sources from which each specimen was obtained is also provided, but the detail behind these sources is minimal, since each sources is listed as either human, animal, food, environmental, or missing. (Ward et al. 2004) While these source types provide some degree of clarity, the lack of specific details pertaining to these sources is unhelpful for further analysis of the relationship between sources and strains.

An important first step in sorting the impact of listeria on humanity is examining the variance of severity among different strains. It has been shown (Muchaamba et al. 2021) that strains possessing serovar 4B have led to longer virus survivability in organisms (zebrafish were used), which is instrumentally tied to worse clinical symptoms overall. (Muchaamba et al. 2021) This kind of distinction is of great importance to the medical examination of listeria, since any implication of differing severity across strains could better inform our decisions when faced with the isolation of one strain versus another. A specific result from this study showed that LI strains of listeria yielded an 85% mortality rate in the tested zebrafish population, whereas the LII strain only yielded a 17% mortality rate and the LIII strain only yielded a 2.5% mortality rate. These survival rates were fitted using a standard Kaplan-Meier estimation curve. (Muchaamba et al. 2021) From this report, it is clear that not all strains of listeria pose an equal threat to biotic organisms (the study uses a population of zebrafish but suggests that the gravity of the results can be generalized to human populations), but further trends have yet to be identified in this study. Specifically, a crucial point of interest is whether or not different sources of listeria are associated with more or less aggressive strains.

Studies have also focused exclusively on strains of listeria that have proven to be most severe compared to their counterparts. The study lists other subspecies of listeria that almost always do not cause human illness, but then proceeds to discuss the most potent strains of *Listeria monocytogenes* that were encountered. The 26 strains arose in Bucharest, Romania and were all clinically isolated from a total of 24 patients who were presenting with conventional symptoms over the years 2009 to 2013. (Borcan et al. 2014) Three clinical origins were specified: blood cultures, placenta swabs, and cerebrospinal fluid. Over half (16 in total) of the samples came from cerebrospinal fluid, while only a single isolate came from a placenta swab. The most common serovars among these isolates for the hospitalized patients were 1/2a, 1/2b, 3a, and 3b, but serovar 4b was also prominent among the selection of specimens. All of these serovars demonstrated resilience to traditional unifaceted antibiotic treatment, but there was no significant resilience shown against multidrug methods. Collectively, this study again demonstrates that the most severe strains of listeria present specific genetic characteristics that separate them from other strains, yet this approach is limited to solely clinical data, and does not offer insight into where these people could have encountered listeria. (Borcan et al. 2014)

A more expansive approach to investigating the source-severity dynamic of listeria in the human population examines the movement of listeria through several *different* food products across eastern Europe during two distinct time periods (from 2001 to 2005 and 2019 to 2020). (Psareva et al. 2021) These strains were only of two distinct lineages (either LI or LII) and they all came from food products broadly consisting of dairy, meat and poultry, and fish. Strains from the first outbreak period (from 2001 to 2005) were of greater serovar diversity and contained a larger proportion of the LI strains of listeria, which as previously mentioned (Muchaamba et al. 2021) have been associated with worse symptomatic manifestations overall. Furthermore, LI strains were shown to be associated with dairy products in greater proportion than other products included in the study, whereas LII strains were shown to have originated from a more balanced proportion of dairy products versus fish and meats products (combined). It was shown that dairy products have a statistically significant association with greater specimen diversity when compared to the other two groups. (Psareva et al. 2021) Furthermore, the essential conclusion of this study was that dairy products seem to serve as the main origin of LI strains in this outbreak, which is instrumentally tied to the severity of

infections from dairy products since LI strains have been shown to yield greater clinical severity in the past. (Muchaamba et al. 2021) This study therefore points us in an important direction when it comes to matching sources to strains, since it indirectly proves an association between more severe strains and dairy products.

To better understand which specific kinds of listeria are found where, several methods have been used to group (or more accurately cluster) them together. A study conducted less than three years ago made use of single linkage clustering in the process of backtracking and forward checking the propagation of listeria through meat distribution in the case of a particular provider. (Luth et al. 2020) This case's methodology was a promising point of inspiration for us, since we wondered whether or not we could employ similar methods from a more direct, intrinsic source. The study employed single-linkage clustering to group different listeria outbreaks and isolates by genetic makeup through a process called core genome multilocus sequence types (cgMLST). (Luth et al. 2020) The empirical rule for assigning clusters was that two isolates would be placed together as long as they had less than eleven genetic allele differences in accordance with the dictation of cgMLST. While this means that the strains were not directly clustered by isolation source (likely due to the fact that the potential sources were limited to either food, food processing, or clinical detection), the sources were readily comparable with the clusters themselves through several visuals provided in the study. Most notably, two clusters comprise all of the clinical sources in listed in the data, while the remaining 15 clusters have strictly isolates from food or food processing environments.

These two clusters are both part of a single outbreak that occurred in Germany over the years 2013 to 2018, with all of the cluster 2 cases falling between 2015 and 2017 and the cluster 1 cases spanning the entire period. Cluster 1 had 72 cases in total and cluster 2 had 11 cases in total, resulting in the outbreak consisting of 83 total cases. The remaining clusters are not discussed at length, which is not empirically helpful, but it is mentioned that clusters 9, 10, and 12 through 16 all had an identical medical feature (they all had the same virulence factor composition, which consists of genes associated with the outbreak clusters) that tied them to the two outbreak clusters. (Luth et al. 2020) In a general sense, a Mann-Whitney U test done by the study found that gene counts pertaining to virulence in clusters 1 and 2 differed extremely significantly (with a p-value less than 0.0001) from the gene counts pertaining to virulence of all the other clusters. While this realization is important, it is somewhat obvious given the background the study provides, and does not critically examine the potentially significant association between source and strain severity. (Luth et al. 2020) This study made promising progress in sorting the strains of listeria by genetic differences, going even further than sorting by serovar, and its use of traditional clustering methods was relatively successful and highlighting some trends pertaining to the disease, but it is important to explore other, more general methods when studying a disease that had been shown to be multifaceted and intrinsically complicated.

More complex genetic clustering has also been done on strains of listeria, which has yielded some more insight into how useful a newer approach to clustering could be. Whole genome sequencing with SNP clusters (similar to what we discuss in the next section) has been used to provide a more detailed, empirical analysis of similarities and differences between different strains of listeria. (Chen et al. 2017) This data focused on specific similarities between strains found in either environmental samples or from ice cream, all of which were confirmed to have contributed to an outbreak of listeriosis in the United States. This kind of granular analysis is critically important to our approach, as it demonstrates the potential success that can surface from building multifaceted, similarity-oriented clusters from genetic data. One feature that is lacking in this analysis is the incorporation of other information, such as self-same distance or time of isolation (the latter is naturally missing because only one outbreak is considered). Consequently, a generalization of this approach to a broader dataset would be of great use to public health officials, since it would allow for a more in depth examination of strain similarities.

It is clear that clustering analysis is a natural approach when examining the behavior of listeria in the context of public health. The sources we have referenced above make good progress in this regard, but fail to incorporate a broader collection of observations and cannot utilize more dynamic evaluation schemes (like those that make use of weighting). We will build upon this research by constructing unweighted k-means clusters and weighted network clusters using the information we have available, and then compare our models to pre-existing genetic clusters. This will provide a more dynamic picture of listeria's behavior in the context of public health and provide the methodological framework for identifying high risk strains (i.e. those with serovar 4b) and the sources from which they most commonly originate.

Data Exploration

The data we will be working with is taken from the National Library of Medicine which is operated and organized under the auspices of the National Institutes of Health. (NLM:NCBI 2022) The central contributors to this database are the CDC, the FDA, the USDA, and PHE (Public Health England). Collected by numerous different agencies this database encompasses a wide variety of often profoundly unreliably labeled variables, some of which are missing in great quantities as discussed below. This dataset contains an excessive array of variables, many of which are either missing in great quantities or not relevant to our study, so we will begin by discussing the most important variables first. In terms of the collection of this data, researchers submit their results to the NIH for the NLM and report variables that they have for each isolate. As we will see, this leads to many different ways of denoting the same feature for a given observation, which makes comprehensive analysis extremely difficult. We discuss examples of this below. Several variables (for example, serovars) are recorded clinically uses standard biological procedures, whereas other variables are automatically observed by the researchers (like location). It is up to the submitter, however, to properly list each variable, for even if the information is available, it may not always be recorded. This provides a rough outline of how the data in the source was collected and where it came from.

We begin by explaining each covariate and the percentage of entries missing for that given variable (which accompanies each variable name in parentheses). The strain (17%) is given, along with its serovar (81%), the organism's group (not missing), the outbreak (not missing, assuming all empty entries are sporadic), the location of the isolate (12%), the isolation's source (22%), the date of collection (16%), and two minimum distance variables. The distance variables record the minimum distance from a given isolate to an isolate of the same strain (27% missing) and an isolate of a different strain (49% missing). In our analysis, we only use the self-same distance, because this is the most useful metric in inspecting how an isolate spreads around and how far it can spread. We note that these distance variables denote the distance between to isolates in a given SNP cluster (an organization of the isolates by genetic information), which is a variable that we do not use directly, since its influence is felt through the use of the distances. Furthermore, we also consider the number of contigs (genetic components of the isolates, where N50 is a related contig property that represents additional genetic information in our case) a given isolate has.

Naturally, there are many more variables in this dataset that are not of immediate use to us, either due to missingness of bureaucratic irrelevance. We will quickly go through these variables and justify their omission in our analysis. Which initiative (or bioproject) is recorded completely, but since any demographic information would be better represented through location and time data, we omit it. Software version variables and analysis type variables are not of any use to us since they are not tied to the biological properties of listeria or its origin. Enzyme pattern variables (93% for primary and 94% for secondary), host disease (87%), phenotype and genotype variables (100% for both), computed types (100%), the isolate's host (78%), stress genotype (65%), and IFSAC category (65%) are all missing in such profound percentages that using them as covariates would be impractical. Specifically, if we were to try to apply any method to replacing them (either through inverse probability weighting or imputation), we would either be working with an extremely small amount of data (in the case of weighting) or we would be engaging in a excessively and wastefully laborious process (in the case of imputation). Consequently, we only use completed observations that have all of the necessary data, which leaves us with a considerable amount of data anyway.

Already, it is evident that this data will be challenging to work with and analyze numerically. This is obviously due to the fact that so few of the variables we have to work with are numerical, which means direct quantitative analysis difficult. This is one of the primary limitations of the dataset that we will have to circumvent in our analysis, since visualization and statistical analysis dependent heavily on the numerical inputs of a given dataset. Another immediate limitation is the intense missingness of so many variables, as previously mentioned and discussed further below. Many of the variables we would like to consider are missing in drastic quantities, which render them unusable in most modeling contexts and even in some exploratory contexts. As a result, we have to fix our focus on a specific set of key variables that we can explore and investigate, in order to get a better sense of any possible modeling avenues that seem fitting or any basic trends that are worth observing and addressing when it comes time to implement a model.

An immediate adjustment can be made for the sources of listeria, since we technically have two variables to work with. In addition to the source type listed above, there is another variable ("isolation type") that has a smaller percentage of missing entries (11%), but is far less informative since it only lists entries as either environmental or clinical. However, it is better to have *some* information on the source of isolation than none at all, so we can conflate these two variables to improve the missingness of the isolation data. Essentially, if any observation in the specific isolation variable is missing, but there is a general observation (either environmental or clinical), we fill in the empty entry with the general information available. Through this transformation, we reduced the missingness on the isolation source variable, which is a significant improvement, and since there were some entries in this column that were already vaguely listed as either environmental or clinical to begin with, we have not lost a great deal of information by performing this transformation. Now that we have shaped, contextualized, and focused this dataset into a structure that is more manageable, we can begin to explore the trends concerning distributions that arise naturally among the different covariates.

The usage of dates throughout this dataset is spurious and quite challenging to work with. Some of the entries provide exact days and months, some only months, but most of the dates are either missing or provide only the year. However, enough provide yearly information that we can analyze the change in our data over time to a least some extent. We have divided the data into quarter-years and restrict our analysis to data points between the years 2017 and 2022. With this, we can examine the probability densities of our minimum distance variable over this period. Figures 1 and 2 below show the densities of the minimum distance to another identical isolate by each yearly (four-quarter) group from 2017 to 2022. Figures 3 and 4 show the densities of the number of contigs grouped into the same yearly divisions. Note that we include outliers in our visualizations in order to emphasize the wide range of possible values taken on by these isolates (in the case of both variables).

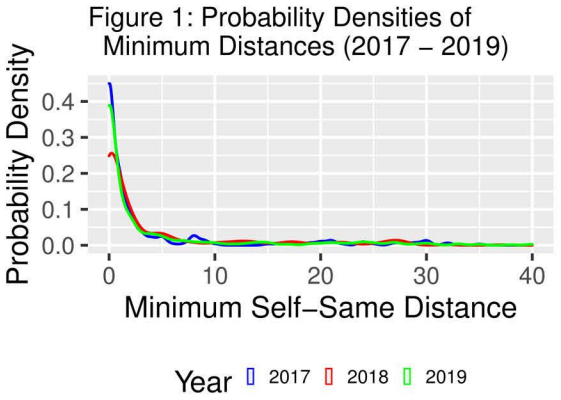


Figure 1: Probability Densities of Minimum Distances (2017 – 2019)

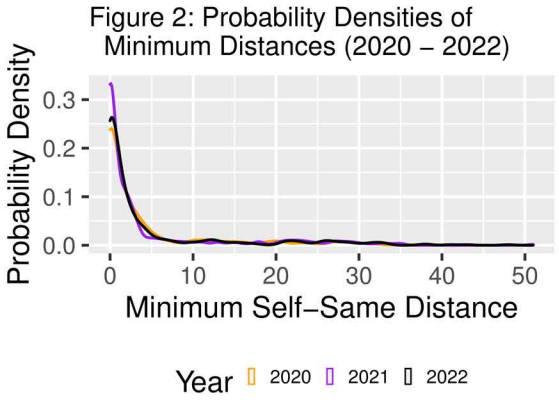


Figure 2: Probability Densities of Minimum Distances (2020 – 2022)

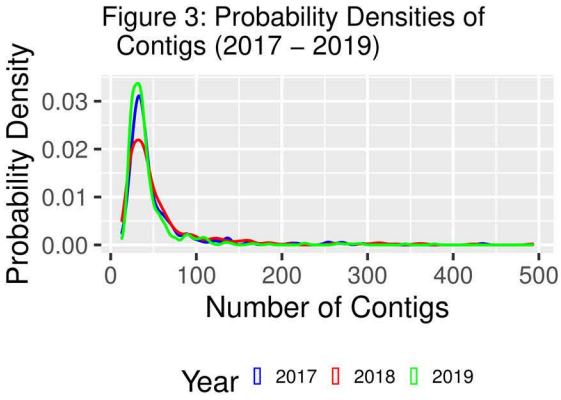


Figure 3: Probability Densities of Contigs (2017 – 2019)

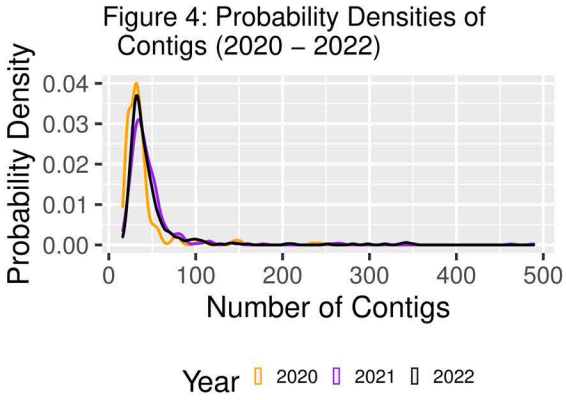


Figure 4: Probability Densities of Contigs (2020 – 2022)

This information is important to us when we are planning our analysis approach of this data. We can see that the years 2017, 2019, and 2021 all had minimum self-same distances that were more likely to be smaller (so that they were closer to identical strains on average) when compared to the other years. These kind of

temporal trends are important to make note of, since certain years appear to exhibit a greater tendency for strains to be close together when compared with other years. Furthermore, similarities in distance densities are also somewhat reflected in the contig densities, though not exactly. Most notably, the contig densities have a shape that closely resembles an F distribution, while the distance densities closely resemble an exponential distribution. Additional exploration can be done regarding the distributions of our variables when grouped by source as well, which provides further insight into the dynamics of the dataset.

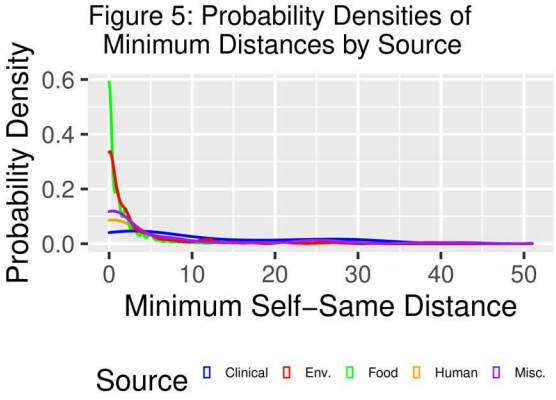


Figure 5: Probability Densities of Minimum Distances by Source

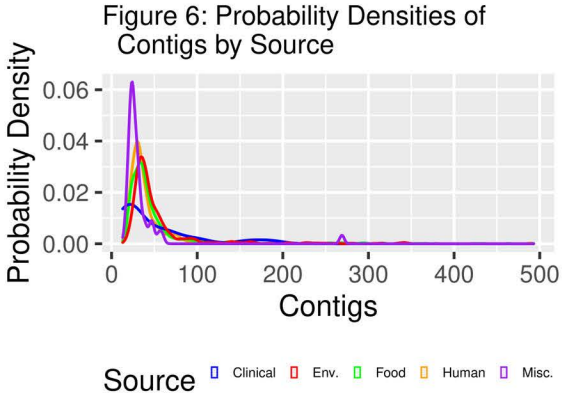


Figure 6: Probability Densities of Contigs by Source

Again, we see that the distributions of these quantities differ noticeably by source, just as in our investigation of yearly effects. Food sources seem to yield isolates that are closest to identical isolates, which makes sense in the context of how listeria emerges in food-processing situations. One bad collection of food products will arrive in the same place and ultimately effect the same region, which would lead to several identical isolates within the same area. Environmental sources also followed this trend, which is justified by the same line of reason used in the case of food products. The remaining sources did not seem to keep with this pattern as evidenced by their densities, which would lead us to believe that other sources do not yield the same centralized effect that these two do. This is an important point that we will address in our clustering analysis.

Finally, we can examine the baseline proportions of serovar 4b strains in the dataset (before any clustering is actually done). Table 1 below contains the proportions of serovar 4b for each source group so that we can see how these proportions relate to the presence of serovar 4b in each of our clusters. As we see, clinical isolates lead in terms of serovar 4b prevalence, which makes sense since a great deal of clinical data is (presumably) extracted post-mortem (as in many cases of cerebrospinal fluid). Human cases also demonstrate a prevalence that is more prominent than other sources (though much less than clinical isolates.

Table 1: Serovar 4b Prevalence by Source

Source	Serovar 4b Prevalence (%)
Human	2.07
Food	0.56
Clinical	39.13
Environmental	0.20
Misc.	0.00

Now that we have established some of the essential trends that surface in the dataset, we can detail the precise methodology through which we will create our two separate clustering models. Note that we have not discussed much about the SNP genetic clusters that exist in our dataset already, since this data will become more relevant in the form of a baseline comparison with our newer models. We can use a variety of methods to compare these two approaches, ranging from elementary to complicated. Collectively, these comparisons will allow us to comment on whether or not (or, more generally, to what degree) we can generalize our approaches to non-genetic information when trying to identify isolation sources that elicit the greatest chance of harm to human beings.

K-Means Methodology, Software, and Diagnostics

Our analysis is conducted in the R programming environment. (R Core Team 2018) Our clustering analysis will be carried out using the `MASS` (Venables and Ripley 2002) and `cluster` (Maechler et al. 2021) packages, with supplemental diagnostics from the `factoextra` (Kassambara and Mundt 2020) and `clValid` (Brock et al. 2008) packages. Visualizations of clusters include color-coded scatter plots to examine cluster wide behavior in view of continuous numerical variables, which allows us to qualitatively assess the behavior of different variables within different clusters. In terms of the technical mechanisms behind our chosen methodology, we will employ the conventional k-means clustering method using contig, N50, minimum distance, year, and numerically indexed source data. To this end, we introduce the following formalisms in our explanation. Let $\mathbf{X} = \{X_m\}_{m=1}^{n}$ be our set of n numerical observation vectors (each having components for each variable listed above) and let $\mathbf{\Sigma} = \{\Sigma_j\}_{j=1}^{k}$ be a set of clusters that exhaustively divide $\mathbf{X}$ for some $k \leq n$. For each set Σ_j let μ_j be the mean of all points contained therein. Then the k-means algorithm finds the set of clusters $\mathbf{\Sigma}^*$ that satisfies:

$$\mathbf{\Sigma}^* = \underset{\mathbf{\Sigma}}{\operatorname{argmin}} \left\{ \sum_{j=1}^{k} \sum_{\{X_m \in \mathbf{X} \ | \ X_m \in \Sigma_j\}} ||\mathbf{X}_m - \mu_j||_2^2 \right\}$$

where $||\cdot||_2$ denotes the L^2 norm whose square ($||\cdot||_2^2$) is the sum of squared components for any given vector input. The implementation of this method corresponds to the `cluster` and `MASS` packages in R. Naturally, a point of concern that must be addressed is the number of optimal clusters to be used in our analysis. The process of choosing this value overlaps with our discussion of diagnostics below.

The `factoextra` and `clValid` packages provide numerous diagnostic methods that not only allow us to assess the performance of our model in selecting $\mathbf{\Sigma}^*$, but also allow us to choose the optimal number of clusters to create. The primary method of examining our model's performance and choosing the optimal number of clusters will be the silhouette coefficient. For some $X_m \in \Sigma_m$, we define the following:

$$\alpha(X_m) = \frac{1}{|\Sigma_m| - 1} \sum_{\{X_r \in \Sigma_m \ | \ X_m \neq X_r\}} ||X_m - X_r||_2$$

$$\beta(X_m) = \min_{\{r \neq m\}} \left\{ \frac{1}{|\Sigma_r|} \sum_{\{X_r \in \mathbf{X} \ | \ X_r \in \Sigma_r\}} ||X_m - X_r||_2 \right\}$$

$$\xi(X_m) = \begin{cases} \frac{\beta(X_m) - \alpha(X_m)}{\max\{\alpha(X_m), \beta(X_m)\}} & |\Sigma_m| \neq 0 \\ 0 & |\Sigma_m| = 0 \end{cases}.$$

For any given observation vector, the silhouette coefficient ($\xi(X_m)$) falls between -1 and 1, with 1 indicating the strongest, well-matched cluster placement for that observation, 0 indicating an indifferent placement of the observation, and -1 indicating the strongest **poorly-matched** cluster placement for the observation. Part of our visuals will be plots of the average silhouette coefficients across all observations for different numbers of chosen clusters. Relevant software is included in the `factoextra` and `clValid` packages, specifically in terms of generating the previously mentioned plots, which are more economic in terms of defending our chosen cluster count when compared to computing silhouette coefficients manually. Our main method of comparing clusters created by different methods will be to compute Rand index scores using the `fossil` package (Vavrek 2011), which we discuss later. This will allow us to measure similarity across different modeling approaches, which is helpful for aggregate comparisons. Specifically, such aggregate comparisons will allows us to determine the degree of redundancy in either of our models. If either our k-means approach or network approach are identical to the SNP cluster model, then there would be no need to create them, since they require additional computational exertion and contribute very little additional information.

Network Clustering Methodology, Software, and Diagnostics

The network clustering approach that we employ is exceedingly complex and nuanced, making use of aspects from graph theory to machine learning. We have accordingly tried to condense our theoretical explanation of the methodological underpinnings of this approach. In short, we follow the methods outlined in the assembly of the `linkcomm` package in R. (Kalinka and Tomancak 2011) This approach allows for the creation of cluster-like "communities" that are more flexible in terms of their membership features. Membership to a given community is not binary, as it is in the k-means approach, but rather a string of proportions (or more stochastically probabilities) that indicated the degree to which a given node belongs to a given community. In an illustrative example, a network may consist of three points that are grouped into two communities. Point 1 could belong entirely to community 1, point 3 could belong entirely to community 2, and point 2 could hold 50% membership in community 1 and 50% membership in community 2. The main benefit that this approach offers is this generalization of cluster membership, which allows for overlapping trends (either because of similar timing, sources, or distance) to be better processed in the grouping of data points.

At the core of this network-style comparison of similarities is a set of weights that we assign to the data we have. Firstly, we note that there are 2290 unique strains that we are considering, which means that there are $\frac{2290^2}{2} = 2622050$ nontrivial strain combinations to consider. This presents issues in itself, which we address later, but for now, we can outline the weights we apply to these combinations. To this end, we define the function

$$[x]_\eta = x \text{ rounded to the nearest order of magnitude } \eta$$

(so if $x = 48946$, $[x]_3 = 49000$, but $[x]_1 = 48950$). Also, note that when we discuss an observation X_m, we actually mean to write:

$$X_m = (\text{N50}_m, \text{Contig}_m, \text{Source}_m, \text{Distance}_m, \text{Quarter}_m).$$

Hence, for each nontrivial observation pair (X_m, X_r), we define the subweights:

$$\omega^{(1)}_{m,r} = \begin{cases} \frac{0.4}{3} & [\text{N50}_m]_3 = [\text{N50}_r]_3 \\ 0 & [\text{N50}_m]_3 \neq [\text{N50}_r]_3 \end{cases} \quad \omega^{(2)}_{m,r} = \begin{cases} \frac{0.4}{3} & [\text{Contig}_m]_1 = [\text{Contig}_r]_1 \\ 0 & [\text{Contig}_m]_1 \neq [\text{Contig}_r]_1 \end{cases} \quad \omega^{(3)}_{m,r} = \begin{cases} 0.3 & \text{Source}_m = \text{Source}_r \\ 0 & \text{Source}_m \neq \text{Source}_r \end{cases}$$

$$\omega^{(4)}_{m,r} = \begin{cases} \frac{0.4}{3} & [\text{Distance}_m]_1 = [\text{Distance}_r]_1 \\ 0 & [\text{Distance}_m]_1 \neq [\text{Distance}_r]_1 \end{cases} \quad \omega^{(5)}_{m,r} = \begin{cases} 0.3 & \text{Quarter}_m = \text{Quarter}_r \\ 0 & \text{Quarter}_m \neq \text{Quarter}_r \end{cases}$$

From these subweights, we define the weights for the interaction between observation X_m and observation X_r:

$$0 \leq \omega_{m,r} = \sum_{q=1}^{5} \omega^{(q)}_{m,r} \leq 1.$$

In terms of our network (a graph with each node representing a specific strain of listeria and each edge between two nodes representing an interaction of similarity), these are the weights we assign to each edge. We do not consider the directions of such edges (meaning that the connection between strain A and strain B is taken to be the same as the connection between strain B and strain A). With these weights, we pair each interaction $(X_m, X_r) = (X_r, X_m)$ with the weight $\omega_{m,r}$, after which we can construct the network itself.

As mentioned, the entire construction process is intensely involved, but the fundamental metrics off which the approach is based are essential in a methodological discussion. In the case of weighted nodes (as in our case), it is important to address the method by which node-sharing links are declared "similar" and to what extent, since this is the foundation of the network approach. Adapting the notation from the documentation

listed in the `linkcomm` package (Kalinka and Tomancak 2011), let a_m be the vector of all weights concerning immediate interactions with strain (node) m. Let $\epsilon_{m,q}$ and $\epsilon_{r,q}$ be two strain interactions (links) that share the common node q. This approach makes uses of Tanimoto's distance metric, defined as:

$$\mathcal{D}(\epsilon_{m,q}, \epsilon_{r,q}) = \frac{\langle a_m, a_r \rangle}{||a_m||_2^2 + ||a_r||_2^2 - \langle a_m, a_r \rangle} = \frac{\langle a_m, a_r \rangle}{\langle a_m, a_m \rangle + \langle a_r, a_r \rangle - \langle a_m, a_r \rangle}$$

where $\langle \cdot, \cdot \rangle$ denotes the conventional vector inner product defined by:

$$\langle v, u \rangle = \sum_{s=1}^{t} v_s u_s, \ \forall v, u \in \mathbb{R}^t \ni \langle v, v \rangle = ||v||_2^2 = \sum_{s=1}^{t} v_s^2.$$

This metric is used to inform the distances created in our network model across all possible strain combinations.

In terms of intra-model diagnostics, there are two quantities of interest that we can examine in after the creation of our model, which we can define using an adapted version of the notation introduced in the `linkcomm` documentation. (Kalinka and Tomancak 2011) Firstly, we consider the community centrality, a measure that incorporates the Jaccard coefficient and will give us an idea of how well our network separates the data. First, we note that the Jaccard coefficient is given by:

$$\mathcal{J}(C_1, C_2) = \frac{|C_1 \cap C_2|}{|C_1| + |C_2| - |C_1 \cap C_2|}$$

where C_1 and C_2 are communities (hereafter clusters) that have been created in our model. (Note the structural similarity to Tanimoto's distance metric). Let M be the set of clusters to which node m belongs, and let $M_R = \{m \in R \mid R \subset M\}$ and $M_{R,S} = \{m \in (R \cap S) \mid (R \cap S) \subset M\}$. From this, we define the community centrality of strain (node) m as:

$$\mathcal{C}(m) = \sum_{m \in M_R} \left(1 - \frac{1}{|M_{R,S}|} \sum_{m \in M_{R,S}} \mathcal{J}(R, S) \right).$$

This measure gives us a sense of the overlap, individuality, and centrality of the connections established for a given node m, which can be helpful when evaluating practicality and functionality in applied settings. Additionally, we can define node modularity for a cluster C, which is given by:

$$\mathcal{M}(C) = \frac{\mathcal{L}^+(C)}{\mathcal{L}^-(C)} \cdot \frac{2\delta}{|C| - 1}$$

where $\mathcal{L}^+(C)$ is the number of interactions in cluster C alone, $\mathcal{L}^-(C)$ is the number of interactions **not** in cluster C, and δ is the average degree (number of interactions for a given strain) across all strains in the network. This measure allows us to inspect the degree of separation for a given cluster, which will aid in determining how cohesive the distinctions we have made truly are, and subsequently, whether or not they were worth making at all.

These two metrics each serve a similar purpose, but on a different scale, which is why it is essential that we include both in our analysis. Centrality is evaluated at the strain level, which means that it is a reflection of separation and connection for given strains of listeria in the context of other isolates. Consequently, this measure possesses a great deal of clinical value, since it is an effective way of examining which strains are most related and which are most different. On the other hand, modularity evaluates differences at the cluster level, which while useful for analytical purposes, restricts investigations of difference to an aggregate level. The goal of creating this clusters is to establish a model in which one can freely move from levels of generality (strain to cluster to dataset, and vice-versa), so modularity is included primarily to compare the broad degree of separation in a network and how this compares to other models.

Dimensionality Issues and Sampling Methods

Traditional k-means methodology is easily applied (in the computational sense) to a dataset of the size we are working with. To this end, no further methodological action is needed. However, network clustering is a far more intensive procedure as we have a previously mentioned. Though our refined dataset only has 2290 distinct observations (strains), this directly requires us to evaluate a total of 2622050 nontrivial strain interactions. Initial attempts at this (from a purely analytical perspective) proved impossible, since it took in excess of 30 minutes to make basically trivial computational progress. Additionally, it is impossible to render any kind of useful network visualization with over 500 strains due to practical font sizing issues and shading limitations. To resolve these problems, we introduce an elementary sampling approach for the partial creation and evaluation of our network clustering model. First, we randomly sample 1000 observations from the original dataset of 2290 observations when performing network clustering analysis. This allows us to create a recompiled network interaction dataset with 999000 entries. From here, we apply the following sampling schemes.

Visualization Sampling

Our approach to visualization is fairly straightforward. Since the highest number of strain combinations that can be easily visualized using a Fruchterman-Reingold (FR) graph (Fruchterman and Reingold 1991) is 500, the algorithm for which is the mathematical underpinnings that assemble our clusters as well. While the details behind this process are beyond the scope of this paper, the visuals are essentially large groups of connected nodes accompanied by small pie charts indicating cluster membership percentages. It is essential that we preserve the readability and presence of these small pie charts, since they allow for quick, qualitative assessments of strain-to-strain dynamics to be made. However, this is not possible for the entire set of possible strain pairings. Therefore, we adopt the following partial analysis sampling scheme to create two sub-visualizations of the data:

$$\begin{cases} \text{[1] Draw 500 randomly sampled strains from the dataset and cluster them into an FR graph.} \\ \text{[2] If the algorithm is successful, draw another 500 randomly sampled strains without replacement.} \\ \text{If not, repeat step 1.} \\ \text{[3] Cluster the second sample into an FR graph. If the algorithm is successful, stop.} \\ \text{If not, repeat steps 2 and 3.} \end{cases}$$

We note that the FR graph algorithm may not always converge due to overt difference within a "bad" sample. Hence, we simply repeat the process until this works by testing different random seeds (see below). The outcome here is two FR graphs showing the interactions between 500 randomly selected strains (for a total of 1000 interactions). This is only a small number of the possible interactions, and our analysis is certainly limited, but at least these visuals will allow us to make some general remarks about the behavior of these strains of listeria.

Analytic and Diagnostic Sampling

If we do not need to render a visualization of the networks we create, then we have a greater degree of flexibility regarding the number of strains we can computationally handle in a subnetwork. Initial trial-and-error implementations showed that samples of up to 1000 strains are plausible if only diagnostic and analytic computations are needed. Therefore, we can still implement FR network clustering as we did before, but now with 1000 randomly sampled strains, which allows from a greater deal of possible interactions and more comprehensive metrics. The only other method that was first considered plausible is to set edge limits on visualizations, but not only did this prove incredibly limiting, it also yielded little to no improvement due to the immensity of our graph and the nuance of our weighting system. The only drawback to our sampling approach is the amount of time and trial-and-error required to find a convergent seed across a sufficient number of sub-networks, an issue that we will discuss in greater detail later on. Hence, we implemented the following similar analytic sampling scheme:

[1] Draw 1000 randomly sampled strains from the dataset and cluster them into an FR graph.
[2] If the algorithm is successful, draw another 1000 randomly sampled strains without replacement.
If not, repeat step 1.
[3] Cluster the second sample into an FR graph. If the algorithm is successful, **continue**.
If not, repeat steps 2 and 3.
[4] Repeat steps 1, 2, and 3 until **five** FR graphs have been successfully constructed
(**without** replacing previous samples).

This will allows us to get an idea of the broader behavior between strains in the dataset, in that we can compute diagnostics and appropriate metrics for each subnetwork and compare them. Additionally, we it comes to computing specific centrality values, we will simply select 10 random strains from a given network and compare the values overall, with aggregate measures also included. Finally, we list the seeds we used in drawing these samples below in table 2.

Table 2: Seeds Used for Sampling

Action	Seed
Drawing 1000 Observations From the Original Dataset	3
Constructing Visual Networks With 500 Interactions	18
Constructing Analytic Networks With 1000 Interactions	7

Inter-Methodological Comparative Diagnostics

Finally, we can compare the performance of our k-means model, network model, and the original SNP model provided in the dataset. Our main method of comparing similarity will be the Rand index, which we define through some additional notation. Let $\mathbb{M}$ denote our k-means model, $\mathbb{N}$ denote our network model, and let $\mathbb{S}$ denote the pre-existing SNP model. Let $O(\triangle^{\pm}, \square^{\pm})$ denote the set of observations within the same (+)/different (-) cluster in model $\triangle$ and also within the same (+)/different (-) cluster in model $\square$ (for example, $O(\mathbb{M}^{+}, \mathbb{N}^{-})$ denotes the set of observations that are in the **same** cluster in the k-means model ($\mathbb{M}$), but in different clusters in the network model ($\mathbb{N}$)). Then the Rand index for two models, $\triangle$ and $\square$ is given by:

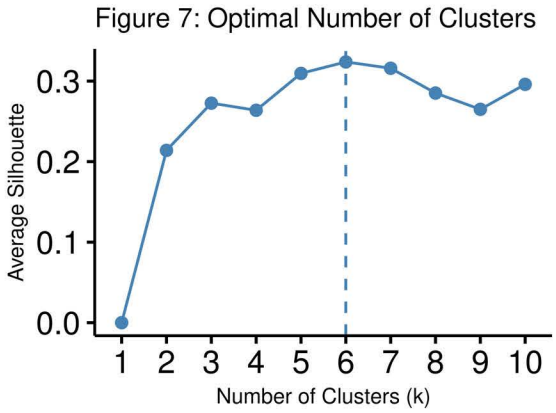
$$\mathcal{R}(\triangle, \square) = \frac{O(\triangle^{+}, \square^{+}) + O(\triangle^{-}, \square^{-})}{O(\triangle^{+}, \square^{+}) + O(\triangle^{-}, \square^{-}) + O(\triangle^{+}, \square^{-}) + O(\triangle^{-}, \square^{+})}.$$

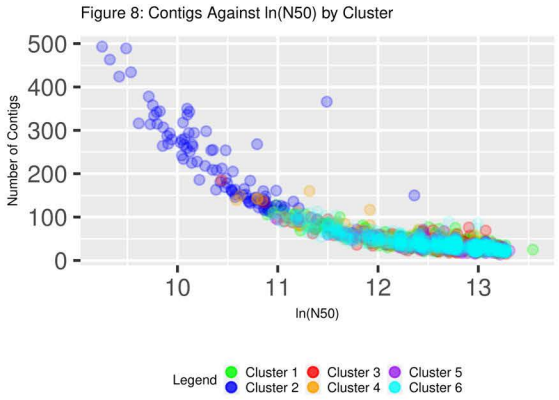
Naturally, we will compute $\mathcal{R}(\mathbb{M}, \mathbb{N})$, $\mathcal{R}(\mathbb{M}, \mathbb{S})$, and $\mathcal{R}(\mathbb{N}, \mathbb{S})$, which will allow us to gauge the similarity of our models with each other and with the pre-existing model. High levels of similarity between either of our two models and the SNP models would be important and promising for clinical applications, since this would essentially suggest that the basic SNP genetic clustering method provides as much nuance as more multifaceted approaches like k-means and network clustering. This would also allow us to assess redundancy in our approach, which is important given the computationally expensive nature of our models.

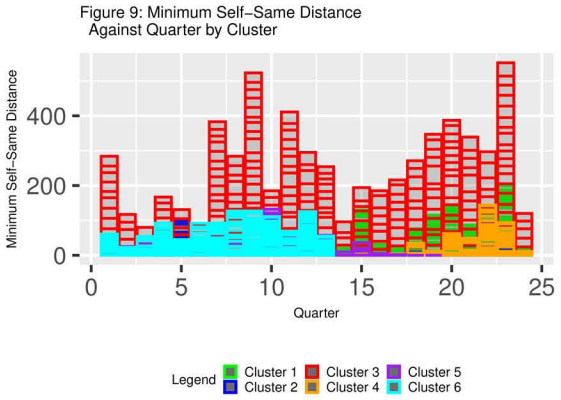
Finally, some elementary summary statistics are of great importance to us despite their simplicity. For each cluster in our k-means approach, we compute the proportion of strains with serovar 4b, which gives us an idea of relative severity between such clusters. Additionally, we examine the dominant sources for each k-means cluster, so that we can determine bivariate trends in regards to severity and source. In our network clustering model, we can also compute these values for each of the five 1000 sample networks. These values may become tiresome to compute and list (as in the case of the SNP clusters), so it is more instructive to compute serovar 4b prevalence values at the network level in order to get a rough idea of the severity dynamics.

Results

We begin by assessing the results from our k-means model. From our initial investigations of silhouette coefficients for 1000 sample draws, the optimal number of clusters fell consistently within the range of 3 to 9, averaging around 6 (as in the case with a seed of 3, which is what was used to create figure 7). As such, we created 6 k-means clusters as shown below.







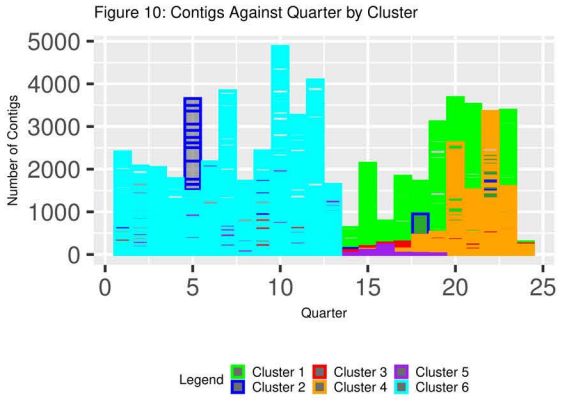


Table 3: Serovar 4b Prevalence by Cluster

Cluster	Dominant Sources (Source Total/Cluster Total)	Serovar 4b Prevalence (%)
1	Human (232/604), Food (372/604)	2.32
2	Human (128/251), Food (102/251)	1.05
3	Human (115/833, Food (709/833)	1.20
4	Environmental (229/269), Misc. (34/269)	0.00
5	Food (62/95), Environmental (22/95)	0.74
6	Environmental (232/238), Misc. (6/238)	0.84

As we can see, clear trends between variables and clusters exist, specifically in regard to figures 2, 3, and 4. Clusters 2 and 6 consistently show lower minimum self-same distances and contig counts, yet seem to oppose each other in terms of N50 values, with cluster 2 exhibiting much higher N50 values and cluster 6 exhibiting much lower N50 values. Serovar 4b prevalences differ noticeably across the six clusters, but not in a way that shows a consist trend in terms of our numerical variables (for example, clusters 4 and 1 both consist of observations from closely positioned quarters, yet opposingly show that lowest and highest serovar 4b prevalence, respectively). To better corroborate these results, we turn to the outcomes of our network clustering model, which allows for more elaborate interactions.

The two FR graphs we obtained from our sampling analysis are presented below in figures 5 and 6. We note that the number of nontrivial interactions (as shown in the graphs themselves) differ vastly between the two models.

Figure 11: First 500–Observation Network (Trivial Links Omitted)

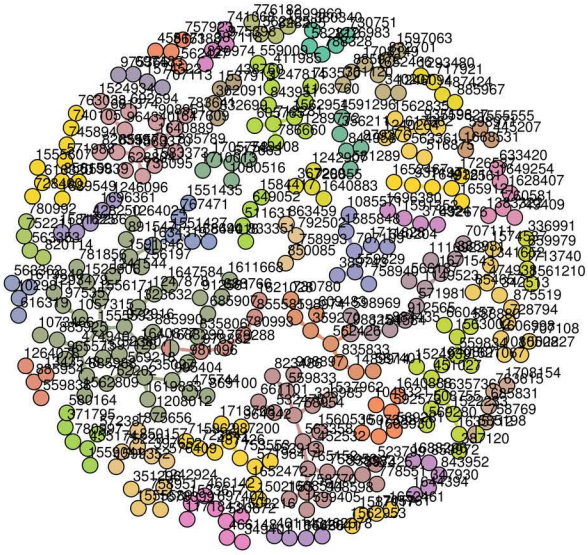
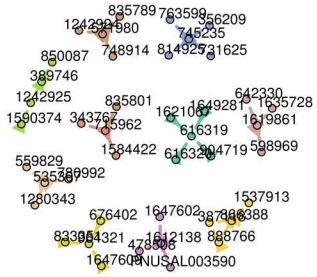


Figure 12: Second 500–Observation Network (Trivial Links Omitted)



These network diagrams clearly showcase the different possibilities concerning network interactions in this dataset. Sometimes, interactions are so limited due to dissimilarity in the data that only small, distinct communities can be established, as in figure 6. In other parts of the dataset, more elaborate (but still fairly separated) networks can be created, with some inter-cluster interaction as in figure 5. This gives us an interesting second viewpoint of our data: while the k-means visualizations showcase intense mixing between clusters in regards to certain variables, these visualizations evidently depict strong separation, which is likely due to the strict weighting scheme we have implemented (which emphasizes timing (quarter) and source). Nonetheless, a clearer analytical picture can be obtained from considering the aggregate diagnostics we have assembled using our five 1000 observation sample network, which include a wider variety of possible interactions. Given that several interactions have weights of 0 (meaning they would not be shown on a visualization like those above), aggregate analytics (mean modularity and mean centrality) will incorporate this effect in a quantitative manner that cannot be achieved qualitatively through the graphs.

As we see in table 4, our networks also reflected the variability of this dataset. Mean centralities for networks 1, 3, and 5 were all comparable, and also had similar standard deviations, though network 5 had relatively more variance than the other two. Network 4 was the most interesting of our sample, since its behavior was the most nonspecific. On one hand, it had a similar number of interactions to networks 1, 3, and 5 (they are not depicted for plausibility), but it was not as sparse as network 2, which possesses incredibly low centrality compared to the other networks. Furthermore, we note that centrality is a **node-level** metric that is evaluated for each strain, which is important to consider in the context of the remaining data. On the other hand, the modularity presented here behaviors opposite in this case. Networks 1, 3, and 5 possess values of magnitude 10^4, while network 4 possesses a value of magnitude 10^3 (though network 3 is fairly close to being in the same magnitude bracket as the others). Network 2 is again the outlier, with shockingly low modularity, again in line with its previous metrics. Notably, network 3 exhibits much higher relative variance to the other networks, and is the only network whose standard deviation of modularity exceeds its mean.

Table 4: Aggregate Network-Level Results

Network	Mean Centrality	SD Centrality	Mean Modularity	SD Modularity	N.S.4b.P. (%)
1	1.04	0.76	3.858688e+04	1.297916e+05	1.38
2	0.01	0.09	5.559012e-01	1.928324e-01	0.00
3	1.03	0.77	1.782799e+04	9.089949e+04	1.27
4	0.51	0.78	5.104336e+03	4.841547e+04	0.60
5	1.10	0.87	2.347899e+04	1.032028e+05	0.71

(Note above that N.S.4b.P. is **network-wide** serovar 4b prevalence). Finally, we wish to compare our two clustering models with the original SNP cluster model as provided in the dataset. When it comes to comparing our k-means model with the original SNP clustering model, we can easily compute the Rand index on the entire dataset: this value was approximately 0.742, indicating decent similarity between the two approaches, as we will discuss later. To make any kind of comparison with our 1000-observation networks, we can compute the Rand index for each network and the corresponding k-means and SNP clusters for the strains from the original dataset. To this end, we obtain 10 distinct Rand indices (5 for each other model), which are summarized in table 5 below.

Table 5: Rand Indices for All Network Model Combinations

Network	K Means Rand Index	SNP Rand Index
1	0.727	0.911
2	0.905	0.571
3	0.712	0.892
4	0.678	0.929
5	0.740	0.926

Evidently, there are varying levels of disagreement across our models and the original SNP model presented in the dataset. For the most part, the network clustering methods agrees quite strongly with the SNP clustering model, with the exception of network 2, which only agrees with the SNP model for a little over half the time. The results are interestingly reversed when we consider the Rand indices for the comparison between the networks and the k-means model. All of the models exhibit decent levels of agreement with the networks, placing observations similarly for about 70 percent of the time. However, network 2 demonstrates a much higher level of agreement with the k-means model than any other network, the opposite of the response to the SNP model. This is especially strange since network 2 was by far the smallest network of the five, which could perhaps be an indication of the tendency of k-means clustering to mix clusters (as shown in figures 2, 3, and 4) due to strict membership criteria when compared to the network approach. Small networks (like the one shown in figure 6) exhibit a similar strict community formation tendency, which could be why this approach has a high Rand index value when compared with the k-means model.

Discussion, Limitations, and Conclusion

It is clear from all of the models discussed herein that source and timing play important parts in contributing to the severity of any manifestation of listeria, at least when one uses serovar 4b prevalence as the dominant metric of severity. Serovar 4b prevalence differs noticeably across the five 1000-observation networks we have created for analysis, just is it differs considerably across the k-means clusters. Of course, our analysis of the networks is inherently limited by the sheer immensity of the structures: the entire network is implausible to construct as we discuss below, and even the 1000 interaction graphs can be challenging to deconstruct. However, we can see that different portions of the data exhibit greater degrees of severity, and small-scale analysis can be done on clusters extracted from our visualizations of the networks in figure 6. For example, the cluster consisting of strains 1590374, 1242925, 389746, and 850087 has no strains with serovar 4b and consists entirely of isolates originating from human sources. By comparison, the cluster consisting of strains 559829, 535307, 780992, and 1280343 also has no strains with serovar 4b, but draws 75% of its sources from human origins and 25% from environmental origins. These kinds of detailed, small scale analyses could be useful when dealing with condensed datasets, since they provide important source-severity associations that elucidate the behavior of listeria. However, conducting many of these kinds of dissections is tedious; it would be inefficient to go through all the clusters in figure 5 for example, let alone a 1000 observation network.

In terms of addressing our first goal, it is clear that variables beyond genetic information alone factor into the severity of listeria, most notably the source from which an isolate originated and the timing of when that isolate we obtained. Our k-means model shows a great degree of separation of clusters in view of certain variables; for example, the bivariate behavior concerning timing and minimum self-same distance (as depicted in figure 3) shows similar behavior for clusters 2 and 6, despite the fact that these two variables differ in terms of their N50 and contig values (which are two of the only pieces of genetic information available). Moreover, table 3 does confirm the pre-existing trends established in the literature review in that human and food sources do seem to present high serovar 4b prevalence when compared to other sources. However, serovar 4b is also present in environmental and miscellaneous sources, which is why additional numerical information is essential in determining associations between serovar 4b prevalence and sources. This concept extends to our network models as well, since the relatively high mean modularities (except in the case of network 2) would suggest that a great degree of disconnectedness exists between clusters in each of the four other networks. As such, this means that our weights (which give priority to isolates from the same source and quarter) sufficiently divide the clusters into a non-overlapping partition (as shown in figures 5 and 6, which most pie-charts are solid color, implying exclusive membership to a single group). Hence, we can deduce that numerical factors beyond genetic information alone play an important part in distinguishing isolates, which subsequently lends itself to evaluating problematic sources and other non-genetic factors. In terms of our second goal, we see that our network approach was very similar by the Rand index to the SNP model (with the exception of network 2) and it was fairly (though much less) similar to the k-means model as well (except network 2 which was very similar. This means that while the SNP approach may be reminiscent of the network approach, it is decently different from the k-means approach.

Our approaches have numerous limitations that should be addressed in future endeavors. First, the dataset is sparse and most of the data is non-numerical. This makes any form of analysis excruciating and inherently limited, so it is advisable to seek out more reliable data for other projects. Second, we were unable to use all of the remaining data to construct a full network analysis due to dimensionality issues. While our sampling approach offered a rough idea of our ideal result, it seems that this form of network clustering is ill-suited to this kind and size of data, and should only be applied to small, complete datasets. Third, our time scale (quarters) was not as granular as it should be; a more continuous time scale would be better for a clustering analysis. Fourth, our source groups were broad and vague; a more detailed analysis should focus on one of our groups (perhaps food alone for example) and divide that group into detailed sub-sources (e.g. dairy, plant-products, meat, etc.). Finally, different weighting schemes should be tested for additional network models, since ours is one of many possible setups. This would better explore the effects of the variables on interactions. Nonetheless, our models have shown that sources and both genetic and non-genetic information play a role in strain severity (serovar 4b prevalence), and we have also assessed their similarity with pre-existing genetic clustering models, while providing statistical framework that could be useful in examining and potentially preventing listeria-oriented food-borne illness.

References

Borcan, A. M. et al. 2014. "Listeria Monocytogenes – Characterization of Strains Isolated from Clinical Severe Cases." *Journal of Medicine and Life* 7. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4391364/.

Brock, Guy, Vasyl Pihur, Susmita Datta, and Somnath Datta. 2008. "clValid: An R Package for Cluster Validation." *Journal of Statistical Software* 25 (4): 1–22. https://www.jstatsoft.org/v25/i04/.

Chen, Y. et al. 2017. "Assessing the Genome Level Diversity of Listeria Monocytogenes from Contaminated Ice Cream and Environmental Samples Linked to a Listeriosis Outbreak in the United States." *PLoS ONE* 12. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0171389.

Fruchterman, Thomas M. J., and Edward M. Reingold. 1991. "Graph Drawing by Force-Directed Placement." *Software: Practice and Experience* 21 (11): 1129–64. https://doi.org/https://doi.org/10.1002/spe.4380211102.

Kalinka, Alex T, and Pavel Tomancak. 2011. "Linkcomm: An r Package for the Generation, Visualization, and Analysis of Link Communities in Networks of Arbitrary Size and Type." *Bioinformatics* 27: 2011–12. https://doi.org/10.1093/bioinformatics/btr311.

Kassambara, Alboukadel, and Fabian Mundt. 2020. *Factoextra: Extract and Visualize the Results of Multivariate Data Analyses*. https://CRAN.R-project.org/package=factoextra.

Luth, S. et al. 2020. "Backtracking and Forward Checking of Human Listeriosis Clusters Identified a Multiclonal Outbreak Linked to Listeria Monocytogenes in Meat Products of a Single Producer." *Emerging Microbes and Infections* 9. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7473094/.

Maechler, Martin, Peter Rousseeuw, Anja Struyf, Mia Hubert, and Kurt Hornik. 2021. *Cluster: Cluster Analysis Basics and Extensions*. https://CRAN.R-project.org/package=cluster.

Muchaamba, F. et al. 2021. "Different Shades of Listeria Monocytogenes: Strain, Serotype, and Lineage-Based Variability in Virulence and Stress Tolerance Profiles." *Frontiers in Microbiology* 12. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8764371/.

NLM:NCBI. 2022. "Pathogen Detection Listeria Monocytogenes."

Psareva, E. et al. 2021. "Diversity of Listeria Monocytogenes Strains Isolated from Food Products in the Central European Part of Russia in 2000–2005 and 2019–2020." *Foods* 10. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8617672/.

R Core Team. 2018. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rogalla, D., and P. Bomar. 2022. "Listeria Monocytogenes." https://www.ncbi.nlm.nih.gov/books/NBK534838/.

Vavrek, Matthew J. 2011. "Fossil: Palaeoecological and Palaeogeographical Analysis Tools." *Palaeontologia Electronica* 14 (1): 1T.

Venables, W. N., and B. D. Ripley. 2002. *Modern Applied Statistics with s*. Fourth. New York: Springer. https://www.stats.ox.ac.uk/pub/MASS4/.

Ward, T. et al. 2004. "Intraspecific Phylogeny and Lineage Group Identification Based on the prfA Virulence Gene Cluster of Listeria Monocytogenes." *Journal of Bacteriology* 186. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC451661/.

Link to Repository: https://github.com/NTProvost/PHP-2550-Project