Methods and Analysis Plan

Nathan Provost and Antonella Basso

Updated Question and State of Data

Our questions of interest have changed slightly from our last update for a variety of reasons pertaining to our original method plan which we discuss below. Specifically, we want to answer the following main questions pertaining to the behavior of *Listeria monocytogenes* using the data at hand:

Is there an association between specific isolation sources, years, distances between given isolates and other isolates, and/or genetic contig information and particularly harmful strains of listeria, specific strains with serovar 4b?

How consistent are pre-existing SNP cluster assignments with clusters formed using *k*-means clustering methods grouped by year, contig and N50 information, minimum distances, and sources?

Our updated dataset is almost identical to the one we used predominantly in our previous literature review and exploration, but we have changed our variable selection slightly. Miscellaneous isolate identifiers will still not be used and outbreak classifications will not be used, since so few data entries are classified as outbreak isolates to begin with. General minimum distance will follow the same pre-processing methods outlined as before, in which the minimum of the minimum distance between an isolate and another identical isolate and the minimum distance between an isolate and a different isolate is computed. In our analysis, we only use data for which there are recorded serovars and recorded distances. Furthermore, we will only consider isolate data from the United States to narrow our analysis down to a more feasible set of data. Additionally, as stated in our review, we will only consider the time period from 2010 to 2020, with grouping of isolates with differently labeled dates grouped into whole year groups (for instance, isolates from "April 2015" would be grouped into the 2015 group) to our best approximate capability. Building upon the exploration that we conducted, we will only consider the top 10, non-overlapping isolation sources from the United States in our analysis. These will be labeled numerically in order of most prominent to least prominent in order to be integrated into our analysis. Isolates without listed sources are not considered.

While we have emphasized the omission of missing data coming from this dataset on account of there being so few numerical variables to work with, there is an additional goal tied to our questions that implicitly arises. The patterns of missing across certain variables, specifically the given year, is important to consider in the context of listeria's behavior. Consequently, part of addressing our two questions will be commenting our certain years where different data is missing in lesser or greater quantities. This analysis will be more qualitative in nature and will be deeply tied to our review of literature, where several public health studies of specific events and periods of time are included. Appending this to the answers to our questions will provide the reader with a better picture of the data at hand in light of the fact that we will not be using the majority of it due to extensive missingness and analytical limitations due to unreliable variable naming.

Finally, it is essential to note that we have reintroduced the SNP cluster variable into our analysis. This is to allow for a comparison between the clusters that we observe and the clusters that have already been constructed using SNP information. We will be able to compare the attributes for each clustering approach (number of contigs, distance, serovar prevalence, etc.), which allows us to compare the overarching trends of severity under each methods. This further motivates the use of complete results only in our analysis, since it is vital for SNP cluster data to be available for comparison. However, there are over 800 different SNP cluster groups for fully observed data points alone, so we will introduce additional methodology to account for this when contrasting this with our clusters (which will be far less numerous).

Software, Methodology, and Diagnostics

Our analysis is conducted in the R programming environment. Visualizations are rendered using the `tidyverse` and `ggplot2` packages, while tables are rendered using the `table1, knitr, IRdisplay` and `kableExtra` packages. Our clustering analysis will be carried out using the `MASS` and `cluster` packages, with supplemental diagnostics from the `factoextra` and `clValid` packages. Exploratory visuals are likely to keep in line with what was present in our literature review and exploration, mainly consisting of density plots. Visualizations of clusters will include color-coded scatter plots to examine cluster wide behavior in view of continuous numerical variables.

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 $\boldsymbol{\Sigma} = \{\Sigma_j\}_{j=1}^{k}$ 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 that we will implement finds the set of clusters $\boldsymbol{\Sigma}^*$ that satisfies:

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

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 $\boldsymbol{\Sigma^*}$, 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:

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

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

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

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. Elementary statistical inference may also be applied for the purpose of supporting the overarching trends that might arise between our clusters. Specifically, analysis of variance applied to mean minimum distance or contig count might be used, and categorical χ^2 tests of association might be used to examine differences in serovar prevalence between clusters and differences in source prevalence between clusters.

Application Outline and Expected Results

The collective steps we must first take to properly cluster the data fall in line with what we have done previously, with a few additional nuances added. After creating a subset of the data the consists of only our numerical variables (year, contigs, N50, minimum distance, and numerically coded source), we employ `kmeans()` to this numerical subset with k^* clusters, where k^* is deemed to be the optimal number. We obtain this value using `fviz_nbclust()` on a sample of 10,000 randomly selected observations, which will yield a plot of the average silhouette coefficient against the number of clusters used (where we will select the number of clusters that yields the highest silhouette coefficient value as explained previously). The output will essentially be a new variable, a cluster assignment taking on a values in the set $\{1, \ldots, k^*\}$.

After this, we will create a variety of visuals depicting any possible trends arising between clusters. To begin, each cluster will be assigned a given color. Then, scatter plots for minimum distance against contig count, minimum distance against N50, and contig count against N50 will be created with the appropriate color coordination. Density estimation plots will also likely be provided for these variables, but more importantly, densities will be examined for each cluster by year and source (since these are finite, countable variables). This may come in potentially different forms (we may decide to display cluster densities for the years 2010, 2015, and 2020, for example, to get an idea of how data in each cluster behaves in different years while still preserving brevity).

Then, we will focus our efforts on tabulating the trends and prevalences of serovars by cluster, investigating the possible differences in the prevalence across different clusters by performing a two sample z test of proportions on the mean number of "dangerous" strains for each cluster. This kind of approach, while elementary, would allow us to present an easily understandable idea of how each cluster differs from the others in terms of the public health threat its isolates posed. As for how we will deem a strain to be "dangerous", an approach we are leaning towards is simply looking at the proportion of isolates with serovar 4b, or simply the number of isolates with serovar 4b in a given sample. It is clear from our literature review that serovar 4b isolates are the dominant cause of listeriosis in human beings, so it would make sense to focus on its presence in a given cluster. Similarly, we will also perform an test of proportions for the percentage of given observations from a specific source by cluster. In a similar manner, this would allows us to compare source prevalence by cluster and could possibly point to a relationship between serovar 4b and specific sources.

Finally, we want to compare the behavior of our clusters in view of the pre-existing SNP clusters that are already included in the dataset. Specifically, we want to compare the clusters on the basis of strain danger, which boils down to examining the serovar 4b prevalence in each SNP cluster when compared to our new clusters. An essential question that should be answered to foment the relevance of our approach is: *are these two sets of clusters exactly the same in terms of prevalence of dangerous isolates*? To examine this, we can iteratively employ some more elementary tests. For each cluster, if SNP clustering is truly different from our clustering, we would expect the probability of a given observation in one of our clusters belonging to a certain SNP cluster would be the same as the probability of the observation belonging to any other SNP cluster. Hence, we can use a χ^2 test for association to determine if there is any degree of overlap between the two clustering methods. We can also use less rigorous methods of examining overlap, like creating bar plots for each cluster showing prominent SNP cluster assignments for that given cluster.

We may or may not experiment with the implementation of network modeling using the `linkcomm` package in R in order to further compare our results with the SNP clusters and the k-means clusters and possibly cement any trends we might observe. However, this aspect remains ancillary and whether or not we include it is dependent upon performance, functionality, and feasibility. Collectively, we will produce a variety of bivariate visualizations color coded by cluster that will help the reader understand any outstanding patterns in the data for listeria. Additionally, we will produce diagnostic parameters (specifically the average silhouette coefficient, though we may also provide heuristic diagnostics like the "elbow method" to provide a greater variety of diagnostic information) that provide insight into the effectiveness of our model. Finally, all of the summaries statistics (specifically p-values and test statistics) for elementary tests done to compare SNP clusters to our new clusters will be provided in full. Potentially, we might experiment and discuss subclustering (clustering by just *one or a subset* of the numerical variables) with appropriate silhouette plots if this avenue proves informative in our research.

Methodological Justification

The methods we have outlined clearly address the two main questions we seek to answer. First, consider our approach to answering the first question pertaining to any possible association between various numerical covariates (with an emphasis on source type) and serovar 4b prevalence. By clustering by these covariates and performing an test of proportions across the serovar 4b counts, we will be able to tell whether or not there is a significant difference in serovar presence in each cluster. Furthermore, if we decide to supplement our main clustering model with additional subclustering models (where the data is clustered by one covariate at a time), we could offer further support through a test of proportions for differences in serovar prevalence exist when clusters based on individual covariates are formed. At a minimum, we will certainly include a subclustering model for isolation source type alone, as this is the most important covariate in our analysis of serovar prevalence. For the second question, a comparison of distributions for numerical covariates between the SNP clusters and our clusters, along with the elementary statistical tests we have outlined, will clearly illuminate and discrepancies between our clustering model and the pre-existing SNP clusters. We will also likely provide direct assessment of the serovar 4b prevalence in the SNP clusters themselves in order to get a baseline sense of severity trends, but this will only be done for a fixed number of the most prominent SNP clusters. Frankly, the excessive number of SNP clusters that are present in even the complete dataset almost completely undermines their usefulness to any practitioner, since a collection of over 800 clusters many of which have very few entries to begin with (as little as 2) provides little extractable information in the first place. In this regard, we already have answered our second question since our clusters will be far less numerous and easily accessible, allowing for direct examination as needed in a clinical setting. The second question will further be answered by the series of elementary statistical tests that we will perform between the two (or more) clustering models, since there results will allow us to comment on the significance of any differences in serovar prevalence between them.

Initial Analysis and Commentary

We begin with a partial analysis to illustrate the functionality of our methods and ensure that our approach is sound and reasonable. Herein, we cluster for only three of the five variables for this demonstration: N50, contigs, and minimum distance. We begin by determining the optimal number of clusters by using the silhouette coefficients produced for a random sample of 10000 observations from our data. Figure 4 on the next page showcases this result. As we can see, the optimal number of clusters we should use is $k^* = 3$, which will can apply going forward. Following this, we use the k-means algorithm to separate the data into three clusters using these three selected numerical covariates. We choose the number of random starting sets to be 20 as a test case, though we will likely experiment with this parameter in our analysis. Figures 1, 2, and 3 below show the differences in clusters for contigs against distance, N50 against distance, and contigs against N50. The plots show the divisions that exist between the three clusters based on the covariates used. Some of these divisions are quite pronounced, as in Figure 2, while the separations in other cases are more subtle, as in Figure 1. Following from this, we can compute the prevalence of serovar 4b within each of the clusters. About 23.1% of cluster 1 isolates had serovar 4b, 12.0% of cluster 2 isolates had serovar 4b, and 17% of cluster 3 isolates had serovar 4b. Performing a test of proportions for cluster 1 versus cluster 2, cluster 1 versus cluster 3, and cluster 2 versus cluster 3 yielded the following results. The proportion of serovar 4b isolates in cluster 1 differed significantly from the equivalent proportion in cluster 2, and the equivalent proportion in cluster 3 differed significantly from that in cluster 2 as well (with p-values of 0.0363 and 4.97×10^{-6} at the 95% confidence level, respectively). There was no significant difference in serovar 4b proportions between cluster 1 and cluster 3. On the next page is a small sample of the visuals we would produce in our final result, but their contents illustrate the effectiveness of our method in achieving our goals.

Table 1: Source Prevalence by Cluster

Source (Cluster 1)	Prevalence	Source (Cluster 2)	Prevalence	Source (Cluster 3)	Prevalence
Food	0.1987076	Food	0.2148370	Food	0.1553908
Environmental	0.1760905	Environmental	0.3255385	Environmental	0.2126197
Clinical	0.1631664	Clinical	0.2066582	Clinical	0.0770003

Figure 1: Contigs Against Distance

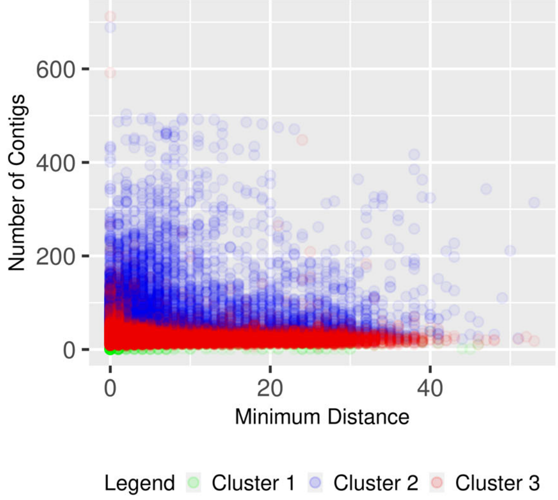


Figure 2: N50 Against Distance

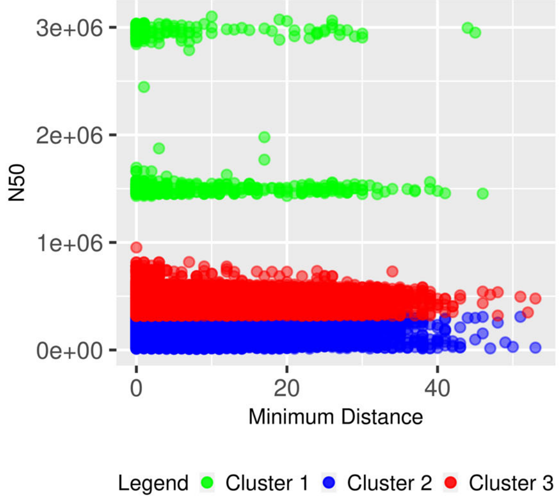


Figure 3: Contigs Against N50

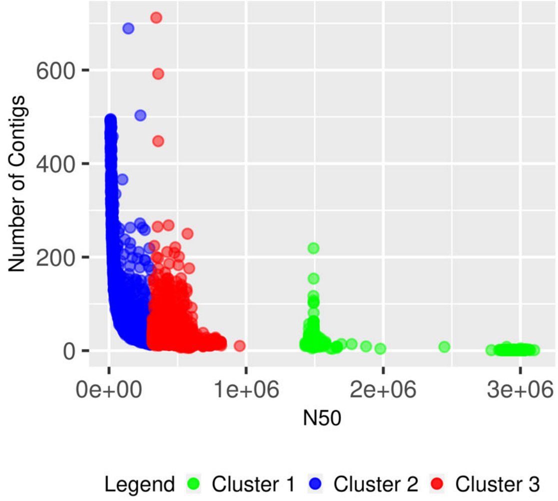
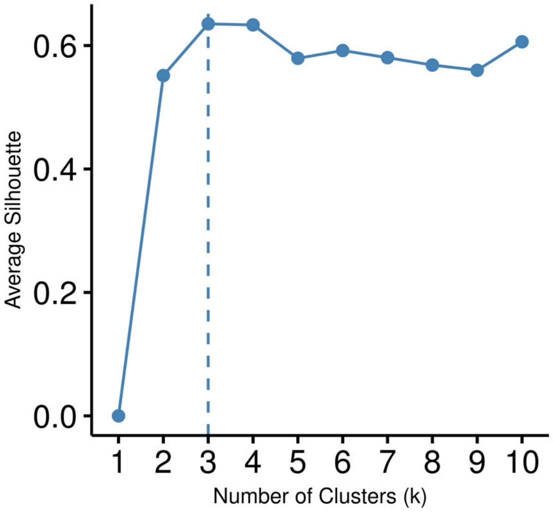


Figure 4: Optimal Number of Clusters



As we see in Table 1, the prevalence of source varies considerably by cluster. In our full analysis, we would test for significant differences in source prevalence by cluster, and specifically, we would compare this result to our subcluster created with solely source data by examining serovar 4b prevalence in our source-only cluster and comparing it to the relationship seen here. Finally, we should briefly touch on a comparison of our clustering model to the model created using SNP information. Grouping by common sources across the three clustering subsets, we want to know if there is a difference in distribution amongst the pre-existing SNP clusters within our own clusters, which would give us an idea of whether or not there is a tendency for at least one SNP cluster the arise with greater likelihood in one of our clusters. A χ^2 test of association yielded a p-value of 2.2×10^{-16}, which would indicate that there is not an equal likelihood of an observation within any one of our clusters being in any one of the SNP clusters. Hence, while we cannot say our clusters are independent of the pre-existing model, they do provide greater context into the trends of severity of a given isolate (from the point of view of considering serovar 4b the most dangerous of all the strains).

While there is much more for us to do here, this surface level analysis proves that our plan to analyze this data is model. As we move along, we will implement these practices with greater detail and refine our results and approaches as needed. We will also experiment with other approaches for comparison to better support our work. A link to our updated repository is given here: https://github.com/NTProvost/PHP-2550-Project