A Comparison of 1918 Flu and COVID-19 Pandemics in Missouri Counties

Stat 9530 Final Project

Aaron Bogan

1 Introduction

It is not hyperbole to say the entire planet has been impacted by the SARS-CoV-2 virus (COVID-19). In early 2020, the virus went from a relatively obscure reference in a news article to a household name in a few short weeks. The human toll of this pandemic will take years to fully realize. In an effort to begin to better understand novel coronavirus, epidemiologists, historians, and anthropologists are investigating the Flu Pandemic of 1918 for similarities. Often referred to as the "Spanish Flu" due to unsubstantiated assumptions concerning the pandemic's origins (Trilla et al. 2008), here it will be referred to as the 1918 Flu Pandemic. The 1918 Flu is estimated to have infected 500 million people (roughly one-third of the global population in the early 20th century), with a death toll of roughly 50 million (Centers for Disease Control 2018). As of this writing, the SARS-CoV-2 virus has infected an estimated 495 million people resulting in 6.1 million deaths worldwide (Johns Hopkins University 2022).

This project seeks to aid in pandemic comparison at the county level in Missouri. Using county-level COVID death counts from various sources compiled by the state of Missouri, and death certificate data from 1918 to 1920, the mortality time series for both pandemics can be observed and compared (see figure 1). Additionally, socioeconomic variables for each county sourced from the 1910 and 2010 census are utilized for comparison. Since county-level divisions are a somewhat arbitrary means of spatial aggregation, more meaningful spatial divisions may assist in pandemic comparisons. Specifically, it was of interest to investigate if counties could be clustered based not only on regional proximity, but also along socioeconomic similarities. Such clustering not only reduces the space dimension of the investigation (providing some simplification) but also has the potential benefit of offering a glimpse into disease progression across time for differing socioeconomic groups. Differences in disease progression observed across the varying regional/socioeconomic clusters provide a window into the pandemics that may help hypothesize potentially important related factors. These insights can be further aided by analyzing the relative importance of the factors in determining county cluster membership.

2 Methods

2.1 Clustering

Ward clustering was performed on county-level socioeconomic variables obtained from the 2010 and 1910 decennial census using the ClustGeo (Chavent et al. 2021) package in R. It was desired to select similar socioeconomic variables for each time period to provide more meaningful comparisons. The 1910 variables include the literacy rate for all voting age males, the average value (in dollars) of all property per farm, proportion of males, number of physicians per 100,000 people, the ratio of 20 to 44 year-olds to those older than 65, and proportion white. The same variables were used for the 2010 analysis plus the proportion of Hispanics. Additionally, the "literacy rate" for 2010 is the proportion of persons age 25 or higher with at least a high school diploma. Using a Ward dendrogram of all possible cluster divisions based on socioeconomic

features as well as input from project collaborators, a four-cluster scheme was selected for both time periods. The four clusters of counties based solely on the feature space fails to account for regional/spatial similarities associated with county proximity (see figure 2). To account for spatial similarities, county centroids were used to introduce a contiguity constraint. The constraint provides a means of selecting the amount of influence physical proximity has on the socioeconomic homogeneity of each cluster.

Selecting a reasonable weighting ("alpha") of the physical space dimension relative to the feature space can be accomplished by choosing a value that minimizes the reduction in feature similarity while maximizing the influence of spatial proximity. This concept is demonstrated visually in figure 3. For the 1910 socioeconomic setting, an alpha value of 0.15 was selected. For the 2010 socioeconomic clustering, alpha was 0.22.

Applying the addiitonal proximity constraint, a much more spatially contiguous grouping of counties is achieved. Based on recommendations from anthropology experts on the research team, the four 1910 clusters were identified as "North," "Central," "Southwest," and "Southeast." The 2010 contiguity constrained clusters were identified as "North," "South," "Urban," and "Bootheel." (figure 4)

2.2 Variable/Feature Importance

To better understand the influence of particular socioeconomic variables, a random forest algorithm (Liaw and Wiener 2002) was used on the clustered counties to evaluate relative importance. For the 1910 clusters, average farm value, proportion white, and literacy were highly influential predictors of county cluster assignment. For 2010, the same three variables were estimated to be the most important, but proportion Hispanic and the age ratio variable also had a reasonable amount of influence (see figure 5).

2.3 Mortality Measures

For the 1918 Flu Pandemic, researchers examined Missouri death certificates dated between January 1918 and December 1920. Daily death counts attributed to the pandemic were sorted by county. These counts were standardized for county population, and the cumulative daily deaths per 100,000 people was used as a measure of mortality.

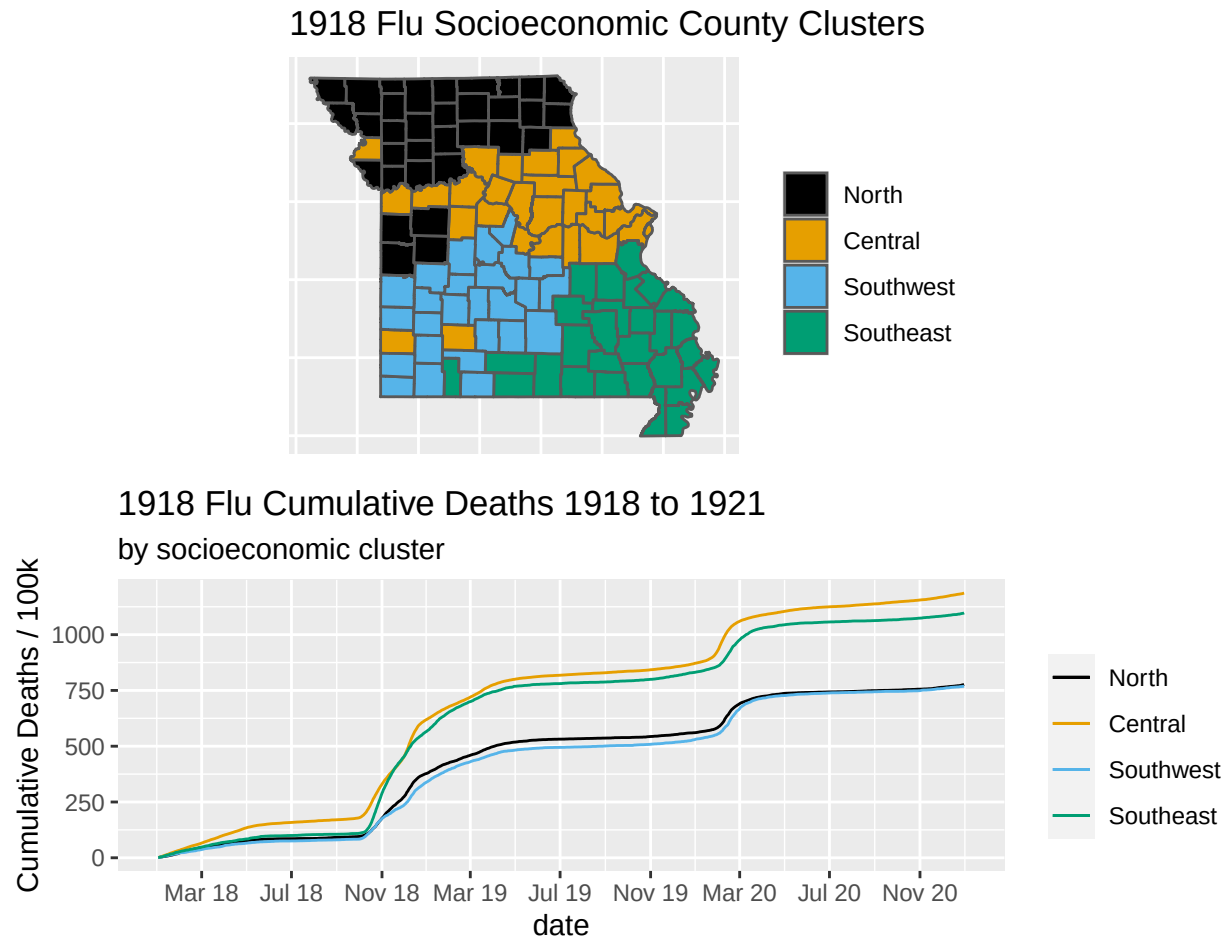
For COVID, the state of Missouri compiled daily counts of deaths attributed to the SARS-CoV-2 virus from various sources across the state. Death data from January 1st, 2020 to December 9th, 2021 was used for this analysis. Similar to the 1918 Flu data, cumulative daily counts were sorted by county and standardized for population. Figure 1 provides an illustration of the cumulative time series for both pandemics.

3 Findings

Time series visualizations of both pandemics' mortality were created to evaluate potential trends/similarities/differences between the socioeconomic clusters.

3.1 The 1918 Flu Pandemic

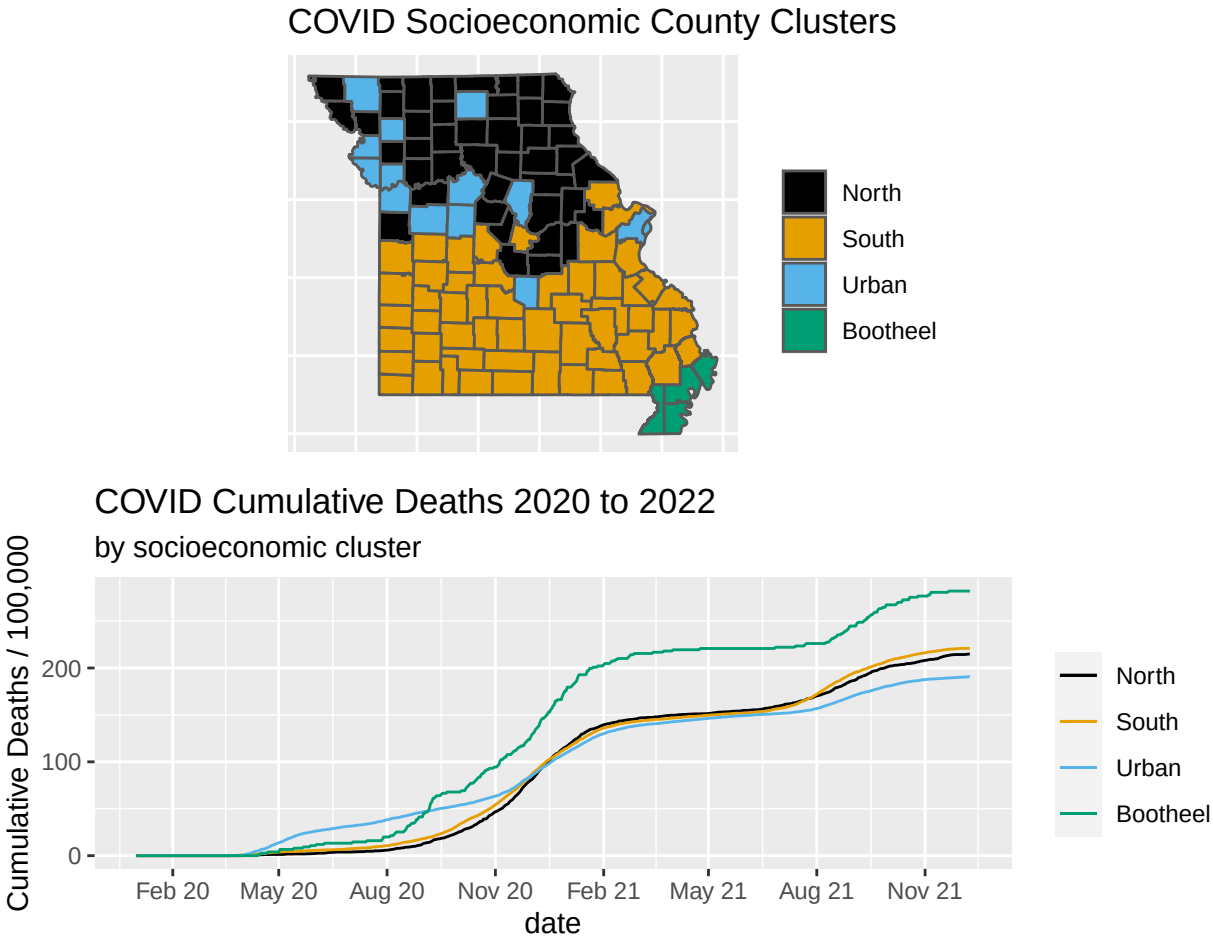
Cumulative death counts / 100,000 for the central and southeast clustered counties suggest a similar time series progression. The north and southwest clusters also have similar trajectories. The central/southeast appears to have suffered increased mortality from roughly November 1918 to April 1919 when compared to the north/southwest clusters. This suggests a more severe flu season in these counties when compared to those socioeconomically grouped in the north/southwest clusters.



A comparison of these combined clusters' important socioeconomic features (average farm value, proportion white, and literacy) shows statistically significant differences in mean literacy and percent white. Simple, one-sided T-tests show the central/southeast clustered counties have significantly lower mean literacy rates and proportion of whites (pvalues ≈ 0 for both). Average farm values for these combined regions are not significantly different at the .05 level (pvalue $= 0.06$). Figure 6 shows boxplot comparisons of these variables.

3.2 COVID

A similar analysis of the COVID time series shows visibly increased mortality in the "Bootheel" region (southeast corner of the state) between August 2020 and roughly March 2021. Additionally, urban counties have visually higher mortality from the pandemic's start in March of 2020 to approximately November of the same year. At that point, the north, south, and urban clusters demonstrate fairly similar cumulative patterns over time. This suggests an earlier initial surge in mortality for urban counties, but an overall similar pandemic experience for all of the non-Bootheel counties.



As with the 1918 Flu, mean comparisons were conducted using simple, one-sided T-tests on the combined clusters of Missouri counties. Based on visually similar cumulative pandemic time series, the north, south, and urban clusters were combined and compared to the far southeast counties in the Bootheel cluster.

The Bootheel-clustered counties have significantly lower literacy levels (pvalue ≈ 0) and percent white (pvalue $= 0.01$) than the rest of the state. The average farm value for Bootheel counties is significantly higher (pvalue $= 0.001$). Proportion of Hispanics and age ratios are not significantly different. It should be noted that the Bootheel cluster includes only a small number of counties (4 out of 115 total Missouri counties) which lends less credibility to these mean comparisons due to possible undetected issues with distributional assumptions required for the statistical tests. However, as the small sample does not appear to be heavily skewed (see figure 7), there is still reasonable evidence to support the objective validity of these conclusions.

4 Discussion

For both pandemics, the southeast region of Missouri seems to have been more negatively impacted than the rest of the state. In the 1918 pandemic, this includes a relatively large number of counties in this spatial region. For COVID, only four "Bootheel" counties in the extreme southeast corner of the state are included. Based on county-level data from both the 1910 and 2010 census, these clustered counties have statistically significantly lower literacy rates as well as percent whites for both pandemics. These results are interesting in that while separated by a century, counties that seem to have experienced higher mortality for both 1918 Flu and COVID demonstrate similar regional/socioeconomic differences.

4.1 Recommendations

Additional clustering schemes should be attempted in order to lend support to the somewhat arbitrary choice of four clusters. K-means, principle components, or perhaps the multiple Bayesian elastic net (MBEN) which uses a Dirichlet process to determine the number of clusters based on prior distributional assumptions and the data may be useful here (Yang 2010).

In an attempt to cluster counties based on the response data (cumulative deaths per 100,000), the method employed in this project did not produce useful results. Counties seemed to cluster somewhat randomly based on their overall pandemic mortality. However, other methods may well be able to detect relationships among the counties based on their individual pandemic time series (see Schweinberger et al. (2017) for an example of detecting relationships within high-dimensional spatio-temporally correlated data).

More sophisticated means of comparing cluster time series would lend additional support to the combined cluster comparisons. Similarly, more complex modeling is necessary to better deal with the space-time correlations inherent in these data (Wikle et al. 2019).

It is beyond the scope of this particular project to make conjecture(s) concerning possible causal relationship(s) between the pandemics and socioeconomic differences. An obvious next step would be to consult with subject matter experts concerning broader interpretation of the results.

5 Conclusion

Researchers are interested in comparing the 1918 Flu and COVID pandemics to provide a better understanding of disease spread. Identifying important factors that influence transmission is key to reducing the cost of future pandemics. Using socioeconomic variables common to both time periods and readily available from the decennial census (1910 and 2010), Missouri counties were grouped together using a Ward clustering algorithm with a weighted contiguity constraint. A four-cluster grouping of counties for both eras was selected based on Ward dendrograms and subject matter expert input. Visual time series analysis of mortality demonstrated noticeable differences between pandemic progression for some of the clusters. More specifically, southeasetern Missouri counties tended to be more negatively impacted by both pandemics. Utilizing basic T-tests, these counties were shown to have lower mean literacy rates and percentage of whites for both time periods. More sophisticated means of modeling the space-time dependencies in the mortality data are necessary to reveal additional, useful relationships as well as potentially important disease transmission factors.

6 Figures

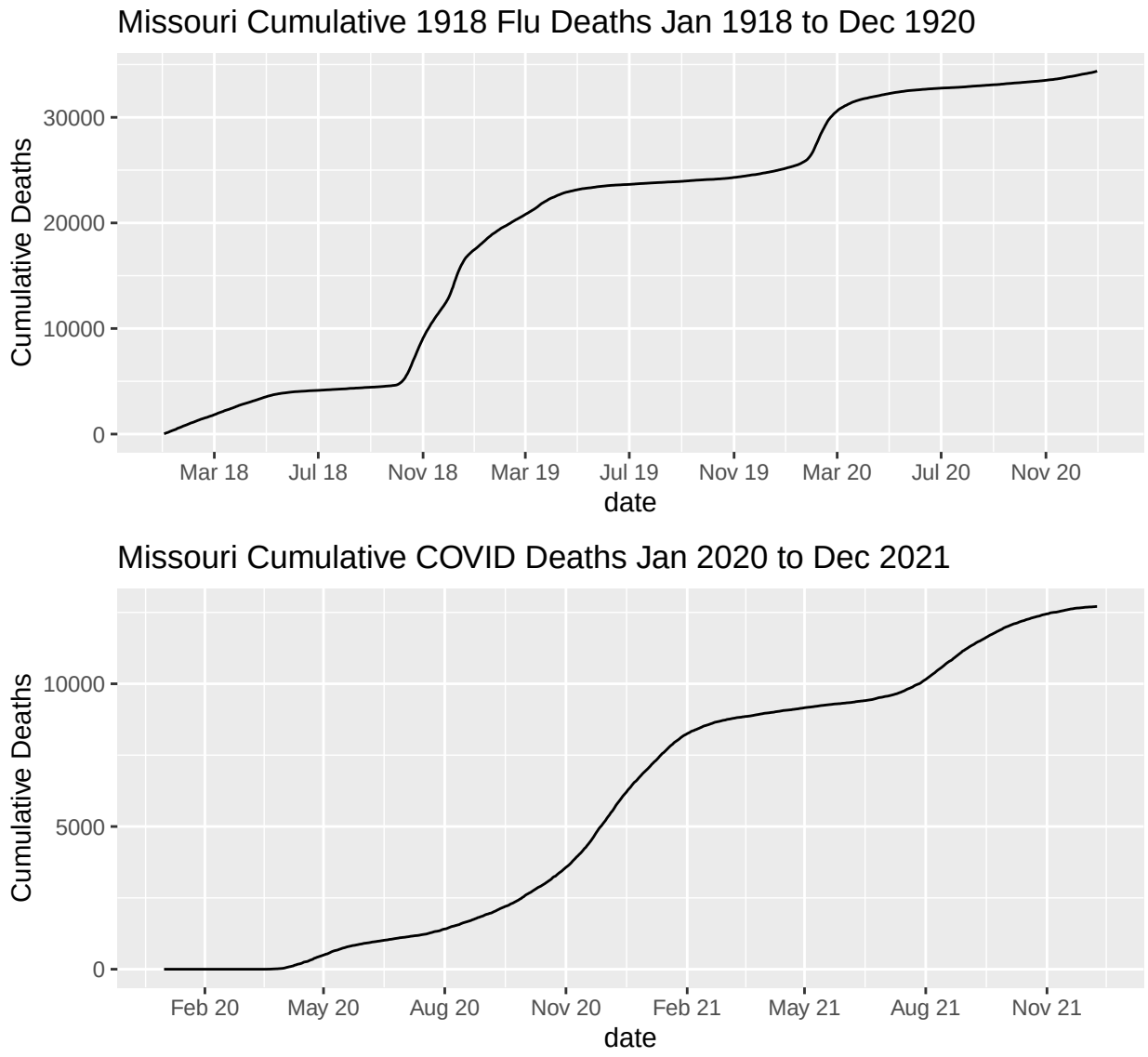


Figure 1: Statewide 1918 Flu and COVID Deaths.

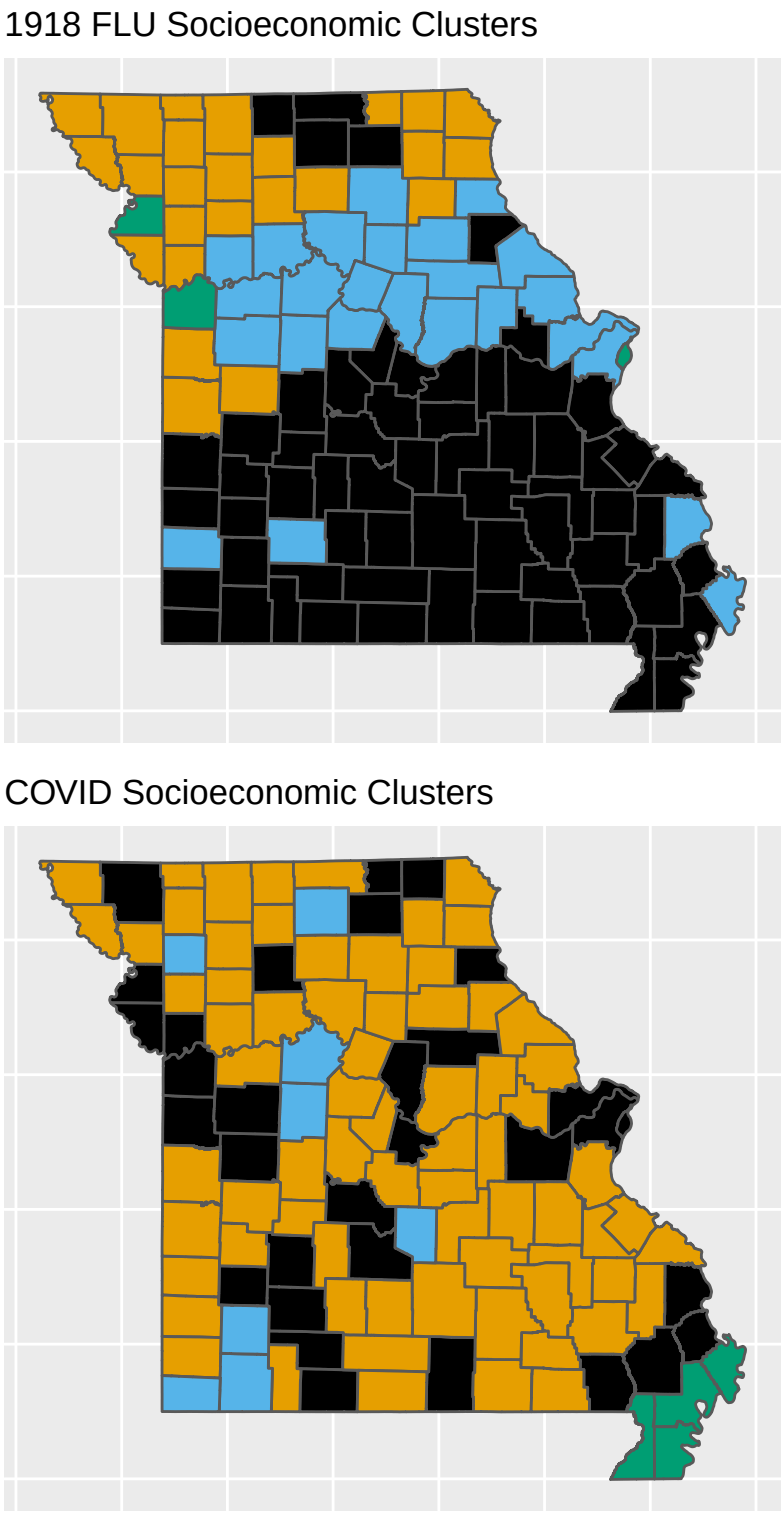


Figure 2: County clusters based on socioeconomic similarity alone (no contiguity constraint)

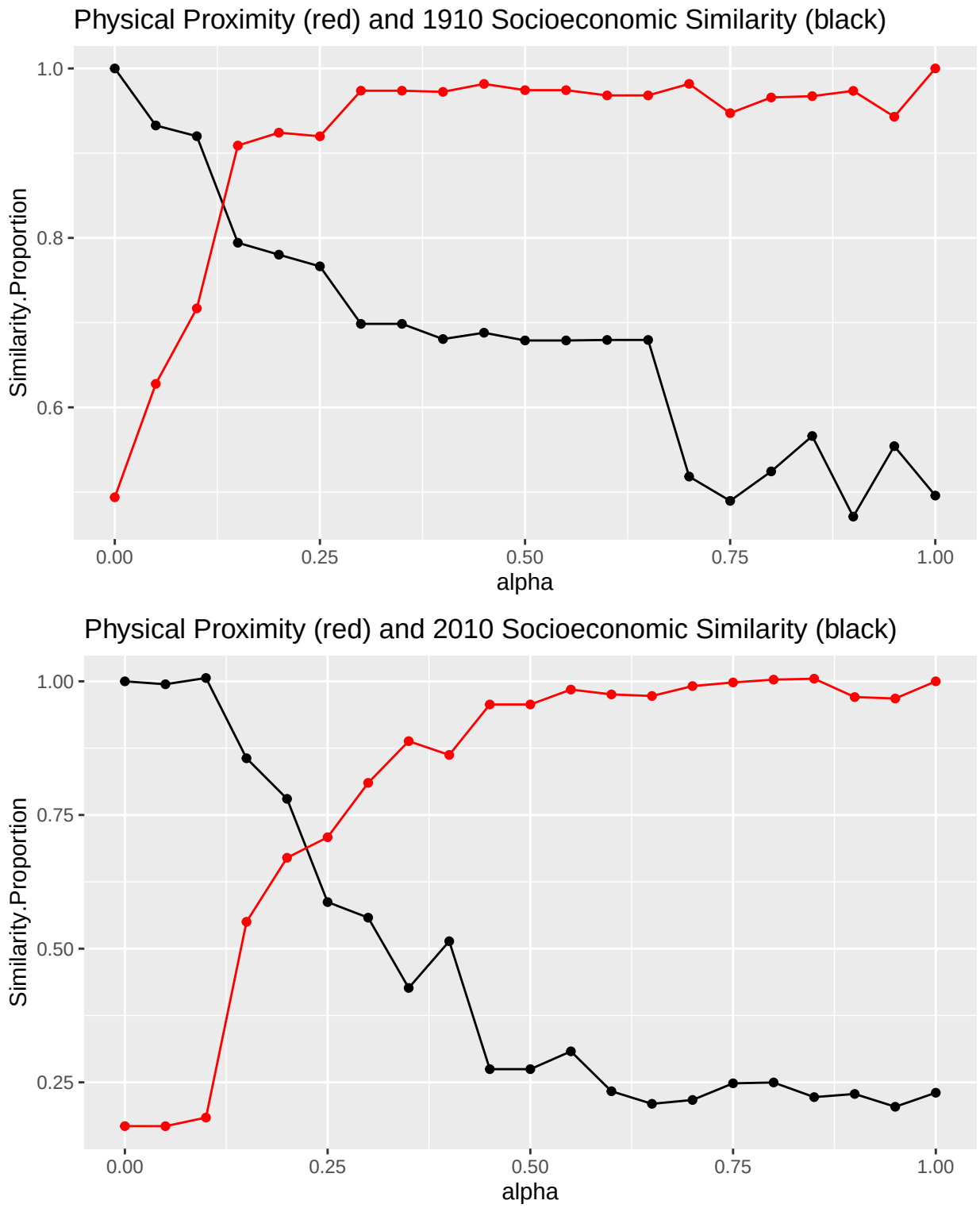


Figure 3: County similarity based on physical proximity (in red) and socioeconomic variables (in black) for various alpha weights

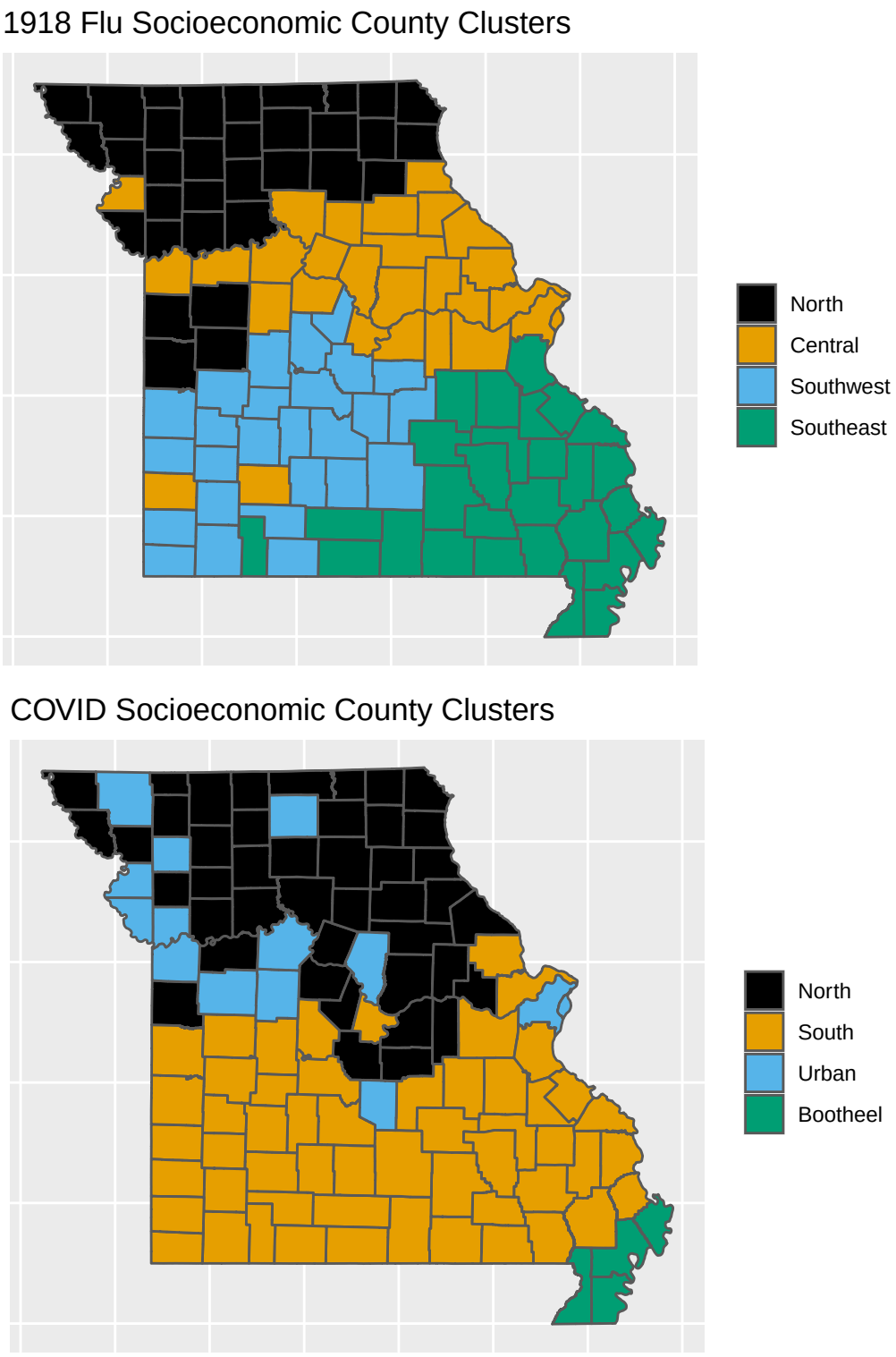


Figure 4: Contiguity constrained county clusters based on socioeconomic similarity

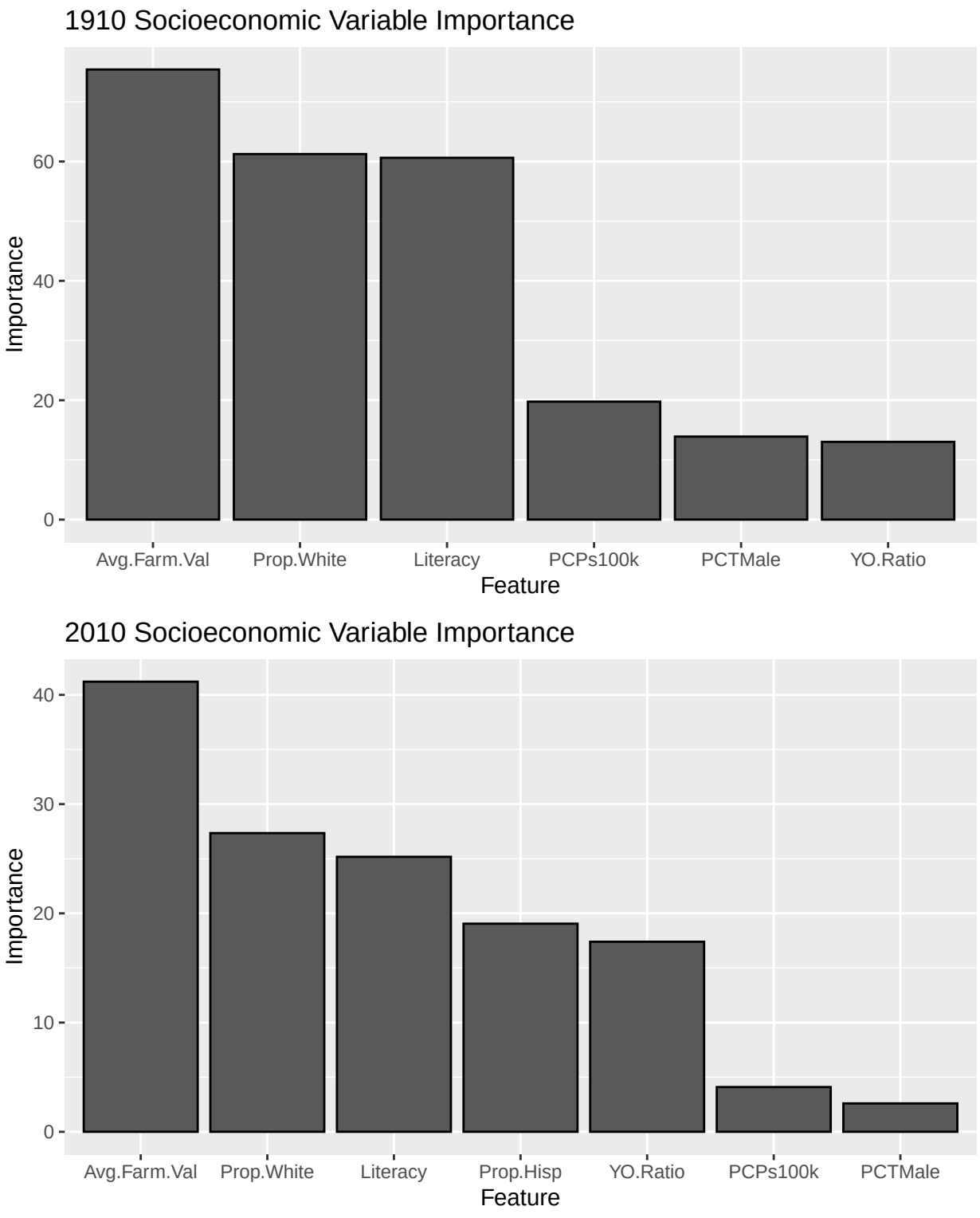


Figure 5: Variable importance based on mean accuracy decrease for 1910 and 2010 socioeconomic clusters

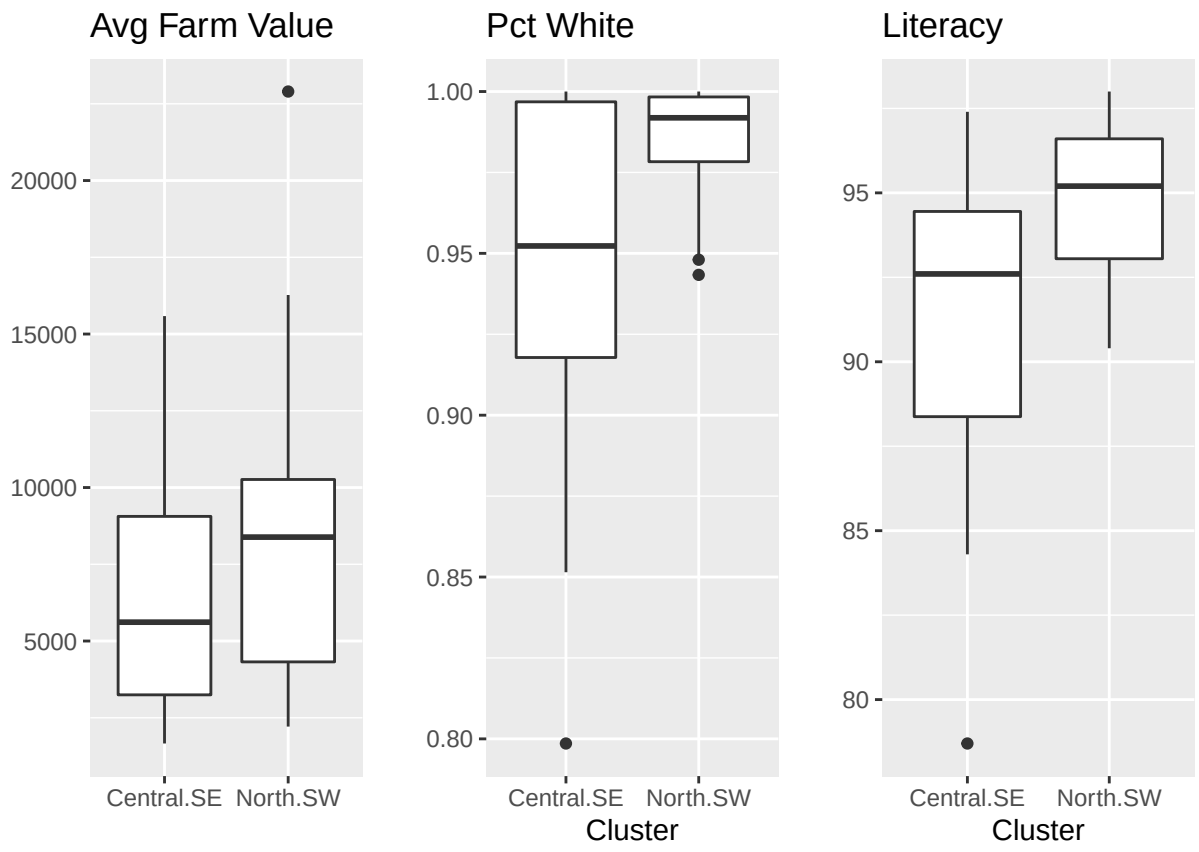


Figure 6: 1910 Socioeconomic variable comparisons by mortality similarity combined clusters

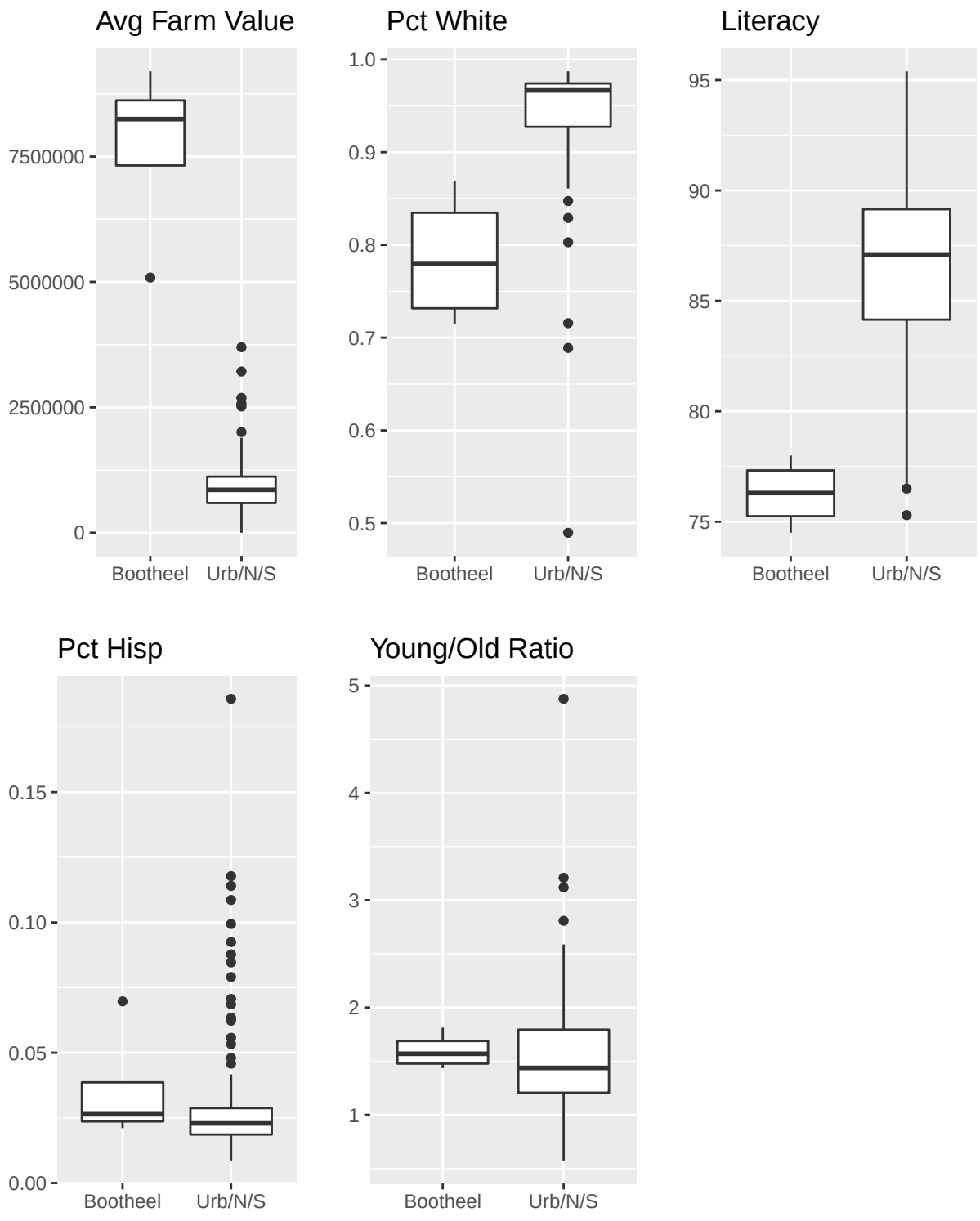


Figure 7: 2010 Socioeconomic variable comparisons by mortality similarity combined clusters

References

Centers for Disease Control (2018), "History of 1918 flu pandemic," Available athttps://www.cdc.gov/flu/pandemic-resources/1918-commemoration/1918-pandemic-history.htm.

Chavent, M., Kuentz, V., Labenne, A., and Saracco, J. (2021), *ClustGeo: Hierarchical clustering with spatial constraints*.

Johns Hopkins University (2022), "COVID-19 dashboard," Available athttps://www.arcgis.com/apps/dashboards/bda7594740fd40299423467b48e9ecf6.

Liaw, A., and Wiener, M. (2002), "Classification and regression by randomForest," *R News*, 2, 18–22.

Schweinberger, M., Babkin, S., and Ensor, K. B. (2017), "High-dimensional multivariate time series with additional structure," *Journal of Computational and Graphical Statistics*, Taylor & Francis, 26, 610–622. https://doi.org/10.1080/10618600.2016.1265528.

Trilla, A., Trilla, G., and Daer, C. (2008), "The 1918 'Spanish Flu' in Spain," *Clinical Infectious Diseases*, 47, 668–673. https://doi.org/10.1086/590567.

Wikle, C. K., Zammit-Mangion, A., and Cressie, N. (2019), *Spatio-temporal statistics with r*, Chapman; Hall/CRC.

Yang, H. (2010), "Nonparametric bayes models for high-dimensional and sparse data," PhD thesis, Duke University.