# Clustering County-Level Mortality Curves in the United States

Colton Syndergaard[a], Brian Hartman*[a], Robert Richardson[a], and Chris Groendyke[b]

[a]Department of Statistics, Brigham Young University, Provo, UT, USA
[b]Department of Mathematics, Robert Morris University, Moon Township, PA, USA

August 19, 2024

**Abstract**

Regions with similar characteristics often exhibit comparable mortality patterns. Recent papers show that similarity can be expressed spatially (areas closer together are more likely to have similar experience). That similarity can also be expressed in other dimensions like rurality, political affiliation, health, and socioeconomic status. In this study, we explore mortality similarity independent of spatial location in the United States. We construct county-level mortality curves for both males and females in the contiguous United States using cubic splines and derive regression coefficients from these splines. These coefficients form the basis for clustering counties with similar mortality patterns, revealing that three clusters are generally optimal. Clustering is performed for each individual year as well as for all years from 2000 to 2021 combined. Our findings indicate that while the cluster-level mortality curves exhibit broadly similar shapes, significant differences emerge particularly at ages 5 – 35. We use multinomial logistic regression and a random forest to analyze the differences between these clusters based on several covariates collected from the constituent counties, such as population density, race, marriage level, household size, unemployment rate, and education. The results suggest that the clusters have significant differences with respect to these covariates and that the clusters largely reflect an urban-rural divide. Additionally, we examine how the compositions of these clusters change over time. Finally, we compare the performance of the cluster-level mortality curves to those formed at the state level at predicting future mortality and find that the cluster-based mortality curves are generally superior to the state-level models in this regard. Practicing actuaries can use these clusters to build mortality models at the county level, enabling better predicting, pricing, and risk management.

## 1 Introduction

Accurate mortality inference, prediction, and forecasting are essential for many stakeholders. Community leaders are interested in population aging to understand demand for social services and the impact of public health policies (Kindig and Cheng, 2013; Masters et al., 2015). Private organizations and actuarial teams allocate resources such as pensions, life insurance rates, and predicted payouts (Brown and Orszag, 2006; Wilmoth and Horiuchi, 1999).

In this analysis, we focus on mortality rates in the United States. U.S. mortality rates are the focus of many recent papers (Ho and Hendi, 2018; Woolf et al., 2018; Case and Deaton, 2017). But modeling the mortality rates of the entire United States is an oversimplification. The people of the United States are not a monolith. They vary on many dimensions (Xu et al., 2020) including spatially (climate, elevation, humidity), culturally (politics, religion, community), economically (income, wealth, poverty), and healthcare (access, quality, price). Building a model without allowing for flexibility on these dimensions can lead to poor model fit and potentially poor decisions.

Many of the current papers modeling county-specific mortality do not use the spatial relationships to borrow strength between counties (Dwyer-Lindgren et al., 2016; Monnat, 2018; Currie and Schwandt, 2016; Clark and Williams, 2016). Other recent papers use the spatial relationships between counties to improve the mortality predictions (Gibbs et al., 2020; Yang et al., 2015). But some counties may be

*Corresponding author - email: hartman@stat.byu.edu

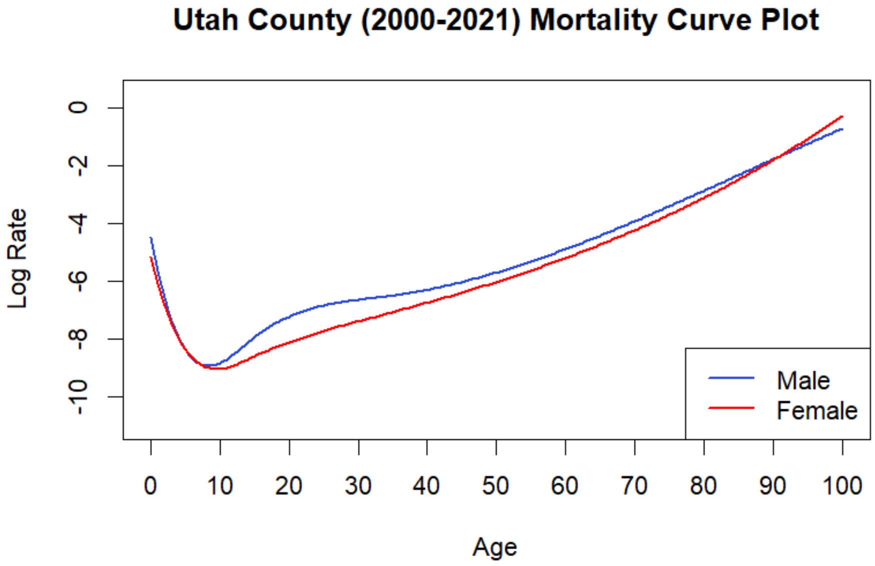


Figure 1: Example County Level Mortality Curves, Utah County

very similar even if they are not especially proximate. For example, a relatively rural county in Western Washington may be more similar to a rural county in Nebraska than it is to a neighboring county which is mostly a suburb of Seattle (and perhaps less rural). This paper groups counties more flexibly, based on their mortality experience, relying less on their spatial location. Singh and Siahpush (2014) references urban and rural disparities in mortality, while Chetty et al. (2016) discusses the impact of income on mortality, suggesting that mortality grouping may rely more on characteristics of a region rather than a borrowing strength spatially.

Initially we will investigate the shape of the curves as a function of age alone. Due to data constraints, we estimate this curve by aggregating the mortality information of ages which are close together; we then use the methods described in Section 2.2 to construct a continuous mortality curve as a function of age. This process produces curves that look like the example in Figure 1, which has the male and female mortality curves from data aggregated over the period from 2000-2021.

For many counties, however, this aggregation is insufficient to gain insight into the true mortality rate. This issue is particularly prevalent in counties that have very small populations, which may have too few individuals to accurately assess mortality risk. There are some common practices that have been suggested in the past, such as aggregating the data for a single county over many years, or by using state level data (Kleinman, 1977). However these methods can fail to meaningfully account for a variety of factors; aggregating data over multiple years can cause abnormal events (such as a global pandemic) to give a skewed perception of true typical mortality, and using state level data does not capture the difference in varying parts of the state, such as the rural-urban divide.

This is one issue that this paper seeks to address. Specifically, we seek to define some method of determining mortality curve similarity across counties in the continental United States, finding natural groupings or clusters of counties. We can then aggregate the mortality information within these clusters to properly assess the rate of mortality in these "similar" counties. A secondary objective is to then interpret these clusters, to try to understand how these clusters can be characterized, as well as how the composition of the clusters changes over the years.

One apparent difficulty with this goal is the functional nature of the data that we are dealing with. Although data is collected discretely, aging is a continuous process; as such, the mortality curve (a function of age) is also continuous. This can present difficulties for traditional statistical methods. Several methods have been developed for representing and clustering such data. For example, Tarpey and Kinateder (2003) discussed the capabilities of $k$-means clustering when representing each curve in the finite dimensional subspace spanned by the eigenfunctions of its covariance kernel. More recently Bouveyron and Jacques (2011) explored model-based clustering methods for univariate curves, specifically time series, noting that finite dimensional curve representations often are either discretized forms of the continuous process (measurements at intervals), or represented by spline basis functions. We use the cubic spline representation of the curves which are univariate functions of age, and perform a linear regression on them as this is a straightforward way of representing these curves (Faraway, 1997), and then use the resulting coefficients to perform the clustering.

The remainder of this paper is organized as follows: Section 2 describes the data and methodology used in this study; Section 3 gives the results of the clustering and compares the mortality predictions of cluster-based models to those of state- and national-level models; Section 4 uses regression and random forest analyses to lend interpretations to the clusters; and Section 5 offers some concluding remarks.

# 2 Methodology

## 2.1 Data

The data used to create the mortality curves was collected from the Division of Vital Statistics of the National Center for Health Statistics (2023), which itself is a subdivision of the Centers for Disease Control and Prevention; these data contain demographic and mortality information for every individual who died in the United States between 2000 and 2021. The total population of each demographic group was obtained via the estimates made by the United States Census Bureau. The census data aggregates ages into 18 groups of five years, with group one being ages 0-4, group two being ages 5-9, continuing until the age of 85, over which all are assigned to the eighteenth group (Gibbs et al., 2020). As a result, our data is also binned and to increase interpretability the groups 1-18 were taken as the mean age in the group; for the first 17 groups this was simply the median age (2 for group one, 7 for group 2, etc.), until the group 18 which we assigned to be 90.

There are other issues with the data. Several counties were very small and had population estimates of zero at some point in time for different age groups. Others had more reported deaths in an age group than people estimated to be in that age group. Finally, certain counties changed their designation or underwent boundary changes over the course of the 20 years that data were gathered. To solve these issues, we combined counties where the issues were present with their respective most populous adjacent counties; details are given in Table 8 and can be seen on the map in Figure 17, both of which can be found in Appendix A.

We also have collected various covariates for each county relating to the county's demographic characteristics. These covariates, which include proxies for race, household size, education level, unemployment rate, marriage rate, land area, and population density, are used to help interpret the clusters of counties that result from our analysis. These variables are described in Section 4.1.

## 2.2 Curve Definition and Estimation

Almost 200 years ago Benjamin Gompertz noted that mortality increases exponentially, particularly after age thirty (Gompertz, 1825). Traditional literature posits that after age 85 there is a deceleration in mortality increases, however more recent studies disagree with this assertion, claiming that mortality continues to increase at an exponential rate (Gavrilov and Gavrilova, 2011). Because of this, a mortality curve can be defined as the natural logarithm of the mortality rate for a given age, so that after age 30 the curve is a relatively linear function of age.

However, for forecasting and understanding mortality, this fact alone is insufficient. Building off of Gompertz's work, Makeham proposed that mortality is best modelled by the sum of the previously described age component as well as an additional age-independent component (Makeham, 1860). This secondary component (which itself is the result of many disparate parts such as accidents, disease, war, etc.) is of particular interest to many mortality researchers, such as Marmot et al. (1981), where the effects of excessive alcohol on mortality were studied.

Mortality curves can be represented by various functions that model continuous processes, such as polynomials and splines. In our case cubic splines are preferable, as they allow us to model differing behaviors over various ages, accounting for relatively high infant and young adult mortality. For example, consider the mortality curve in Figure 1, but now with the two interior knots shown in Figure 2. Before the first interior knot we see a steep decline in the likelihood of death after infancy. Between the interior knots, covering individuals from puberty to young adulthood, the curve increases more steeply than expected from age alone. After this, the rate of increase slows dramatically, settling into relative linearity after the upper knot. This increased mortality in youth is particularly pronounced in male

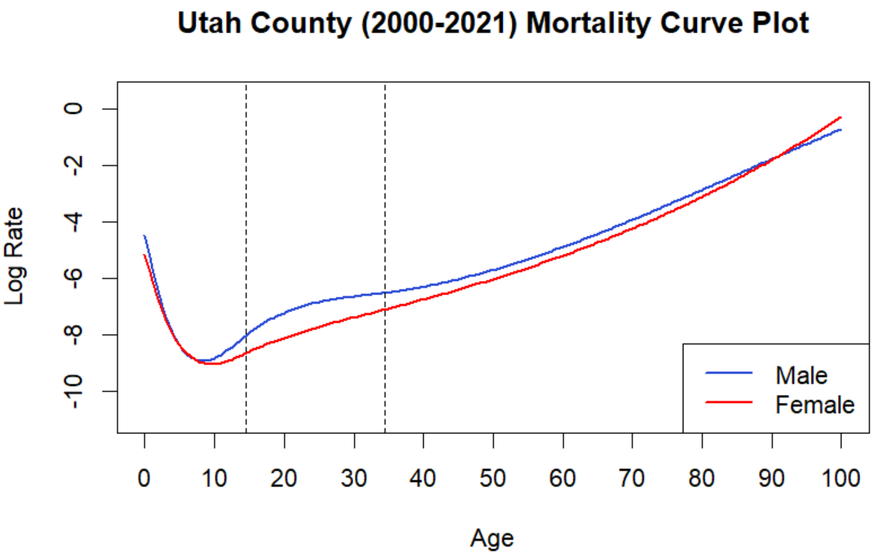


Figure 2: Utah County Mortality Curve Splines with Knots Defined

curves. Because this hill-like structure can at least partially be attributed to an increase in risk-taking behaviors, this phenomenon is colloquially known as the accident or young adult mortality hump; for more discussion of causes of this hump see Remund et al. (2018). As a result of our mortality curves having distinct forms and features over these three age ranges, we have chosen the interior knots for the splines at ages 15 and 35.

To avoid undefined values in counties with no recorded deaths (or individuals) at certain ages, we introduce a small correction to the traditional form, as shown in Equation (1).

$$M(a) = log\left(\frac{(\#\text{ Deaths at age } a) + 1}{(\#\text{ Individuals at age } a) + 1}\right) \tag{1}$$

Letting this functional corrected log death rate at age $a$ be $M(a)$, we have that in general the expected value of the mortality curve function for a geographical region may be represented as Equation (2):

$$\mathbb{E}[M(a)] = \sum_{j=0}^{5} \beta_j b_j(a), \tag{2}$$

where the $b_j(a)$ functions ($j = 0, \ldots 5$) are a set of basis functions for $0 \leq a \leq 90$, with $b_0(a) = 1$, an intercept term. These basis functions are represented by the curves in Figure 3; the $\beta_j$ terms are regression coefficients on these basis functions. Further discussion of the basis functions can be found in Beer et al. (2020). This is the model used when we consider the overall (i.e., not varying by year) mortality for a county. Each county's set of regression coefficients $(\beta_0, \ldots, \beta_5) \in \mathbb{R}^6$ form the basis for the overall county clustering; this process is performed separately for the male and female models.

We also use yearly data to fit a hierarchical random coefficients model to model the curves of each county, treating age as a fixed effect and year (2000-2021) treated as a random effect. However, we note that the COVID-19 pandemic significantly impacted mortality across the United States — the pandemic was a year-specific phenomenon but not a random effect. Thus we add a COVID indicator into the model. We treat the presence of COVID as a fixed effect on the model, and we explore its consequence with each of the spline basis functions via an interaction effect, with the main effect being absorbed into the intercept.

This results in the model in Equation (3), where the total regression coefficient on the $j^{th}$ basis function, $b_j(a)$ is the sum of $\beta_j$, the fixed effect that age has on county mortality over the $j^{th}$ basis function, and $\alpha_{jy}$, the deviation that the $y^{th}$ year has from that main effect, and the fixed COVID effect, $\gamma_j \mathcal{I}(y)$, where $\mathcal{I}(y)$ is the indicator function defined in Equation (4).

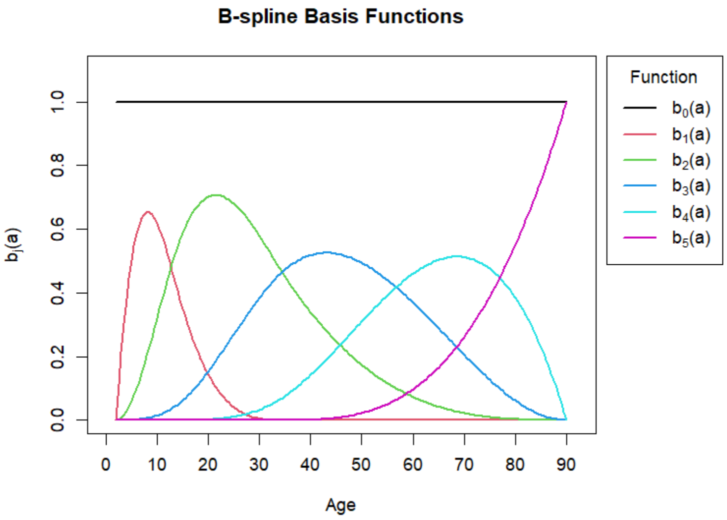


Figure 3: Basis Functions for County Splines

$$\begin{aligned}\mathbb{E}[M(a,y)] &= \sum_{j=0}^{5}\gamma_j\mathcal{I}(y)b_j(a) + \sum_{j=0}^{5}(\beta_j+\alpha_{jy})b_j(a) \\ &= \sum_{j=0}^{5}(\beta_j+\alpha_{jy}+\gamma_j\mathcal{I}(y))b_j(a)\end{aligned} \tag{3}$$

$$\mathcal{I}(y) = \begin{cases} 1 & \text{if } y \geq 2020 \\ 0 & \text{if } y < 2020 \end{cases} \tag{4}$$

These coefficients are then used as coordinates in $\mathbb{R}^6$, $(\beta_0+\alpha_{0y}+\gamma_0\mathcal{I}(y), \beta_1+\alpha_{1y}+\gamma_1\mathcal{I}(y), ..., \beta_5+\alpha_{5y}+\gamma_5\mathcal{I}(y))$, which can now be used to cluster similar counties together for each year separately, finding natural groupings of counties that exhibit, in some sense, similar shapes.

## 2.3 Curve Clustering and Exploration

We primarily employ hierarchical clustering (Ward Jr, 1963) to group counties based on mortality curve coefficients, and the results shown are based on this method. We note, however, that our results do not seem to be sensitive to the clustering method used, as $k$-means clustering (MacQueen et al., 1967) produced very similar results. Using the fixed effects, we determine overall county clusters, while the random effects allow us to cluster each individual year to observe how the clusters evolve over time.

Choosing an optimal number of clusters is challenging due to the absence of “ground truth” groups for county mortality against which to compare. Metrics like the Calinski-Harabasz Index (Caliński and Harabasz, 1974) and the GAP statistic (Tibshirani et al., 2001) help in selecting the number of clusters, but their effectiveness varies. The choice of metric can influence the decision for the number of clusters, and metrics can be inconsistent when clusters are not clear.

Charrad et al. (2014) proposed using an ensemble of relative-performance metrics, creating the `R` (R Core Team, 2024) package `NbClust`. The optimal partition is defined as the one suggested by the plurality of metrics. When determining the number of clusters to use for further analysis, we use the clustering suggested as optimal as explained below.

We then compare aggregated cluster curves to state-level curves in predicting log mortality rates using mean squared error (MSE) to assess accuracy. Each county’s log mortality is predicted using its overall cluster’s curve, and the MSE from these predictions is calculated. This process is repeated annually, comparing the predictions to those using state-level curves for each county, determining if clustering offers predictive advantages over using state curves.

# 3 County-Level Clustering Results

## 3.1 Overall Clustering Results

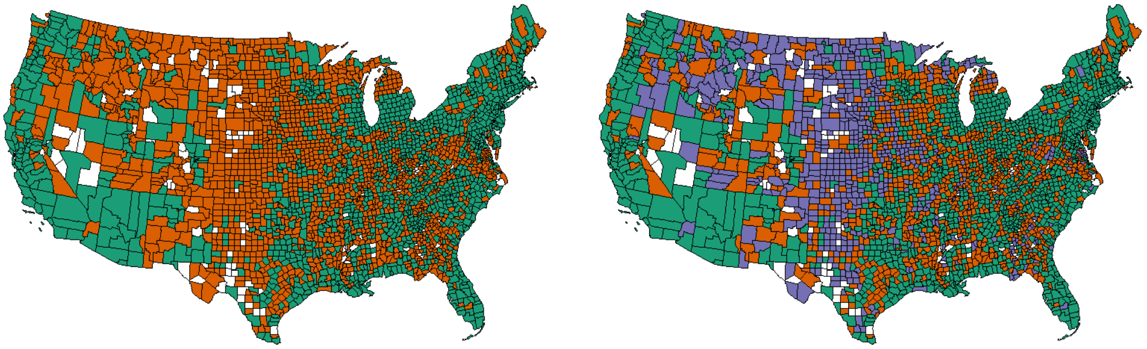
After creating the curves, counties were clustered as described in Section 2. We explored different numbers of clusters, ranging from 2 to 10, for our hierarchical clusterings. The `NbClust` metric results are shown in Table 1. Recall, these metric numbers are simply the number of metrics which were optimized under a given number of clusters. The metrics indicate that either 2 or 3 clusters are optimal for both the male and female clusterings. Note that the spike in metric votes suggesting 10 clusters is due to the upper limit of our search, causing metrics with small complexity penalties, optimized by more granular clusters, to select 10.

Female curves either form 2 poorly separated clusters or 3 close clusters as illustrated in Figure 4. In both the 2 cluster and 3 cluster maps, the green cluster tends to be comprised of more urban counties, while the other cluster(s) are more rural. We can see that when the female map goes from 2 to 3 clusters, the new purple cluster is primarily composed of counties previously in the orange cluster; the counties in this purple cluster are mostly located in the middle and western portions of the U.S.

For the male clusterings, we see similar pictures (see Figure 5), the biggest difference being that the male clusterings feature more counties in the orange clusters.

| Clusters | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10+ |
|---|---|---|---|---|---|---|---|---|---|
| Female | 6 | 7 | 1 | 4 | 1 | 1 | 0 | 0 | 2 |
| Male | 3 | 14 | 1 | 1 | 1 | 0 | 0 | 1 | 1 |

Table 1: Suggested Clusterings



(a) 2 Clusters

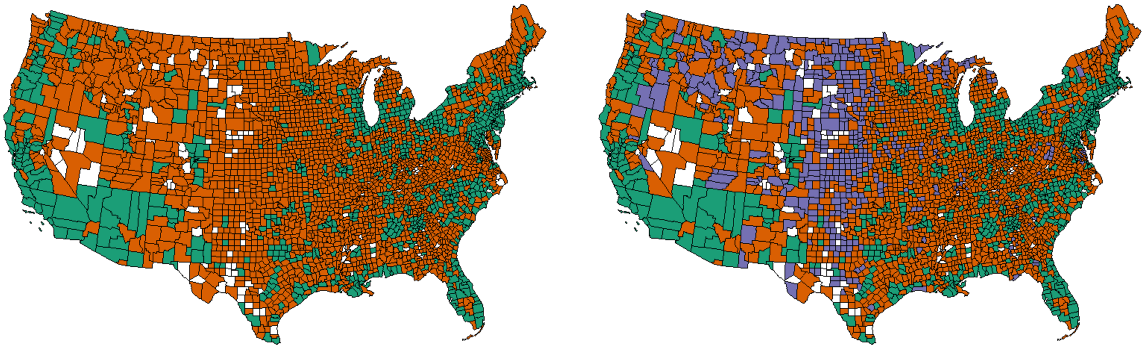
(b) 3 Clusters

Figure 4: Female Clusterings

We choose the clustering method we use for further analysis based on the `NbClust` suggestion of the most consistently optimal partitioning. Table 1 shows that a clustering of 3 clusters is suggested for both men and women, but much more strongly for the female clustering. Thus we will proceed with our further using 3 clusters for both the male and female models; using the same number of clusters for the two models also allows us to more easily compare the male and female clusterings.

We note that the clusterings of counties for the female and male mortality curves resulted in 2220 of the 3007 counties (73.8%) falling in the same overall male and female clusters. Moreover, from Table 2, we can see that the vast majority of the clustering differences are the result of counties which fall in the orange cluster in the male clustering falling one of the other two clusters (mostly the green cluster) in the female model. The reverse is not the case: only 7 of the 1024 counties in the orange cluster in the female model are in another cluster in the male model. We also note that there are no instances of the male model placing a county in the green cluster that the female model puts into the purple cluster, or vice-versa.

Interestingly, across all clusterings, a spatial effect is evident even without imposed spatial structure. Urbanized counties tend to cluster together (green), rural counties are marked as similar (orange), and a consistent band of middle American rural counties (purple) emerges in both three-cluster configurations.



(a) 2 Clusters

(b) 3 Clusters

Figure 5: Male Clusterings

| | | Female Cluster | | | |
|---|---|---|---|---|---|
| | | Green | Orange | Purple | Total |
| Male Cluster | Green | 765 | 1 | 0 | 766 |
| | Orange | 582 | 1017 | 198 | 1797 |
| | Purple | 0 | 6 | 438 | 444 |
| | Total | 1347 | 1024 | 646 | 3007 |

Table 2: Counties by Overall Clusters

The shapes of the curves themselves reveal intriguing qualities. To highlight these, we compare the differences across the curves for the selected clusterings in Figure 6, using the orange cluster as the reference. Female curves show that the purple cluster is quite similar to the orange (both relatively rural) in overall mortality levels, but the green (typically urban) cluster has lower overall mortality. The shapes of the orange and purple mortality curves differ, however, with the orange cluster having better mortality in the younger ages and the purple counties having better mortality in the older ages. The largest differences in the curves are at the younger ages; beyond age 50, the mortality rates converge a bit.

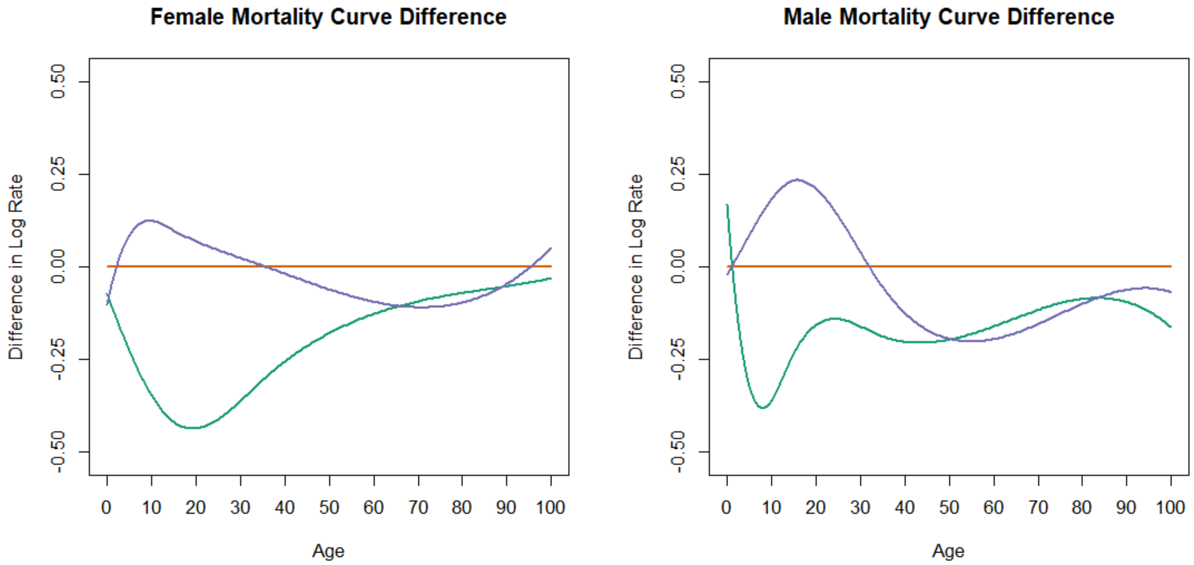
Male curves show a similar pattern, differing significantly until about age 30, then become similar. Specifically, the purple curve exhibits high mortality until age 25, then more closely aligns with the orange curve. The green curve shows lower mortality than the other two throughout, but particularly until age 25.

### 3.2 Yearly Clustering Results

Using these results, we move forward with the clustering with 3 clusters for the year effects. While the precise clusters for each individual year vary, the underlying themes in the overall clusterings were maintained annually, namely a distinct urban-rural divide and a less distinct separation between the two more rural clusters.

The yearly clusterings were relatively stable in consecutive years (with the notable exception due to COVID discussed below), with, on average, just under 70% of counties staying in their same cluster in consecutive years; the results were similar for the male and female clusters, though the female cluster compositions tended to vary slightly less for the female clusters than for the male clusters.

When considering the movement of counties between clusters from year to year, the orange cluster appears to be a "transitory" or "intermediate" cluster between the green and purple clusters for both the male and female clusterings. That is, counties frequently moved between the green and orange clusters, and between the orange and purple clusters, but very rarely directly between the green and purple clusters. A total of 372 counties (out of 3007) remained in the same female cluster for the entirety of the 22 years, while 295 counties remained in the same male cluster for this entire time span. We note that of the counties that stayed in the same cluster for throughout, they were split nearly evenly between



(a) Female Curve Differences

(b) Male Curve Differences

Figure 6: Cluster Mortality Curve Differences

the green and purple clusters, with only a handful staying in the orange cluster for the entirety.

Considering the mortality curves, it is interesting to note that mortality was slowly trending downwards for all ages in each of the clusters from year to year, until 2020 when the COVID-19 pandemic hit, causing a sudden, predictable spike in mortality across all age groups. COVID also influenced the clusters themselves. In Figures 7a and 8a (i.e., 2019, immediately preceding the pandemic) we can see 3 distinct clusters following the trends expected from the overall clustering (Figures 4 and 5). In 2020, however, the makeup of the clusters changed, with a sharp increase in the number of (particularly rural) counties whose mortality experience behaved similarly to the (typically) urban counties (Figures 7b and 8b). We posit that as COVID affected the entire nation, many counties exhibited similar mortality outcomes. By 2021, as lockdowns and other measures were lifted, clusters largely reverted back to what they were before the pandemic, but mortality rates stayed relatively high (Figures 7c, 8c). Thus, at least in its first year, the COVID pandemic appears to not have impacted the mortality in the U.S. in a uniform manner, causing counties which do not typically have similar mortality to cluster together for the year 2020. These effects are very similar for the male and female models.

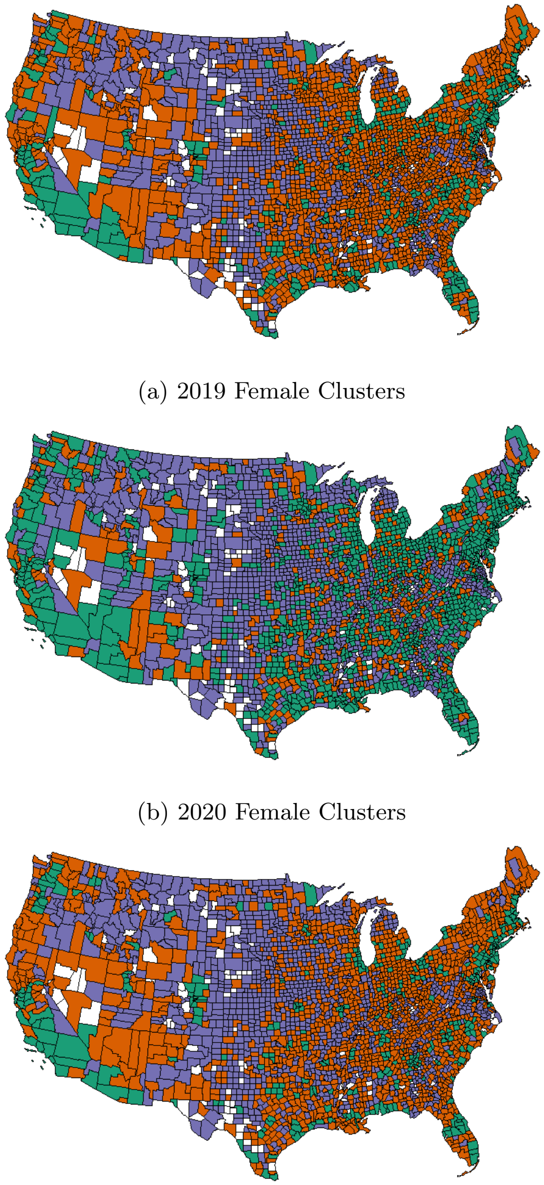
## 3.3 Cluster to Standard Curve Comparison

By clustering similar counties, we seek to gain efficiency in the prediction of future expected mortality. For example, one mortality curve per county means 3000 different curves. Regional clustering is already employed for mortality, often in the form of one single mortality curve used for each state. Many states are a mix of urban and rural areas, and so we explore mortality predictive performance when the nation is split by state versus when it is split by our clusters.

We compare the selected clustering with state and national level curves using the standard mean squared error (MSE) between true log mortality rates and predicted log mortality rates. (We also explored using other metrics such as a population-weighted mean squared error, and obtained broadly similar results.)

We evaluate two prediction methods for our models. The first method — which we name the “cumulative approach” — aggregates the data from all previous years (for all counties within the appropriate cluster) to predict a given year, along with the overall cluster identity. The second method (the “yearly approach”) predicts the mortality of a year based only on the previous year’s cluster and mortality information. Table 3 presents the model performance for these various methods.

Using 3 clusters consistently outperforms state-level models (and also national-level models, which is not surprising, as the latter can be viewed as a single-cluster model) on MSE despite using only three



(a) 2019 Female Clusters

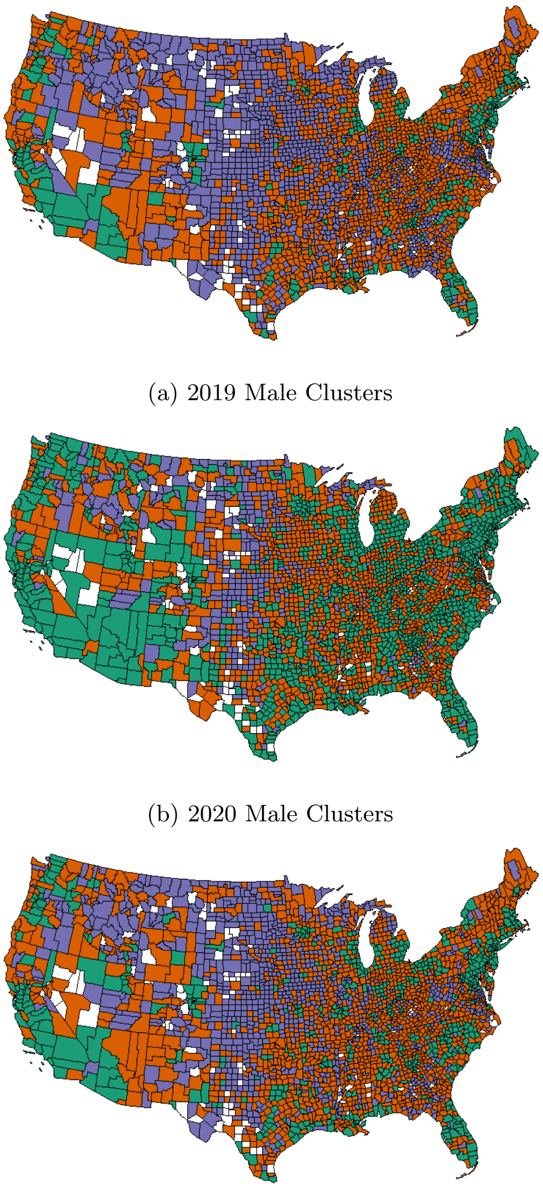
(b) 2020 Female Clusters

(c) 2021 Female Clusters

Figure 7: 2019-2021 Female Clusters

| | Cumulative | | | Yearly | | |
|---|---|---|---|---|---|---|
| | Nation | State | Cluster | Nation | State | Cluster |
| Male | 1.137 | 1.060 | 0.946 | 1.194 | 1.105 | 0.964 |
| Female | 1.735 | 1.605 | 1.408 | 1.794 | 1.652 | 1.440 |

Table 3: MSE for Prediction Models



(a) 2019 Male Clusters

(b) 2020 Male Clusters

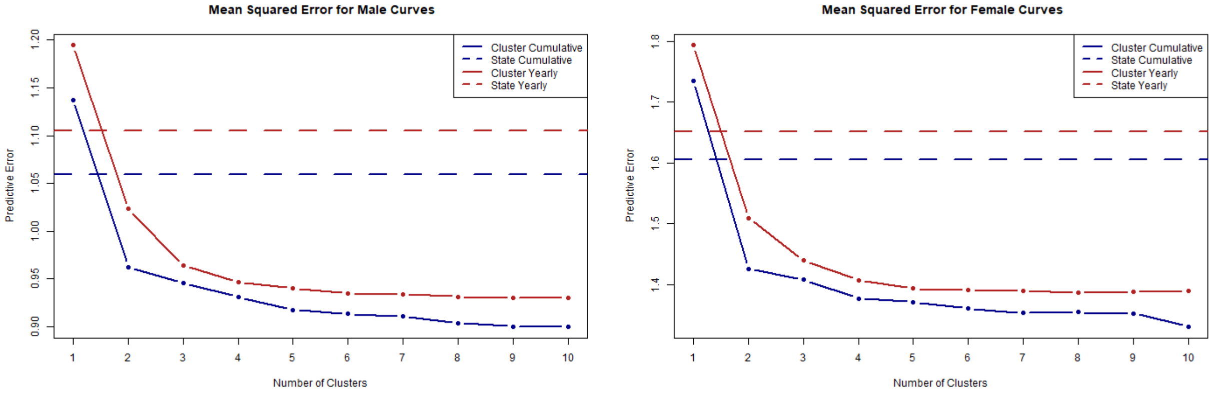
(c) 2021 Male Clusters

Figure 8: 2019-2021 Male Clusters

instead of 48 different mortality models. This efficiency suggests that three archetypal mortality curves can provide more accurate predictions for a county's mortality than the state-specific curves, significantly improving efficiency.

We do note that considering weighted MSE does cause the state-level models to slightly improve their performance relative to the cluster-level models. However, weighting MSE by population may not be the best metric. This approach primarily judges curve performance based on how well the curves predict population centers, which do not typically need aggregation. States with many rural counties often have one or two large metropolitan areas, such as their capital, and so the state aggregated curves predict these centers well but fail in rural areas. By treating each county individually, unweighted MSE reveals that low population density areas are better predicted with cluster curves, needing only two clusters to outperform state-level curves.

Figure 9 shows that increasing the number of clusters does indeed enhance predictive performance, but this improvement tends to be minimal beyond three clusters. Both the cumulative and yearly methods outperform their respective state-level models with as few as two clusters. We again note that the results are similar for the male and female models.



(a) Male MSE

(b) Female MSE

Figure 9: Curve Predictive Performance

To further analyze prediction performance, we calculate the partial MSEs of each clustering — these are shown in Table 4. We notice a few things from these results. First, the cluster-level models outperform their state-level counterparts in the areas represented by both the orange and green clusters, whereas the opposite is true for the green cluster, where state-level curves perform better. Given the nature of the clusters, this is perhaps not surprising. Consider a specific urban county in a given state. As the green cluster is composed of more urban counties, and these urban counties comprise the bulk of their states' populations, the state-level model is largely using the urban data from its own state to predict this (urban) county's mortality, whereas the cluster-level model is combining data from mostly urban counties in many states. On the other hand, the fact that the cluster-level model outperforms the state-level model in the areas represented by the orange and purple clusters indicates that there is real value to be gained in predicting rural mortality by borrowing strength from rural areas in other states. Finally, we note that both models perform much better in the areas represented by the green cluster than in those belonging to the orange cluster, and worse yet in those represented by the purple cluster.

| | Green Cluster | | Orange Cluster | | Purple Cluster | |
|---|---|---|---|---|---|---|
| | State | Cluster | State | Cluster | State | Cluster |
| Male | 0.189 | 0.232 | 0.880 | 0.791 | 3.291 | 2.805 |
| Female | 0.450 | 0.541 | 1.504 | 1.245 | 4.217 | 3.509 |

Table 4: Partial MSE for Each Cluster

Overall these results indicate that three clusters seems to be the optimal choice; the decrease in MSE is a matter of significantly diminishing returns after 3 clusters, and this clustering is also simple enough to allow for meaningful cluster interpretation, which we will explore further in Section 4.

# 4 Cluster Interpretation

We have previously suggested that the broad categories we find when we perform county level mortality clustering are urban versus rural, but there may be more nuanced causes of this divide. Here we attempt to gain further insights into the drivers of the differences in the cluster mortality curves and to try to lend interpretations to the clusters themselves where possible; we do this by examining the counties (and their underlying properties) that comprise each cluster.

## 4.1 Covariates and Clusters

We considered a number of county-level covariates, which we used to help characterize and interpret the clusters. In particular, we collected, for each year of the analysis, variables related to education level (measured by percentage of adult county residents with a bachelor's degree), percentage of county residents who were married, unemployment rate, race (measured by percentage of the heads of household who is white), and population density. In addition, we also considered average household size (by number of occupants) and land area of the county, whose values were only available for 2010. While the clustering of counties was performed both on a yearly basis in addition to the overall (composite) clustering spanning all years, we use the overall clusterings as we proceed to analyze and interpret the clusters in terms of the covariates of their counties. While the composition of the clusters is, in general, relatively stable, the movement of counties between clusters does have some effect of obscuring the relationships between the clusters and covariates.

Figure 10 shows the mean population densities (calculated as number people per square kilometer of land area) of the clusters through time. We very clearly see that the green cluster represents counties with greater population densities (i.e., urban counties), whereas the other two clusters have counties much lower population densities (i.e., more rural counties). We find that the effects for this — and all other covariates — are more or less similar in the male and female models.

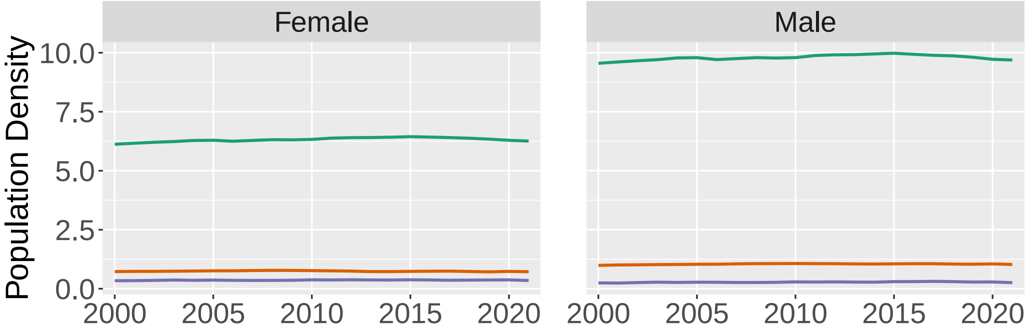


Figure 10: Population Density by Cluster

Figure 11 gives the mean land areas of the counties by cluster, which do not vary through time. Looking at these land areas, we see that the purple counties are the largest but, perhaps somewhat surprisingly, the green cluster has larger counties than does the orange cluster, at least on average. This is likely due to the presence of many counties in California and Arizona — which tend to be larger in area — in the green cluster (see Figures 4 and 5).

Figure 12 shows the percent of heads of household who are white, by cluster. The three clusters are well separated with respect to this variable, with the purple cluster have the greatest percentage of white heads of household throughout, and the green cluster having the least. We also see that there is an overall decreasing trend over time for all clusters. We note, though, that the slope of this decrease is slightly flatter for the orange cluster than for the other two clusters. (In general, the cluster curves tend to be nearly parallel through time for all of our covariates; this is the only real exception, and it is not extremely pronounced.)

Figure 13 gives the mean unemployment rates over time by cluster. Covariates tended to move smoothly and monotonically through time; unemployment rate was, not surprisingly, the lone exception.

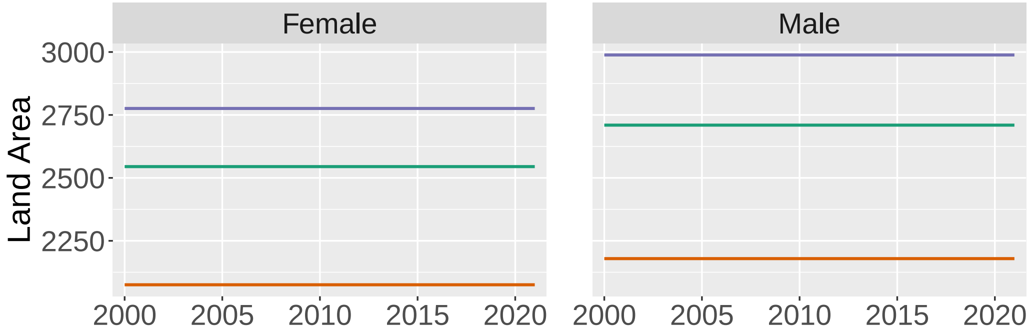


Figure 11: Land Area by Cluster

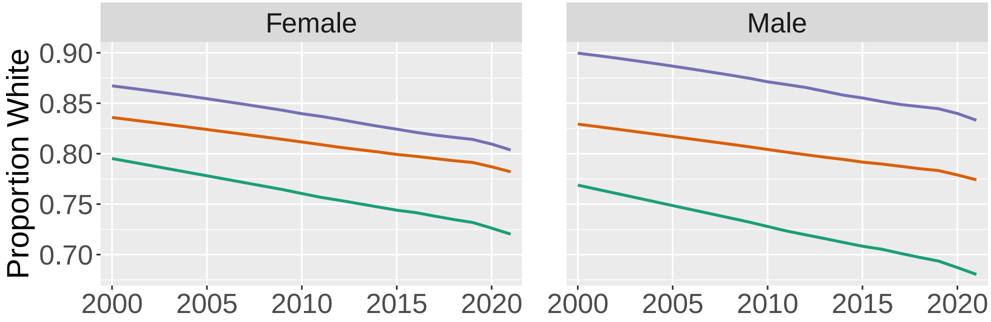


Figure 12: Proportion White Heads of Household by Cluster

Throughout the timeframe analyzed, the green and orange clusters had very similar unemployment rates, while the purple cluster was the outlier, having significantly lower rates of unemployment.

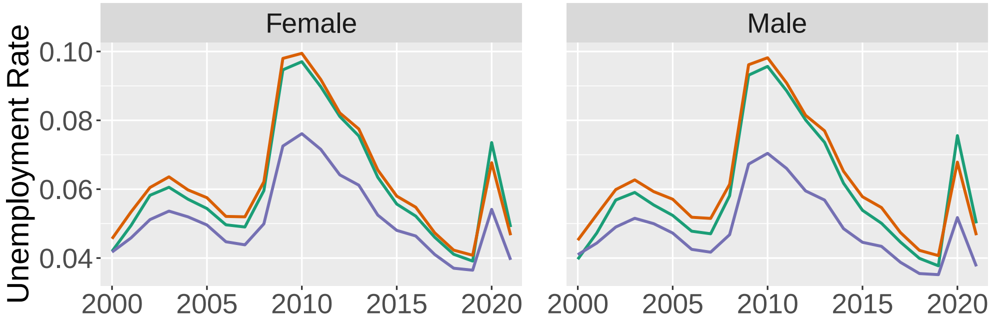


Figure 13: Unemployment Rate by Cluster

Figures 14 and 15 give the percentage of adults who have a bachelor's degree and who are married, respectively, by cluster. We can see that the former variable has an increasing trend through time, while the latter variable has the opposite trend. In terms of education level, the green cluster lies significantly above the other two. The purple cluster has the highest level of marriage, followed by the orange cluster, and then the green.

These results indicate that, as we continue to further analyze the clusters using the covariates, it is not necessary to consider each year's covariate values; rather, it is sufficient to utilize representative covariate values in our analyses. Further, we will continue to use the overall clusters rather than the clusters from individual years, in order to eliminate the noise in individual year cluster covariate values caused by the yearly movements of counties between clusters. The following sections proceed to use regression and random forest analyses to help provide interpretations to these overall clusters in terms

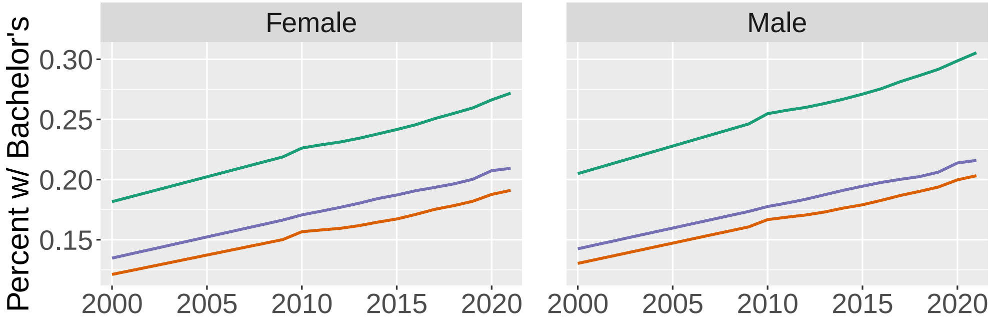


Figure 14: Education Level by Cluster

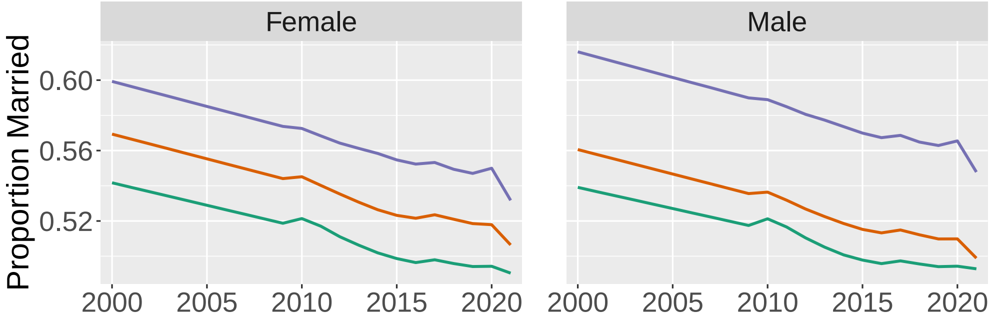


Figure 15: Marriage Level by Cluster

of the various county covariates.

## 4.2 Logistic Regression

As a means of analyzing cluster separation and interpretation, we fit a multinomial logistic regression on the overall clusters using a number of county level variables. We used all of the variables described in Section 4.1 (specifically, their 2010 values, standardized). The regression was done using the purple cluster as the base case, so the coefficients can be interpreted as the change in the log odds of being in a given cluster (as compared to the purple cluster) resulting from an increase of one standard deviation in the value of that covariate, holding all other variables constant. The values of the coefficients, and their associated standard errors and p-values are in Table 5.

There are a few notable aspects of these results. First, we notice that the coefficients for most of the these covariates are highly significant, meaning that they highly influence the probably of a county belonging to the various clusters. (Or at least the difference in probabilities of belonging to the purple cluster, as opposed to one of the other clusters, though we can see that there are also significant differences between the coefficients for the green and orange clusters.) We note that the results are broadly similar between the male and female models; almost all coefficients are the same sign, and most are of very similar magnitudes. The values of the coefficients are in general unsurprising, given the relationships discussed above. We also note that most of these covariates are themselves significantly correlated.

## 4.3 Random Forest

In addition to the regression described in Section 4.2, we also used a random forest approach to analyze and interpret the three clusters of counties. A random forest is an ensemble tree method introduced by Breiman (2001) which performs classification (or regression) by creating trees from randomly chosen observations and randomly chosen subsets of predictor variables. Methods using this type of bootstrap aggregation (or "bagging") have many benefits, including variance reduction and minimizing the chances of overfitting models (Breiman, 1996). We utilized the `randomForest` package in R (Liaw and Wiener, 2002) for this analysis.

| | Male Model | | | Female Model | | |
|---|---|---|---|---|---|---|
| | Estimate | Std. Error | P-Value | Estimate | Std. Error | P-Value |
| Intercept (Green Cluster) | 0.8059 | 0.0912 | $***$ | 1.0329 | 0.0713 | $***$ |
| Intercept (Orange Cluster) | 1.8219 | 0.0842 | $***$ | 0.7037 | 0.0778 | $***$ |
| Bachelor's (Green Cluster) | 1.0383 | 0.1011 | $***$ | 1.1148 | 0.0925 | $***$ |
| Bachelor's (Orange Cluster) | 0.0131 | 0.0953 | 0.8910 | -0.0284 | 0.0991 | 0.7741 |
| Married (Green Cluster) | -0.6059 | 0.1052 | $***$ | -0.6255 | 0.0882 | $***$ |
| Married (Orange Cluster) | -0.7471 | 0.0955 | $***$ | -0.2686 | 0.0867 | 0.0020 |
| Household Size (Green Cluster) | 1.1458 | 0.0969 | $***$ | 1.1087 | 0.0817 | $***$ |
| Household Size (Orange Cluster) | 0.8752 | 0.0873 | $***$ | 0.7397 | 0.0793 | $***$ |
| Unemployment (Green Cluster) | 0.8947 | 0.0913 | $***$ | 0.9056 | 0.0763 | $***$ |
| Unemployment (Orange Cluster) | 0.6437 | 0.0780 | $***$ | 0.6477 | 0.0731 | $***$ |
| White (Green Cluster) | 0.3995 | 0.1070 | 0.0002 | 0.7546 | 0.0864 | $***$ |
| White (Orange Cluster) | 0.6049 | 0.0946 | $***$ | 0.5246 | 0.0825 | $***$ |
| Land Area (Green Cluster) | -0.0861 | 0.0593 | 0.1464 | -0.0513 | 0.0527 | 0.3302 |
| Land Area (Orange Cluster) | -0.2127 | 0.0556 | 0.0001 | -0.1996 | 0.0628 | 0.0015 |
| Pop. Density (Green Cluster) | 0.8043 | 0.5345 | 0.1324 | 0.7919 | 0.4633 | 0.0874 |
| Pop. Density (Orange Cluster) | -0.2457 | 0.5471 | 0.6533 | 0.0021 | 0.5421 | 0.9969 |

Table 5: Coefficients for Male and Female Multinomial Regression Models. Entries marked $***$ indicate p-values less than $10^{-8}$.

We made standard choices for the tuning parameters; with 7 predictors (the same ones as were used in Section 4.2), we randomly selected $\lfloor\sqrt{7}\rfloor = 2$ of them to try at each node (Bernard et al., 2009). We used 63.2% of the 3007 observations (counties) for each subsample. We used 500 trees, and the errors were converging well within this time frame in all cases.

For both the male and female models, the random forest did very well at classifying the counties into their actual clusters. The confusion matrices are given in Table 6 for the counties, based on their out of bag samples.

Table 6: Confusion Matrices for Male and Female Random Forest County Classifier Models

(a) Male

| | | Predicted | | | Error |
|---|---|---|---|---|---|
| | | Green | Orange | Purple | Rate |
| Actual | Green | 682 | 84 | 0 | 0.1097 |
| | Orange | 56 | 1705 | 36 | 0.0512 |
| | Purple | 0 | 87 | 357 | 0.1959 |

(b) Female

| | | Predicted | | | Error |
|---|---|---|---|---|---|
| | | Green | Orange | Purple | Rate |
| Actual | Green | 1236 | 110 | 1 | 0.0824 |
| | Orange | 99 | 850 | 75 | 0.1699 |
| | Purple | 1 | 87 | 548 | 0.1384 |

The overall misclassification rates were 8.8% and 12.4% for the male and female models, respectively. We note again that for both models, these results indicate that the orange cluster was an intermediate cluster between the green and purple clusters. That is, there was some misclassification between the green and orange clusters, and between the orange and purple clusters, but virtually none between the green and purple clusters. The one county from the green cluster in the female model that was misclassified as purple was Aroostook County, ME, an unusually large county on the Canadian border; the one county from the purple cluster in the female model to be misclassified as green was Poquoson, an independent city with a small land area, located on the Virginia Peninsula. Overall, 72.2% of counties were predicted to be in the same cluster for the male and female models, as compared to the actual value of 73.8%.

The random forest model also allows us to consider the relative importance of the various predictors. In particular, for each predictor in both models, we calculated the mean decrease in out of bag prediction accuracy when permuting each predictor, while leaving the other predictors unchanged. Figure 16 shows a measure of the importance of the various predictors in the random forest algorithm.

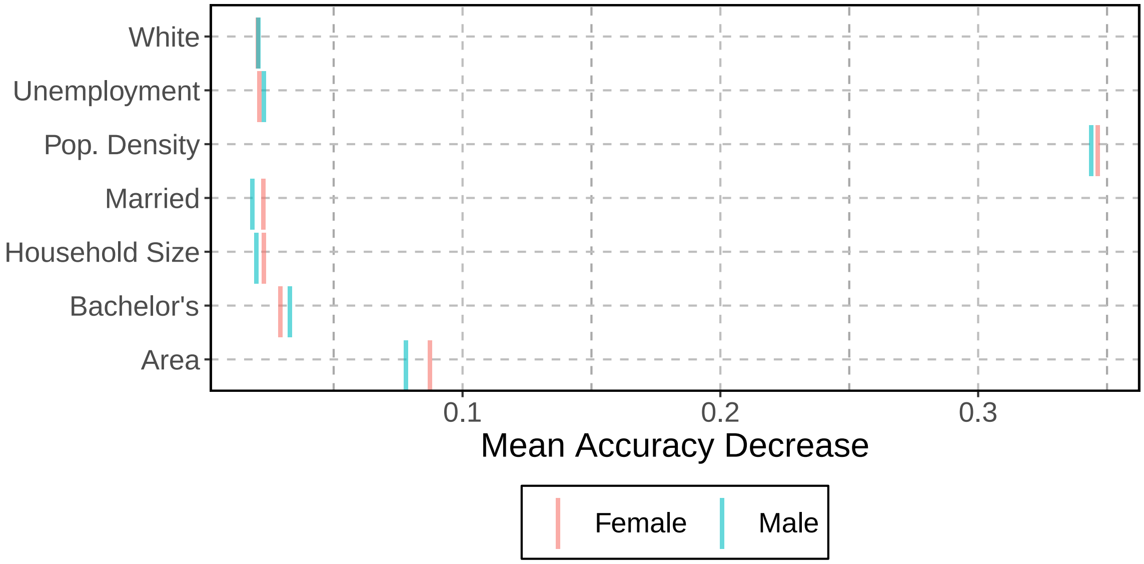


Figure 16: Variable Importance Plot for Random Forest Model

We can see again that the results are very similar for the male and female models. We note that the populations density is by far the most important predictor, with land area being a distant second, and all other predictors far less important. This lends further weight to the notion that the clusters largely differ along urban-rural lines.

## 4.4 Interpreting Model Results

Combining the information from the clustering results of the mortality curves, as seen in Figure 6, as well as the plots and analysis from Sections 4.1 through 4.3, table 7 offers a succinct summary of the differences between these clusters, both in the curves themselves as well as the observed differences in characteristics of the specific counties.

| **Characteristic** | **Green** | **Orange** | **Purple** |
|---|---|---|---|
| **Mortality (Age 40 and under)** | Low | Medium | High |
| **Mortality (Age 40 and older)** | Slightly lower than orange | High | Slightly lower than orange |
| **Population Density** | High | Low | Low |
| **Land Area** | Medium | Low | High |
| **White Population** | Lowest | Medium | Highest |
| **Unemployment Rate** | Similar to orange | Similar to green | Lower |
| **Bachelor's Degrees** | High | Similar to purple | Similar to orange |
| **Marriage Rates** | Lowest | Medium | Highest |

Table 7: Characteristics of Orange, Green, and Purple Groups by mortality curve structure and demographic information.

These groupings are going to be more complicated than these covariates can properly address, but there are still several insights that can be made.

- The green group has the highest population density, suggesting a more urban area, whereas the purple and orange groups had lower population densities suggesting more rural areas

- The purple and orange groups have higher percentage of whites and higher marriage rates, again suggesting more rural areas, but the purple group was higher in both categories than the orange, providing some distinction between the two groups.
- The purple group has low education rates and low unemployment rates, perhaps suggesting a more prominent blue collar work force.
- Even though the green and purple groups have similar mortality for individuals over 60, they are not similar in any other demographic feature.

# 5 Conclusion

Understanding mortality patterns in the United States is a crucial step in informing policy makers in related fields, such as healthcare, social services, and insurance groups. Finding the underlying behaviors, drivers, and trends of mortality in different regions of the country is one step towards this understanding.

We adopted the approach of characterizing counties based on the shape of their overall and annual mortality curves, which we did by representing them using spline regression coefficients. We found that characterizing counties in this way allowed us to discover fairly nuanced trends present in mortality data. We found that there are three primary and fairly distinct patterns that mortality tends to follow in the continental United States; moreover, the three patterns varied not only in their overall mortality level, but also in the shapes of their curves, with respect to age. In addition, we found that, in at least one sense, these three overarching patterns are better at predicting the mortality of a county than even nearby counties in the same state, which may exhibit dissimilar behavior. We gained predictive power by simplifying and using fewer curves, especially with respect to the more rural counties. This streamlined approach offers researchers and policy makers a powerful tool to navigate and understand mortality in the United States more effectively.

We also found that our clusters could be interpreted in terms of various covariates of their constituent counties. Beyond the obvious urban-rural divide, the clusters consistently varied in terms of education, race, marriage level, and even unemployment; the patterns we found remained fairly constant through time. In addition, the covariates were highly predictive of cluster identity, with the population density being by far the most important predictor.

A few directions that future research could be taken in from this point emerge. No spatial information was used in our modelling process at any point; it is possible that integrating a spatial correlation structure into our framework could yield further insight into geographical variations in mortality patterns, and could provide clarity on the regional disparities that appeared without it. Similarly, the random yearly effects were modelled without any temporal correlation; adding this into the modelling procedure could add a more intuitive change from year to year in the patterns of mortality. More covariates could also potentially lend further insights as to the nature of the clusters. Finally, exploring the causes of death, and in particular how they vary by cluster could provide valuable information regarding the reasons for the differences in the shapes of the cluster mortality curves.

# Acknowledgements

The authors would like to thank Davis Dowdle for his contributions to this work, in particular his efforts to gather, clean, and process the data used in this analysis.

# References

Beer, G., Marussig, B., and Duenser, C. (2020), *Basis Functions, B-splines*, Cham: Springer International Publishing, pp. 35–71.

Bernard, S., Heutte, L., and Adam, S. (2009), "Influence of hyperparameters on random forest accuracy," in *Multiple Classifier Systems: 8th International Workshop, MCS 2009, Reykjavik, Iceland, June 10-12, 2009. Proceedings 8*, Springer, pp. 171–180.

Bouveyron, C., and Jacques, J. (2011), "Model-based clustering of time series in group-specific functional subspaces," *Advances in Data Analysis and Classification*, 5, 281–300.

Breiman, L. (1996), "Bagging predictors," *Machine learning*, 24, 123–140.

—— (2001), "Random forests," *Machine learning*, 45, 5–32.

Brown, J. R., and Orszag, P. R. (2006), "The Political Economy of Government-Issued Longevity Bonds," *Journal of Risk and Insurance*, 73, 611–631.

Caliński, T., and Harabasz, J. (1974), "A dendrite method for cluster analysis," *Communications in Statistics-theory and Methods*, 3, 1–27.

Case, A., and Deaton, A. (2017), "Mortality and morbidity in the 21st century," *Brookings Papers on Economic Activity*, 2017, 397–476.

Charrad, M., Ghazzali, N., Boiteau, V., and Niknafs, A. (2014), "NbClust: an R package for determining the relevant number of clusters in a data set," *Journal of statistical software*, 61, 1–36.

Chetty, R., Stepner, M., Abraham, S., Lin, S., Scuderi, B., Turner, N., Bergeron, A., and Cutler, D. (2016), "The association between income and life expectancy in the United States, 2001-2014," *Jama*, 315, 1750–1766.

Clark, C. R., and Williams, D. R. (2016), "Understanding county-level, cause-specific mortality: the great value—and limitations—of small area data," *Jama*, 316, 2363–2365.

Currie, J., and Schwandt, H. (2016), "Mortality inequality: The good news from a county-level approach," *Journal of Economic Perspectives*, 30, 29–52.

Dwyer-Lindgren, L., Bertozzi-Villa, A., Stubbs, R. W., Morozoff, C., Kutz, M. J., Huynh, C., Barber, R. M., Shackelford, K. A., Mackenbach, J. P., Van Lenthe, F. J., et al. (2016), "US county-level trends in mortality rates for major causes of death, 1980-2014," *Jama*, 316, 2385–2401.

Faraway, J. J. (1997), "Regression analysis for a functional response," *Technometrics*, 39, 254–261.

Gavrilov, L. A., and Gavrilova, N. S. (2011), "Mortality measurement at advanced ages: a study of the Social Security Administration Death Master File," *North American actuarial journal*, 15, 432–447.

Gibbs, Z., Groendyke, C., Hartman, B., and Richardson, R. (2020), "Modeling county-level spatio-temporal mortality rates using dynamic linear models," *Risks*, 8, 117.

Gompertz, B. (1825), "XXIV. On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies. In a letter to Francis Baily, Esq. FRS &c," *Philosophical transactions of the Royal Society of London*, 513–583.

Ho, J. Y., and Hendi, A. S. (2018), "Recent trends in life expectancy across high income countries: Retrospective observational study," *BMJ*, 362, k2562.

Kindig, D. A., and Cheng, E. R. (2013), "Even as mortality declines in many places, life expectancy continues to diverge between counties in the United States," *Health Affairs*, 32, 451–458.

Kleinman, J. (1977), "Statistical Notes for Health Planners," *Mortality*, 16, 77–1237.

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

MacQueen, J., et al. (1967), “Some methods for classification and analysis of multivariate observations,” in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, Oakland, CA, USA, Vol. 1, pp. 281–297.

Makeham, W. M. (1860), “On the law of mortality and the construction of annuity tables,” *Journal of the Institute of Actuaries*, 8, 301–310.

Marmot, M., Shipley, M., Rose, G., and Thomas, B. (1981), “Alcohol and mortality: a U-shaped curve,” *The lancet*, 317, 580–583.

Masters, R. K., Link, B. G., and Phelan, J. C. (2015), “Trends in education gradients of ’preventable’ mortality: A test of fundamental cause theory,” *Social Science & Medicine*, 127, 19–28.

Monnat, S. M. (2018), “Factors associated with county-level differences in US drug-related mortality rates,” *American journal of preventive medicine*, 54, 611–619.

National Center for Health Statistics (2023), “Detailed Mortality - All Counties (all states, all counties, plus cities of 100,000 or more population) (2000-2021), as compiled from data provided by the 57 vital statistics jurisdictions through the Vital Statistics Cooperative Program.” .

R Core Team (2024), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

Remund, A., Camarda, C. G., and Riffe, T. (2018), “A cause-of-death decomposition of young adult excess mortality,” *Demography*, 55, 957–978.

Singh, G. K., and Siahpush, M. (2014), “Widening rural–urban disparities in life expectancy, US, 1969–2009,” *American journal of preventive medicine*, 46, e19–e29.

Tarpey, T., and Kinateder, K. K. (2003), “Clustering functional data.” *Journal of classification*, 20.

Tibshirani, R., Walther, G., and Hastie, T. (2001), “Estimating the number of clusters in a data set via the gap statistic,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63, 411–423.

Ward Jr, J. H. (1963), “Hierarchical grouping to optimize an objective function,” *Journal of the American statistical association*, 58, 236–244.

Wilmoth, J. R., and Horiuchi, S. (1999), “Rectangularization of the survival curve in low-mortality countries,” *Population and Development Review*, 25, 239–257.

Woolf, S. H., Chapman, D. A., Buchanich, J. M., Bobby, K. J., Zimmerman, E. B., and Blackburn, S. M. (2018), “Changes in midlife death rates across racial and ethnic groups in the United States: Systematic analysis of vital statistics,” *BMJ*, 362, k3096.

Xu, J., Murphy, S. L., Kochanek, K. D., and Arias, E. (2020), “Mortality in the United States, 2018,” *NCHS Data Brief*.

Yang, T.-C., Noah, A. J., and Shoff, C. (2015), “Exploring geographic variation in US mortality rates using a spatial Durbin approach,” *Population, space and place*, 21, 18–37.

# Appendix A County Merges

| Merge # | State | Original County | | Merged Into County | |
|---|---|---|---|---|---|
| | | FIPS | County Name | FIPS | County Name |
| 1 | AL | 01011 | Bullock County | 01101 | Montgomery County |
| 2 | AR | 05037 | Cross County | 05035 | Crittenden County |
| 3 | CA | 06003 | Alpine County | 06017 | El Dorado County |
| 4 | CO | 08049 | Grand County | 08069 | Larimer County |
| 5 | CO | 08053 | Hinsdale County | 08067 | La Plata County |

| | | | | | |
|---|---|---|---|---|---|
| 6 | CO | 08057 | Jackson County | 08069 | Larimer County |
| 7 | CO | 08079 | Mineral County | 08007 | Archuleta County |
| 8 | CO | 08093 | Park County | 08059 | Jefferson County |
| 9 | CO | 08111 | San Juan County | 08067 | La Plata County |
| 10 | GA | 13007 | Baker County | 13095 | Dougherty County |
| 11 | GA | 13035 | Butts County | 13151 | Henry County |
| 12 | GA | 13037 | Calhoun County | 13095 | Dougherty County |
| 13 | GA | 13053 | Chattahoochee County | 13215 | Muscogee County |
| 14 | GA | 13181 | Lincoln County | 13073 | Columbia County |
| 15 | GA | 13271 | Telfair County | 13069 | Coffee County |
| 16 | GA | 13307 | Webster County | 13261 | Sumter County |
| 17 | ID | 16025 | Camas County | 16039 | Elmore County |
| 18 | ID | 16033 | Clark County | 16051 | Jefferson County |
| 19 | IL | 17069 | Hardin County | 17165 | Saline County |
| 20 | KS | 20081 | Haskell County | 20055 | Finney County |
| 21 | KY | 21005 | Anderson County | 21073 | Franklin County |
| 22 | KY | 21063 | Elliott County | 21043 | Carter County |
| 23 | KY | 21105 | Hickman County | 21083 | Graves County |
| 24 | KY | 21129 | Lee County | 21065 | Estill County |
| 25 | KY | 21165 | Menifee County | 21173 | Montgomery County |
| 26 | KY | 21197 | Powell County | 21049 | Clark County |
| 27 | KY | 21237 | Wolfe County | 21175 | Morgan County |
| 28 | LA | 22035 | East Carroll Parish | 22083 | Richland Parish |
| 29 | LA | 22091 | St. Helena Parish | 22033 | East Baton Rouge Parish |
| 30 | LA | 22107 | Tensas Parish | 22041 | Franklin Parish |
| 31 | MS | 28023 | Clarke County | 28075 | Lauderdale County |
| 32 | MS | 28055 | Issaquena County | 28151 | Washington County |
| 33 | MS | 28063 | Jefferson County | 28085 | Lincoln County |
| 34 | MS | 28069 | Kemper County | 28075 | Lauderdale County |
| 35 | MS | 28097 | Montgomery County | 28043 | Grenada County |
| 36 | MS | 28163 | Yazoo County | 28049 | Hinds County |
| 37 | MT | 30007 | Broadwater County | 30031 | Gallatin County |
| 38 | MT | 30025 | Fallon County | 30017 | Custer County |
| 39 | MT | 30055 | McCone County | 30085 | Roosevelt County |
| 40 | MT | 30069 | Petroleum County | 30027 | Fergus County |
| 41 | MT | 30107 | Wheatland County | 30027 | Fergus County |
| 42 | MT | 30109 | Wibaux County | 30083 | Richland County |
| 43 | NE | 31005 | Arthur County | 31101 | Keith County |
| 44 | NE | 31009 | Blaine County | 31041 | Custer County |
| 45 | NE | 31057 | Dundy County | 31029 | Chase County |
| 46 | NE | 31075 | Grant County | 31031 | Cherry County |
| 47 | NE | 31085 | Hayes County | 31111 | Lincoln County |
| 48 | NE | 31105 | Kimball County | 31033 | Cheyenne County |
| 49 | NE | 31113 | Logan County | 31111 | Lincoln County |
| 50 | NE | 31115 | Loup County | 31041 | Custer County |
| 51 | NE | 31117 | McPherson County | 31111 | Lincoln County |
| 52 | NE | 31171 | Thomas County | 31031 | Cherry County |
| 53 | NV | 32009 | Esmeralda County | 32023 | Nye County |
| 54 | NV | 32011 | Eureka County | 32007 | Elko County |
| 55 | NV | 32015 | Lander County | 32007 | Elko County |
| 56 | NV | 32017 | Lincoln County | 32003 | Clark County |
| 57 | NV | 32021 | Mineral County | 32019 | Lyon County |
| 58 | NV | 32027 | Pershing County | 32031 | Washoe County |
| 59 | NV | 32029 | Storey County | 32031 | Washoe County |
| 60 | NM | 35011 | De Baca County | 35005 | Chaves County |
| 61 | NM | 35033 | Mora County | 35049 | Santa Fe County |
| 62 | NC | 37095 | Hyde County | 37013 | Beaufort County |
| 63 | NC | 37103 | Jones County | 37133 | Onslow County |
| 64 | NC | 37177 | Tyrrell County | 37055 | Dare County |

| | | | | | |
|---|---|---|---|---|---|
| 65 | ND | 38007 | Billings County | 38089 | Stark County |
| 66 | OK | 40057 | Harmon County | 40065 | Jackson County |
| 67 | SD | 46017 | Buffalo County | 46015 | Brule County |
| 68 | SD | 46041 | Dewey County | 46129 | Walworth County |
| 69 | SD | 46063 | Harding County | 46019 | Butte County |
| 70 | SD | 46075 | Jones County | 46085 | Lyman County |
| 71 | SD | 46102 | Oglala Lakota County | 46103 | Pennington County |
| 72 | SD | 46113 | Shannon County | 46102 | Oglala Lakota County |
| 73 | SD | 46137 | Ziebach County | 46093 | Meade County |
| 74 | TN | 47027 | Clay County | 47111 | Macon County |
| 75 | TX | 48033 | Borden County | 48227 | Howard County |
| 76 | TX | 48075 | Childress County | 48197 | Hardeman County |
| 77 | TX | 48087 | Collingsworth County | 48483 | Wheeler County |
| 78 | TX | 48105 | Crockett County | 48465 | Val Verde County |
| 79 | TX | 48137 | Edwards County | 48265 | Kerr County |
| 80 | TX | 48173 | Glasscock County | 48329 | Midland County |
| 81 | TX | 48229 | Hudspeth County | 48141 | El Paso County |
| 82 | TX | 48235 | Irion County | 48451 | Tom Green County |
| 83 | TX | 48243 | Jeff Davis County | 48371 | Pecos County |
| 84 | TX | 48259 | Kendall County | 48029 | Bexar County |
| 85 | TX | 48261 | Kenedy County | 48215 | Hidalgo County |
| 86 | TX | 48267 | Kimble County | 48265 | Kerr County |
| 87 | TX | 48269 | King County | 48207 | Haskell County |
| 88 | TX | 48271 | Kinney County | 48323 | Maverick County |
| 89 | TX | 48283 | La Salle County | 48479 | Webb County |
| 90 | TX | 48301 | Loving County | 48389 | Reeves County |
| 91 | TX | 48311 | McMullen County | 48013 | Atascosa County |
| 92 | TX | 48369 | Parmer County | 48117 | Deaf Smith County |
| 93 | TX | 48383 | Reagan County | 48451 | Tom Green County |
| 94 | TX | 48431 | Sterling County | 48451 | Tom Green County |
| 95 | TX | 48443 | Terrell County | 48465 | Val Verde County |
| 96 | TX | 48507 | Zavala County | 48323 | Maverick County |
| 97 | UT | 49009 | Daggett County | 49047 | Uintah County |
| 98 | UT | 49033 | Rich County | 49057 | Weber County |
| 99 | VA | 51045 | Craig County | 51121 | Montgomery County |
| 100 | VA | 51515 | Bedford City | 51019 | Bedford County |
| 101 | VA | 51720 | Norton City | 51195 | Wise County |
| 102 | WY | 56035 | Sublette County | 56037 | Sweetwater County |

Table 8: The collection of counties which were merged together, with the counties that were merged into others being on the left, and the counties which they were merged into being on the right.

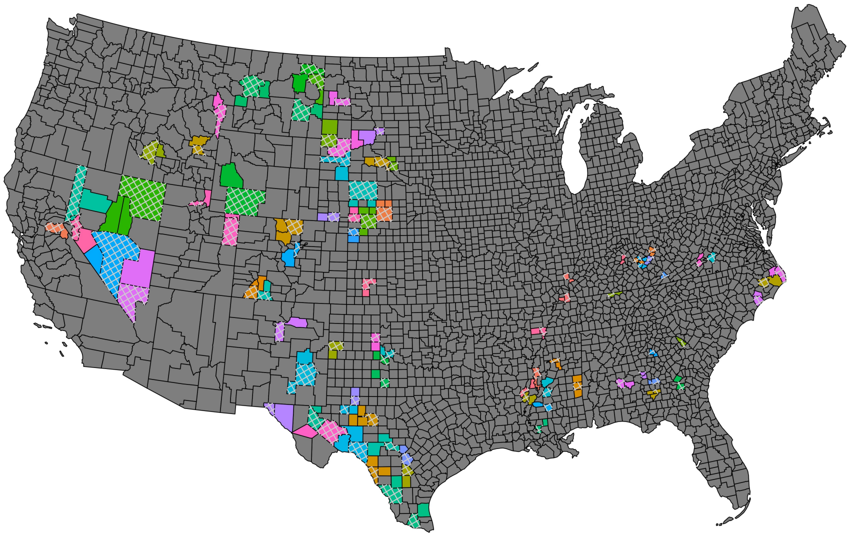


Figure 17: County Merges. Colored counties were absorbed into the cross-hatched counties of the same color.