Identifying Patterns of Abridged Life Table Elements

By

Alice Elizabeth Curtis

Thesis

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

MASTER OF SCIENCE

in

Biostatistics

June 30, 2017

Nashville, Tennessee

Approved:

Robert E. Johnson, Ph.D.

Thomas G. Stewart, Ph.D.

ACKNOWLEDGEMENTS

I would like to acknowledge both Dr. Robert E. Johnson and Dr. Thomas G. Stewart. Thank you, both for guiding me through the research and the writing process for my thesis. I greatly appreciate how generous you have both been in offering me your invaluable time.

I would also like to acknowledge the Department of Biostatistics for accepting me into the Master of Science program. I have gained invaluable information from the incredible professors within the department who taught me and will utilize this knowledge for years to come.

To all of my fellow peers within my cohort, I have had an amazing time with you all, and good luck to all of you in your future endeavours.

Finally, I would like to acknowledge my husband, Austin Ryan Townsend, who has supported me through this challenging process.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

Introduction

1.1 Introduction

The main purpose of this research is to search for patterns of life table elements amongst counties of the United States. Various calculations are computed within an abridged life table, such as the probability of dying, survival probability, and life expectancy. The abridged life table is a specific type of life table which is base upon a hypothetical cohort of 100,000 persons and uses longer age intervals than a complete life table. In order to run the various algorithms to identify separations of life table elements amongst the counties, population counts and the number of deaths by age group were needed. The CDC provides various publicly available datasets online. The multiple cause of death data is a collection of data which contains the necessary information in order to calculate each county's abridged life table elements. The various elements within the abridged life tables were calculated using the formulas detailed within a paper entitled "Life Table and Mortality Analysis" written by Chin Long Chiang (1978). The formulas which are utilized include the proportion of individuals dying, the age-specific death rate, the expected residual life, and the standard error of the expected residual life. The statistical programming language known as "R" was implemented throughout this research to run the various algorithms and create the graphics based upon the results of those algorithms. Methods were developed and tested on "stable" counties and then applied to "unstable" counties. Stability relates to, in part, reasonably small errors of life expectancy estimates. Parititioning Around Medoids (PAM), a clustering algorithm, was implemented in order to identify the various clusters of counties based on the different life table elements. The Multi-dimensional Scaling (MDS) algorithm was used to view the various life table elements within two dimensions to begin to investigate

separation of clusters. Distance matrices had to be calculated before running both the MDS and PAM algorithms in which each entry of the matrix represented the pair-wise distance between two counties. The two distance measures which were used within this research included the Euclidean and Hellinger distances. The Euclidean distance was used for the probabilities of dying, the full life expectancies, the residual life expectancies, and the survival probabilites. The Hellinger distance requires that the inputs for the matrix be based on a probability mass function which sums to one. This restricted the use of the Hellinger distance measure to only the probabilities of dying. Various patterns of life table elements were obtained using the methods detailed in this research. Future work can implement these patterns of life table elements to predict the life table elements in which smaller geographical areas experience these low death counts and unreliable measures.

1.2 Background

Through the prevention and control of injury, disability, and disease in the United States and internationally, the primary goal of the CDC (Centers for Disease Control and Prevention) is to protect public health and safety. Initially, the CDC was formed to combat the spread of malaria and was a new agency under the branch of the United States Public Health Service (PHS). Engineers and entomologists were the only professionals within the initial years of the CDC. In 1957 the Venereal Disease Division of the United States Public Health Service was transferred to the CDC, which expanded its original scope of malaria control to include sexually transmitted diseases. In 1960 Tuberculosis control was also transferred from PHS to the CDC, and in 1963 the Immunisation program was established. The CDC's scope has currently expanded to include injury control, disabilities, workplace hazards, chronic diseases, terrorism preparedness, and environmental health threats.

CDC Wide-ranging ONline Data for Epidemiologic Research (CDC WONDER) makes many health-related datasets available to the public health community by implementing a web application. Users of the web application include the general public, CDC surveillance programs, local and state health departments, healthcare providers, and academic researchers. Topics amongst the 20 collections of public datasets include United States population estimates, vaccinations, Tuberculosis cases, vaccinations, environmental exposures, births, deaths, and cancer diagnoses. The web application provides public access to maps, charts, summary statistics, and most importantly, data extractions. Some of the collections of datasets are updated weekly or monthly, whilst the majority is updated annually. The data are available in a multitude of file formats including spreadsheet files which can easily be imported into R, an open-source statistical programming language, which can then be analyzed with various statistical methods.

The Multiple Cause of Death data is a collection of data available within CDC WON-

DER which displays, in particular, the county-level national mortality and population data. The data are based on United States residents death certificates, filed within all fifty states and the District of Columbia, which are collected from the years 1999-2014. Deaths of non-residents and fetal deaths are excluded from the data. NCHS codes the death certificates from copies of the original documents provided by the State registration offices or are encrypted by the states and provided to NCHS through the Vital Statistics Cooperative Program. Each death certificate includes demographic data, up to a total of twenty multiple causes of death, and a single underlying cause of mortality. "The number of deaths, crude death rates, age-adjusted death rates and 95% confidence intervals for death rates can be obtained by cause of death (4 digit ICD-10 codes, 113 selected causes of death, 130 selected causes of infant death, drug and alcohol related causes of death, injury intent and injury mechanism categories), place of residence (national, region, division, state, and county), age (single-year-of-age, 5-year age groups, 10-year age groups and infant age groups), race (American Indian or Alaskan Native, Asian/Pacific Islander, Black or African American, White), Hispanic ethnicity, gender and year. Data are also available by place of death, month and weekday of death, and whether an autopsy was performed" ("Multiple Cause of Death, 1999-2015 Request - CDC WONDER"). The data obtained for this particular research focuses on county-level death rates broken down by ten-year age groups.

Various policies protect the Multiple Cause of Death dataset collection. First of all, the data are provided for the purpose of statistical analysis and reporting only. The datasets are not allowed to be utilized with other datasets to identify any particular individual. Finally, Sub-national geographical areas with 9 or fewer death or birth counts cannot be presented or published.

The focus of this research is based upon abridged life table calculations which were obtained utilising the ten-year age group multiple causes of death data for all counties in the

United States. One significant concern with abridged life tables is that the number of fatalities reported in various counties may be unrepresentative of that particular area of interest due to unreliable resources. Death counts within small geographical regions tend to be zero or a small number which causes the conditional probability of death to be infinitesimal or zero.

CHAPTER 2

Methods

2.1 Life Expectancy and Abridged Life Tables

Life tables contain mortality and survival statistics on a hypothetical cohort of 100,000 persons, followed from birth to death based on current population age-specific mortality statistics and population estimates. The primary element of the life table is the probability of death over the age interval $(x_i,x_{i+1}]$, given survival to age x_i, which is denoted as q_i. The following calculations were obtained from the paper entitled "Life Table and Mortality Analysis" written by Chin Long Chiang in 1978, which is a standard method for calculating abridged life table elements. A complete life table is constructed using single year age intervals. An abridged life table uses longer, often varying, age intervals. An abridged life table is constructed similarly as a complete life table, except that the main difference is the length of the intervals. An interval, $(x_i,x_{i+1}]$, in an abridged life table is identified by x_i and n_i (length of interval), which is equal to $x_{i+1} - x_i$. A common length for the interval is five years. The fraction of i^{th} age interval of life, denoted as a_i, is the average fraction of the interval lived amongst the people who die at an age within the interval. N_i is known as the number of individuals alive at age x_i, and D_i is the number of people who die within the interval $(x_i,x_{i+1}]$. The following gives the proportion of individuals dying in the interval

$$\hat{q}_i = \frac{D_i}{N_i} \tag{2.1}$$

The ratio of D_i to the total number of years lived by N_i individuals during the interval $(x_i,x_{i+1}]$ is known as the age-specific death rate M_i and is calculated as follows

$$M_i = \frac{D_i}{(N_i - D_i)n_i + a_i n_i D_i}\ , \tag{2.2}$$

where $(N_i - D_i)n_i$ represents the number of years lived by the people who survived within the interval and $a_i n_i D_i$ represents the number of years lived within the interval among those who died. To eliminate N_i from the equation for the proportion of individuals dying in the interval (q_i), we must first solve for N_i in (2).

$$\begin{aligned}
M_i &= \frac{D_i}{(N_i - D_i)n_i + a_i n_i D_i} \\
M_i[(N_i - D_i)n_i + a_i n_i D_i] &= D_i \\
M_i N_i n_i - M_i D_i n_i + M_i a_i n_i D_i &= D_i \\
M_i N_i n_i &= D_i + M_i D_i n_i - M_i a_i n_i D_i \\
M_i N_i n_i &= D_i[1 + M_i(n_i - a_i n_i)] \\
N_i &= \frac{D_i[1 + M_i(n_i - a_i n_i)]}{M_i n_i}
\end{aligned}$$

Then we plug this into equation (1)

$$\begin{aligned}
q_i &= \frac{D_i}{\frac{D_i[1+M_i(n_i - a_i n_i)]}{M_i n_i}} \\
&= D_i \times \frac{M_i n_i}{D_i[1 + M_i(n_i - a_i n_i)]} \\
&= \frac{M_i n_i}{1 + M_i(n_i - a_i n_i)} \\
&= \frac{M_i n_i}{1 + (1 - a_i)n_i M_i}
\end{aligned}$$

We have now obtained the following basic equation for proportion of individuals dying in the interval

$$q_i = \frac{M_i n_i}{1 + (1 - a_i)n_i M_i} \tag{2.3}$$

M_i, the age-specific death rate, can be estimated from

$$M_i = \frac{D_i}{P_i} \tag{2.4}$$

where P_i is equivalent to the mid-year population in the i^{th} age group. All other quantities within the abridged life table are functions of q_i, a_i and the radix l_0 (which is an arbitrary number usually 100,000, the cohort size at birth). The number of cohort deaths, d_i in $(x_i,x_{i+1}]$, and the number of survivors, l_{i+1} at age x_{i+1}, are computed as

$$d_i = l_i \hat{q}_i \tag{2.5}$$

$$l_{i+1} = l_i - d_i \tag{2.6}$$

for $i = 0, 1, \ldots, w-1$, where w represents the starting point for the final age interval. The number of years lived in the interval by the l_i survivors at age x_i, person-years lived, is calculated as

$$L_i = n_i(l_i - d_i) + a_i n_i d_i \tag{2.7}$$

for $i = 0, 1, \ldots, w-1$. The final age interval is an open age interval, which means the information needed to calculate the average number of years lived by an individual beyond age w is not available. L_w is therefore calculated the same as in the complete life table as follows,

$$L_w = \frac{l_w}{M_w}\ , \tag{2.8}$$

where M_w is the age-specific death rate for people age x_w and over. T_i, the total number of years remaining to all the people attaining age x_i, is

$$T_i = \sum_{j=i}^{w} L_j \tag{2.9}$$

The average number of years remaining per cohort person after age x_i is

$$\hat{e}_i = \frac{T_i}{l_i}$$

or

$$\hat{e}_i = \frac{L_i + L_{i+1} + \ldots + L_w}{l_i} \tag{2.10}$$

for $i = 0, \ldots, w$. The quantity $\hat{e}_i$ is called the *expected residual life* after attaining age x_i. The term *life expectancy* usually refers to $\hat{e}_0$. Based on Chiang (1978), the standard error of $\hat{e}_i$ is shown below.

$$s(e_k) = \sqrt{\sum_{i=k}^{w-1} l_k^2 [e_{k+1} + n_k - a_k]^2 \hat{V}(q_k)}$$

where

$$\hat{V}(q_k) = \begin{cases} q_k^2 (1 - q_k)/D_k & \text{if } D_k > 0 \\ 0 & \text{if } D_k = 0 \end{cases}$$

2.2 Elements of Abridged Life Tables

The various elements of the abridged life table are displayed in the chart below. A detailed explanation of each element of the abridged life table follows.

Age Interval	M_i	a_i	$\hat{q}_i$	l_i	d_i	L_i	T_i	$\hat{e}_i$	s($\hat{e}_i$)
$(x_i, x_{i+1}]$									

- M_i: Age-specific death rate. $M_i = \frac{D_i}{P_i}$.
- a_i: Average fraction of interval lived amongst people whom die at an age within the interval.
- $\hat{q}_i$: Conditional probability of dying within the interval. $\hat{q}_i = \frac{M_i n_i}{[1+(1-a_i)n_i M_i]}$.
- l_i: Number of survivors at the beginning of the age interval. Usually $l_0 = 100,000$.
- d_i: The number of cohort deaths over the age interval. $d_i = l_i \hat{q}_i$.
- L_i: The number of years lived in the interval by l_i survivors at age x_i. $L_i = n_i(l_i - d_i) + a_i n_i d_i$. Note: the final age interval is $L_w = \frac{l_w}{M_w}$, where w represents the starting point for the final age interval and M_w is the age-specific death rate for people age x_w and over.
- T_i: The total number of years remaining to all people attaining age x_i. $T_i = \sum_{j=i}^{w} L_j$, where the age intervals are indexed by 0,1,...,w.
- $\hat{e}_i$: Expected years of life beyond age x_i given survival to x_i. Life expectancy (LE) is usually measured from birth, $\hat{e}_0$. $\hat{e}_i = T_i/l_i$, or rather, $\hat{e}_i = \frac{L_i + L_{i+1} + ... + L_w}{l_i}$
- s($\hat{e}_i$): Standard error of $\hat{e}_i$

An assumption that is made in the construction of the abridged life table is that the life expectancy calculation assumes no migration occurs in or out of the county of interest. All

entries and exits into the calculations for the life table are due to births and deaths alone. Another assumption is that the boundaries or areas remain constant over extended periods of time. Many concerns arise on the accuracy and precision when attempts are made to estimate and interpret the life expectancy calculations for small geographical regions. One major concern is that these areas have zero death counts or a limited number of mortality counts within some age groups. When the death counts of these regions are zero or few, this causes the conditional probability of death for a particular age interval to be infinitesimal or equal to zero. Smaller geographic regions may not have reliable resources to collect the actual data representative of that particular area. Another concern is the instability of the population estimates. In smaller regions of the United States, there may be low numbers of persons within some of the age groups. Because these are estimates, they are also subject to variability.

There are various methods to combat against this issue. One such method is to substitute the zero death counts and replace $D_i = 0$ with 0.693 or 3.0 as suggested by Silcocks et al., 2001; Eayres and Williams, 2004. Another method is to replace M_i with a value of a correspondingly larger area which has similar demographic measures. Another suggestion is to aggregate more years of data, which may not always be possible if more years of data are not available. Age groups selected by the user when extracting the data from CDC WONDER can also be collapsed into larger intervals. A final method suggested is to group various geographical areas together to have more reliable estimates of life expectancy.

Another idea is to use the information from reliable geographic regions, i.e. counties which have death counts and mean standard errors that are above a certain threshold, to predict the pattern of survival that the county with unreliable estimates is most probable to fall within. This research demonstrates a method to classify patterns of survival estimates for abridged life table calculations on counties with stable estimates so that counties with unreliable es-

timates can be categorised to a particular pattern. The next few sections explain in detail the procedure of identifying the patterns of stable survival estimates.

2.3 Multidimensional Scaling (MDS)

Before the classification of the counties is acquired, it is of interest to view the data in two dimensions to investigate separation between the counties for the particular abridged life table measures of interest, such as the exact probability of death or full life expectancy. Multidimensional scaling allows for the data to be transformed in a way that allows the user to view the set of data in multiple dimensions. For this research, only two dimensions are specified to look at the separation of the exact probabilities of mortality within each age interval for two particularly distinct counties. However, a higher dimensional view may be required to adequately discern separation of classes.

The level of similarity of individual rows of a dataset can be visualised using a method known as Multidimensional scaling (MDS). An MDS algorithm aims to preserve the between row distances while placing the rows in N-dimensional space. Coordinates are designated to each of the rows which display them in the N-dimensional space. The user can then begin to investigate the similarity and dissimilarity of various observations within the dataset. Metric MDS minimises a residual sum of squares function using a procedure called *stress majorization*. The residual sum of squares function is known as the "stress function".

$$\text{STRESS}_D(v_1, v_2, ..., v_I) = \sum_{i \neq j = 1,...,N} (d_{ij} - ||v_i - v_j||^2)^{\frac{1}{2}}$$

$d_{i,j}$ represents the distances between the i^{th} and j^{th} objects and Δ represents the dissimilarity matrix whose entries are the $d_{i,j}$'s as seen below.

$$\Delta = \begin{pmatrix} d_{1,1} & d_{1,2} & d_{1,3} & \dots & d_{1,I} \\ d_{2,1} & d_{2,2} & d_{2,3} & \dots & d_{2,I} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ d_{I,1} & d_{I,2} & d_{I,3} & \dots & d_{I,I} \end{pmatrix}$$

The goal of Metric MDS is given the dissimilarity matrix, Δ, to find the I vectors $v_1, ..., v_I \in \mathbb{R}^N$ such that $||v_i - v_j|| \equiv d_{i,j} \forall i, j \in \{1, ..., I\}$ where $||\cdot||$ is a vector norm. For this research $||\cdot||$ is strictly the Euclidean norm. However, the Euclidean distances and Hellinger distances (Harsha and Prahladh, 2011) between the rows of the dataset are utilized for the dissimilarity matrix. Metric MDS requires that the dissimilarity matrix implement distances that are considered proper metrics based on various properties. Both the Euclidean and Hellinger distances are considered proper metrics. The Metric MDS algorithm is implemented to view the dissimilarity matrices in two-dimensional space in order to see the separation of two particular counties' exact probabilities of mortality. Metric MDS allows the researcher to investigate separation of datasets, but to classify the counties into various groups based on different abridged life table elements, another method is needed to identify the potential patterns or clusters.

2.4 Pattern Classification Method

The process of classifying the counties with stable estimates begins with identifying the various life table elements that are used for classification. In the life expectancy and abridged life table section above, the measure l_i is the number of survivors at the beginning of the i^{th} age interval, where l_0 in this particular application is assumed to be $100,000$. If we subtract l_{i+1} from l_i and divide by l_0 (the original size of the cohort), then we have an estimate of the probability of dying within each age interval for each county within the United States. A graphic of this measure for all counties within the dataset can be viewed in Figure 2.1. The survival probabilities can also be estimated based on the abridged life table. One minus the cumulative sum of the exact probabilities of dying up to a given age interval represents the estimate of the survival probability within that particular age interval. For example, the survival probability for the age group 5-14 years of age is calculated as follows, $1-(\frac{l_0-l_1}{l_0}+\frac{l_1-l_2}{l_0}+\frac{l_2-l_3}{l_0})$. A graph of these survival probabilities of all counties within the dataset is displayed in Figure 2.1. The final measure of interest is the life expectancies at the lower bound of each age group interval. The calculation for the residual life expectancy is within the Elements of Abridged Life Table section above. The full life expectancy is calculated by adding the final age within each age interval to the residual life expectancy. A graph of the full life expectancies for all counties within the United States is displayed in Figure 2.1.

The classification method which is implemented within this research is known as *partitioning around medoids* (PAM). PAM is a K-medoids algorithm which takes the dissimilarity matrix as its main argument. The K-medoids algorithm is very closely related to the K-means algorithm. K-means initializes various means as the centre of the clusters, whilst K-medoids initializes various data points to be the centres of the cluster. For both K-means and K-medoids, the number of clusters is specified by the user before the algorithm is run. The K-medoids algorithm is outlined below.

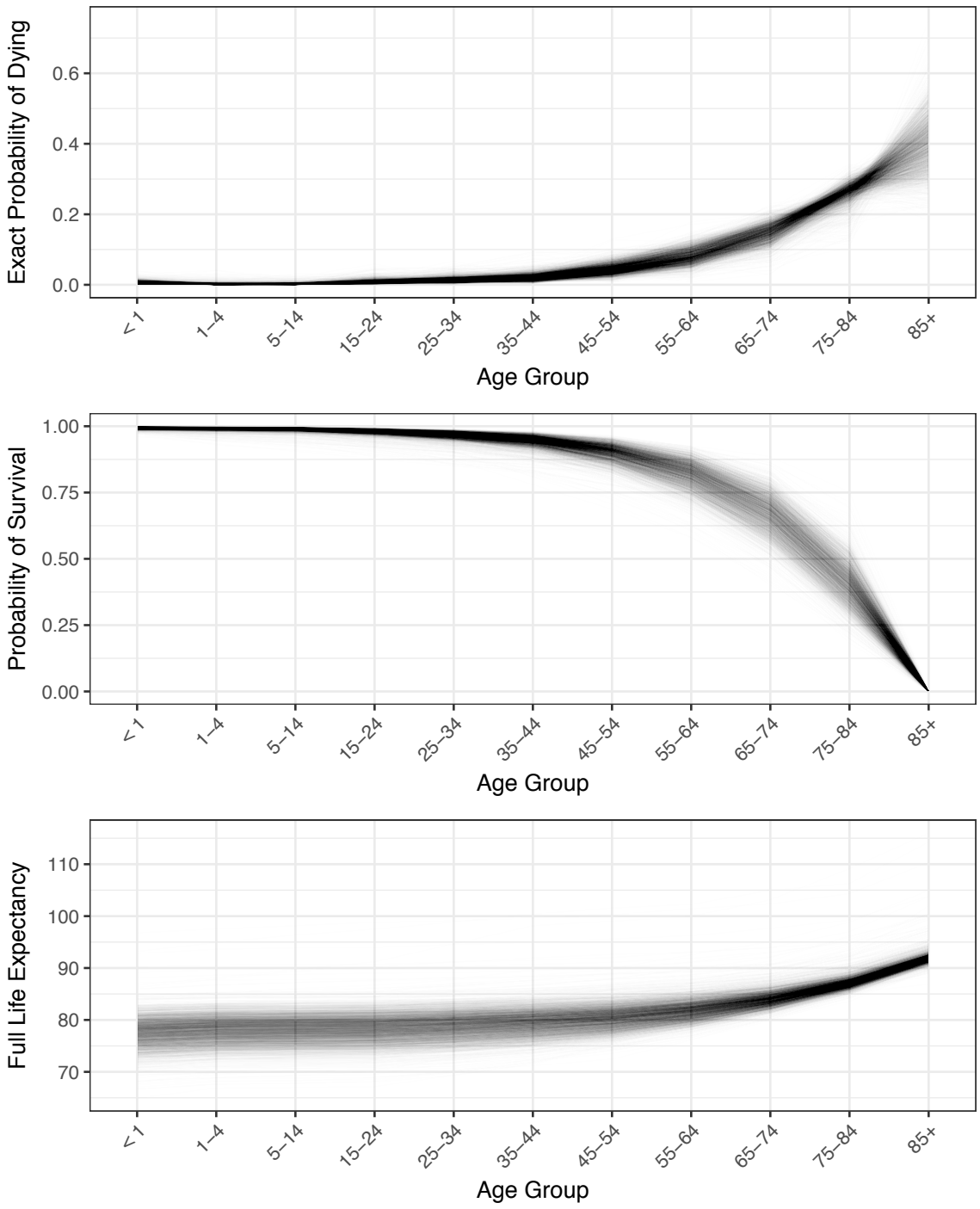

Figure 2.1: Exact Probability of Dying, Survival Probability, and Full Life Expectancy versus Age Interval (for all counties within the dataset)

1. Initialize: set k of the n data points as the medoids.

2. Associate each data point to the closest medoid using an arbitrary metric of distance.

3. While the cost (defined in step (i)) of the configuration decreases:

   (a) For each medoid M, and for each non-medoid data point O:

      (i) Swap M and O, recompute the total cost which is the sum of the distances of points to their medoid

      (ii) If the total cost of the configuration increases, then undo the swap of M and O.

Within R, PAM may be implemented with the function *pam*. The original dataset or the symmetric dissimilarity matrix can be passed to the function. Both Euclidean and Hellinger symmetric dissimilarity matrices are implemented for the purposes of this research in order to classify various life table elements.

In both K-means and K-medoids algorithms, the number of clusters (means or medoids) is specified by the user. In the K-medoids algorithm, a useful tool for determining the number of medoids is referred to as the *silhouette*. The silhouette is a measure of how similar an object is to its own cluster compared to other clusters. The silhouette is defined as the following,

$$s(i) = \begin{cases} 1 - \frac{a(i)}{b(i)} & \text{if } a(i) < b(i) \\ 0 & \text{if } a(i) = b(i) \\ \frac{b(i)}{a(i)} - 1 \text{if } a(i) > b(i) \ , \end{cases}$$

where $a(i)$ represents the average dissimilarity of i^{th} observation with all other observations within the same cluster and $b(i)$ represents the lowest average dissimilarity of the i^{th} observation to any other cluster of which the i^{th} observation is not a member. The silhouette value ranges from -1 to 1, in which -1 represents a bad cluster and 1 accounts for a good cluster. The average silhouette, calculated as $\frac{\sum_{i=1}^{N} s(i)}{N}$ is a measure of how appropriately

the data have been clustered. The closer the average silhouette is to one, then the more appropriately the data have been clustered.

The average silhouette is considered within this particular research. However, a small number of clusters may not be of practical use. It is of greater importance to identify several distinct clusterings of the various life table elements amongst the counties.

CHAPTER 3

Results

It is clear that there are differences between the various counties within the United States with regards to their various life table elements. As displayed in Figure 2.1, we see there are distinguishable patterns in the probabilities of dying, survival probabilities, and the full life expectancies across the various age groups. To test the Metric MDS algorithm, and the PAM algorithm, in identifying patterns, the life tables of two counties within the dataset were chosen as the center of two hypothetical clusters. Figure 3.1 displays the noticeable differences between the probability of dying, survival probability, full life expectancy, and residual life expectancy for Santa Clara County in California and Davidson County in Tennessee.

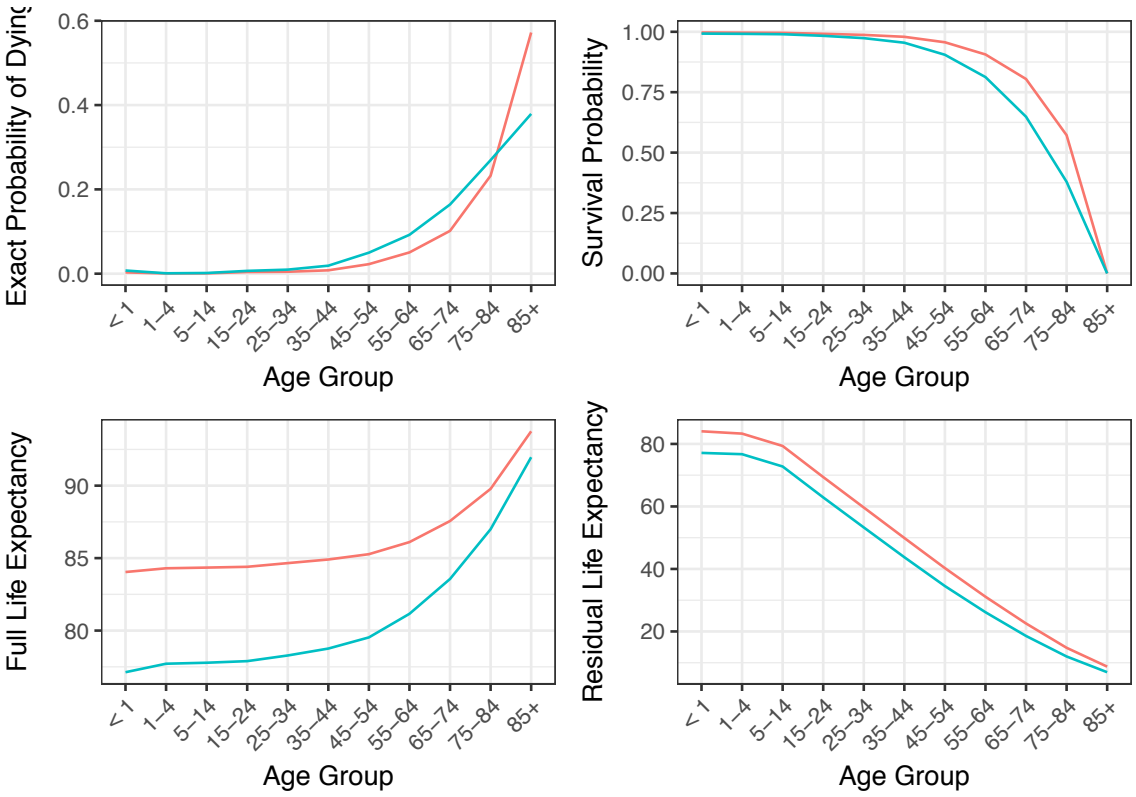

Figure 3.1: Life Table Elements for Santa Clara County (Red) and Davidson County (Blue)

To test if the MDS and PAM algorithms can distinguish between clusters around these two counties, the county values were "jigged". One hundred replicates of each measure of interest were "jigged" by taking the original vector of a county's values, say of the exact probabilities of dying across age intervals, and multiplying it by a vector of exponentiated random uniform variates. The result of the jigging process is a data set consisting of the original measure values and the newly jigged values for one hundred hypothetical counties based on the original two selected counties. The half-width of the interval for the random uniform variates was varied with three different values: 0.25, 0.75, and 5. This resulted in three separate jigged datasets based on the current chosen value of the half-width of the uniform distribution interval. The smaller the value of the half-width of the interval, then the more similar the jigged values remained to the original counties measures.

The new jigged datasets were implemented to see if both the MDS and PAM algorithms can discover the two very distinct groups or clusters. Two separate dissimilarity matrices were obtained for each of the three jigged datasets. For the entries of the first dissimilarity matrix, the Euclidean distances were computed between all pairs of counties. For the second dissimilarity matrix, the Hellinger distances were utilized. The Euclidean distances are calculated as follows.

$$\text{Euclidean } \Delta = \begin{pmatrix} d_{1,1} & d_{1,2} & d_{1,3} & \dots & d_{1,I} \\ d_{2,1} & d_{2,2} & d_{2,3} & \dots & d_{2,I} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ d_{I,1} & d_{I,2} & d_{I,3} & \dots & d_{I,I} \end{pmatrix}$$

An example of calculating the entries is as follows,

$$\begin{aligned} d_{1,2} &= \sqrt{(q_{1,1}-q_{2,1})^2+(q_{1,2}-q_{2,2})^2+...+(q_{1,11}-q_{2,11})^2} \\ &= \sqrt{\sum_{i=1}^{n}(q_{1,i}-q_{2,i})^2}\,, \end{aligned}$$

where $q_{k,i}$ represents the probabilities of dying over the i^{th} age interval in the k^{th} county.

The Hellinger distance may be used to quantify the similarity between two discrete probability distributions. The Hellinger distances are calculated for the probabilities of dying within each age interval as follows.

$$\text{Hellinger } \Delta = \begin{pmatrix} d_{1,1} & d_{1,2} & d_{1,3} & \dots & d_{1,I} \\ d_{2,1} & d_{2,2} & d_{2,3} & \dots & d_{2,I} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ d_{I,1} & d_{I,2} & d_{I,3} & \dots & d_{I,I} \end{pmatrix}$$

An example of calculating the entries is as follows,

$$d_{1,2} = \frac{1}{\sqrt{2}}\sqrt{\sum_{i=1}^{n}(\sqrt{q_{1,i}}-\sqrt{q_{2,i}})^2}\,,$$

where $q_{k,i}$ represents the probabilities of dying over the i^{th} age interval in the k^{th} county. The two dissimilarity matrices for both the Euclidean distances and the Hellinger distances were used separately for the MDS algorithm. The results of the two dissimilarities for all three jigged datasets are displayed in Figure 3.2.

We can see from Figure 3.2 in the first column of the grid plot that there is a clear separation, as shown by applying MDS in two dimensions, between the two clusters of counties, which holds true for both the Euclidean distances and Hellinger distances. We expected the separation in both cases of the dissimilarity measures to be clear due to the choice of

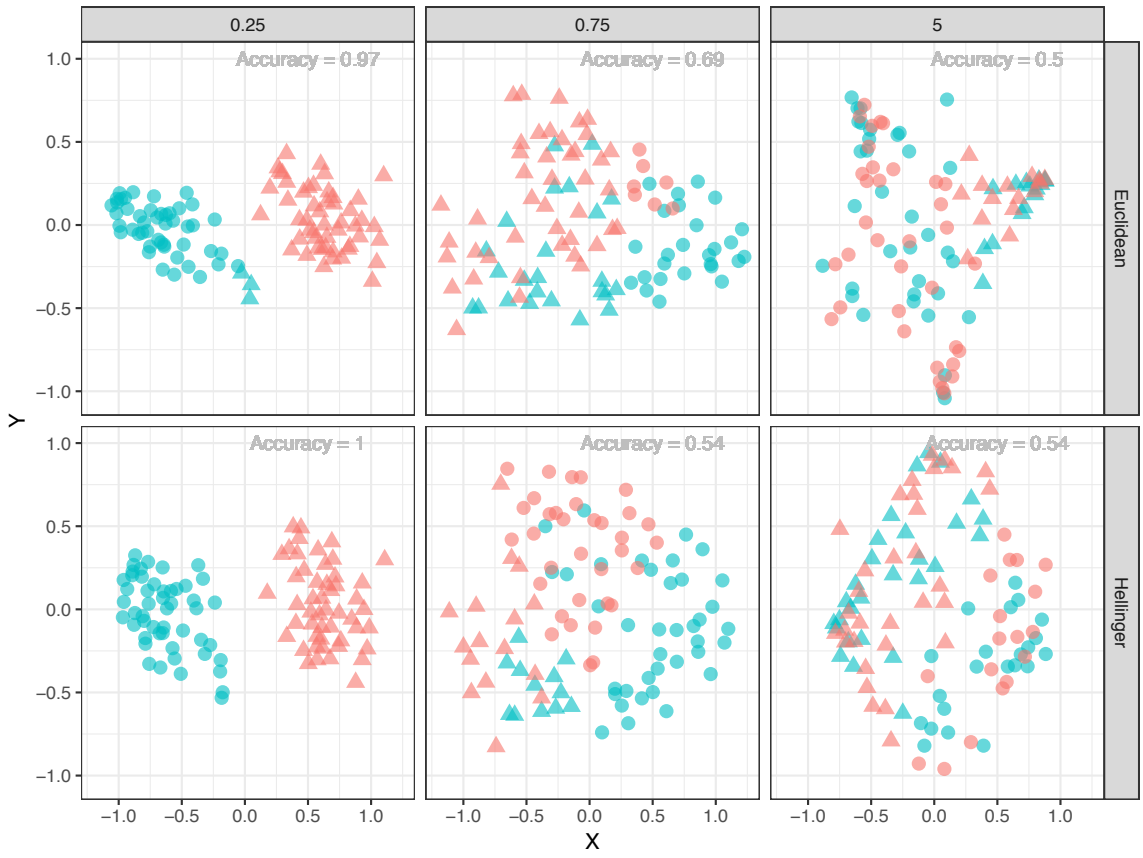

Figure 3.2: Results of MDS algorithm on the three jigged datasets based on Santa Clara County, CA (Blue) and Davidson County, TN (Red). The shapes of the points are based on the classification of the PAM algorithm.

the width of the uniform distribution, 0.5, which was used in the jigging process. We can notice that the separation becomes less clear as the width of the uniform distribution is increased to 0.75 and 5. A single point on the either of the graphs (Euclidean or Hellinger Metric MDS) represents a hypothetical county. Each point can be described as the MDS algorithms attempt to position the points into a two-dimensional plane so that the pairwise distances are preserved as well as possible. The PAM algorithm was then applied each of the three jigged datasets twice, once with Euclidean distances and again with the Hellinger distances. Figure 3.2 also displays the results of the PAM algorithm.

Figure 3.2 details the results of the PAM algorithm for the jigged datasets of the counties Santa Clara and Davidson. The shapes of the points in each of the plots represent the classi-

fication that was assigned by the PAM algorithm. We notice that as the width of the uniform distribution is increased, the accuracy of the PAM algorithm decreases. We would expect the accuracy to decrease because we are forcing the two clusters of counties to be less distinguishable. The results of both the MDS and PAM algorithms, allow us to be confident that the algorithms work on discriminating counties with very different life table elements with both the Euclidean distances and Hellinger distances of certain life table elements.

Seeing as the algorithms worked correctly on the jigged datasets containing just two very distinct counties; the same algorithms were run on a set of counties from the original dataset which was considered "stable" counties. Stable is defined, by the CDC in particular, as a county that has at least twenty deaths within each age group of its abridged life table. The determination of twenty deaths is derived from the relative standard error, which the CDC claims should be no more than 23%. The relative standard error is defined as the standard error of the estimate divided by the estimate and the result is multiplied by 100 so that is is a percentage. If our estimate is the number of deaths within a single age group, denoted D, in which D has a binomial distribution with N trials, then the relative standard error is defined as,

$$RSE(D) = \frac{\sqrt{N\frac{D}{N}(1-\frac{D}{N})}}{D} * 100$$

This equation is equivalent to the following,

$$RSE(D) = \sqrt{\frac{N-D}{ND}} * 100$$

If N is large enough for $\frac{1}{N}$ to be negligible, then the relative standard error may be approx-

imated as,

$$RSE(D) \approx \sqrt{\frac{1}{D}} * 100$$

Solving for D leads to the following,

$$D \approx \left(\frac{100}{RSE}\right)^2$$

Note that when the relative standard error is equal to 23%, this requires that D be about 18.9, which is rounded to 20 in the case of the rule employed by the CDC. When this number of deaths is used to identify the number of stable counties within the original dataset, the sample size is reduced from 3139 counties to only 210 counties. The was considered too restrictive. To increase the number of stable counties within the dataset, a different criterion was employed to define "stable" counties. A rule was determined by investigating the various properties of the sample size, the number of deaths within each age interval, and the standard error of the residual life expectancy. The properties are displayed in Figure 3.3.

The rule for identifying the number of stable counties was for a stable county to have at least one death within each age group and to have a mean standard error for the residual life expectancy less than or equal to 0.5. This increased the subsample size to 1548 counties, from which a random sample of one-fourth of this number of counties was taken. The MDS algorithm is not displayed in this document to view the random sample of stable counties because the separation was not as obvious as the jigged dataset MDS graphic. This is due to a large number of unique counties within the stable counties dataset. The number of dimensions can be increased within the MDS algorithm to begin to see the separation between the "stable" counties. However, there is still not a clear separation within the MDS graph. The PAM algorithm was first applied using pair-wise Euclidean distances between the rows of the stable counties dataset, which were the entries of the dissimilarity matrix.

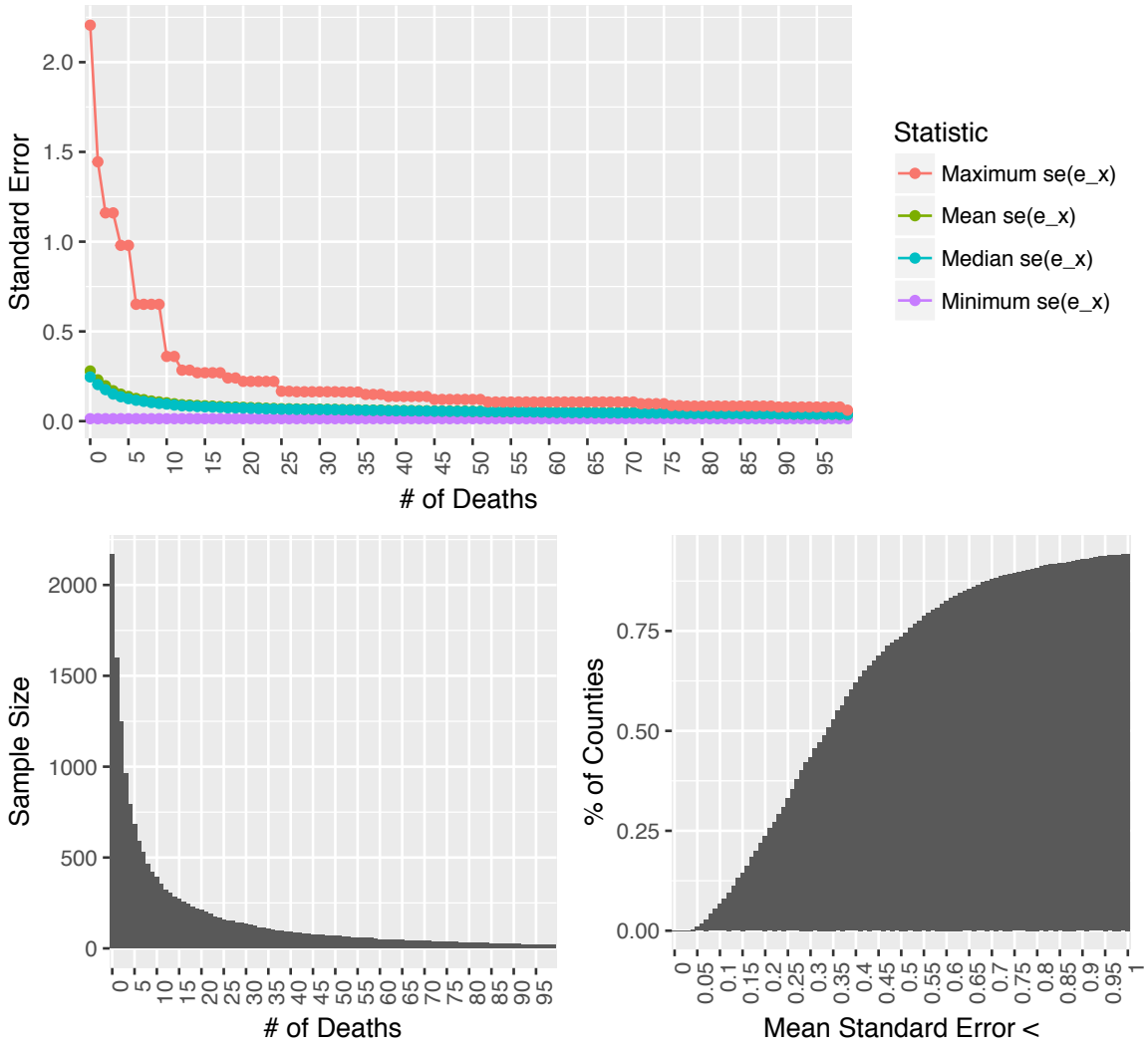

Figure 3.3: Investigating sample size, number of deaths within age intervals, and standard error of residual life expectancy

The probability of dying, the full life expectancy, the survival probability, and finally the residual life expectancy were utilized for the Euclidean distances dissimilarity matrix. Figure 3.4 displays the results of the PAM algorithm for both setting the number of clusters, k, within the PAM algorithm to be two clusters and then three clusters. Each line within each graph represents a separate county's data for the various life table measures. Each graph is created for a different life table measure, in particular, the probability of dying, the survival probabilities, the residual life expectancy, and finally, the residual life expectancy. The x-axis denotes the age intervals. Even though the MDS algorithm did not show a clear

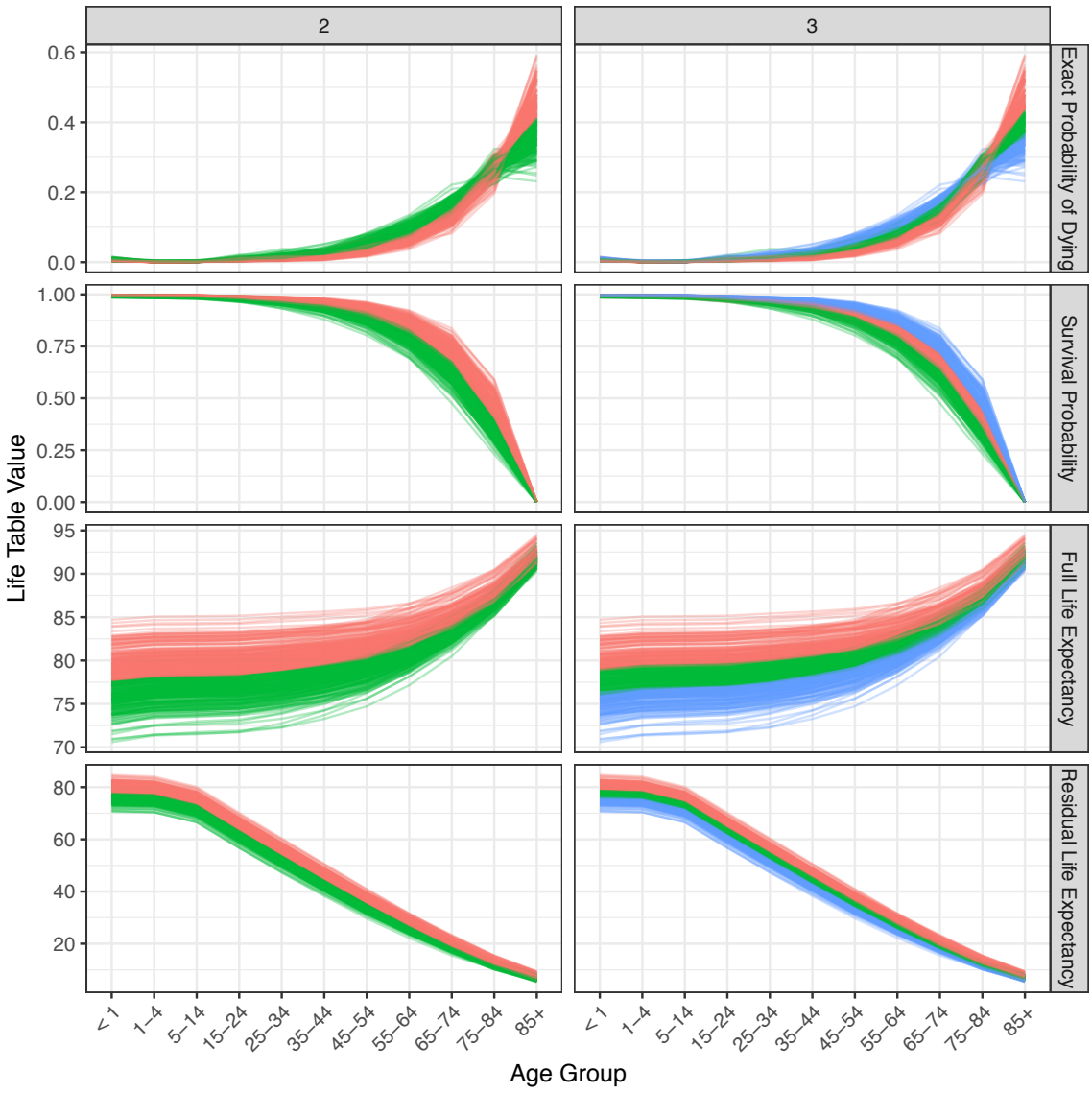

Figure 3.4: Abridged Life Table elements versus Age Groups. Results of PAM Algorithm for Euclidean dissimilarity matrix using random subset of stable counties. Columns of the grid represent the number of clusters specified in PAM.

separation between the counties, we can see from the PAM algorithm that there is a distinct separation for the stable counties based on the various life table elements as displayed by Figure 3.4. It is clear that there is a separation of the various stable counties when looking at the figure, and that these patterns can be characterised by their abridged life table elements.

The pair-wise Hellinger distance was also implemented in the PAM algorithm, however, only the probability of death could be clustered because the Hellinger distance requires the comparison of probability distributions in which the distributions of the probabilities for each county must sum to one. Figure 3.5 displays the results of the PAM algorithm for the pair-wise Hellinger distances. Only one measure was looked at whilst implementing the Hellinger distance, but the number of clusters was varied from two clusters to five clusters. Even as the number of clusters is dramatically increased, we can still see a clear separation of the probabilities of dying for the random sample of stable counties. This distinction is most clear in the final age group due to the amount of variation of the probability of dying within that particular age group. The Hellinger distance for the probabilities of dying also provides a significant distinction between the random sample of stable counties within the dataset.

The average silhouette was examined for both the Euclidean and Hellinger dissimilarity matrices. To see if the model improves, by examining the average silhouette, the number of clusters was varied from two to ten. The average silhouette decreased as the number of clusters increased within the models, for both Euclidean and Hellinger distances. The number of clusters that maximised the average silhouette was two clusters for all models which were examined implementing this particular selection criterion. A model that has clustered the data well should obtain an average silhouette close to one. This was not the case with these particular models. The maximum average silhouette was 0.527, which was obtained by clustering the full life expectancy using the Euclidean dissimilarity matrix. The criterion for selecting the number of clusters, using the average silhouette, was only examined in the case of the random selection of stable counties. Again, as indicated earlier, the main focus of this research is to identify unique patterns of life table elements, such as the distinct patterns of life expectancies displayed within the contrasting counties of the United States. The PAM algorithm outputs the medoids which can be stored and then used on a future dataset to cluster the new counties based on the medoids. Euclidean distance

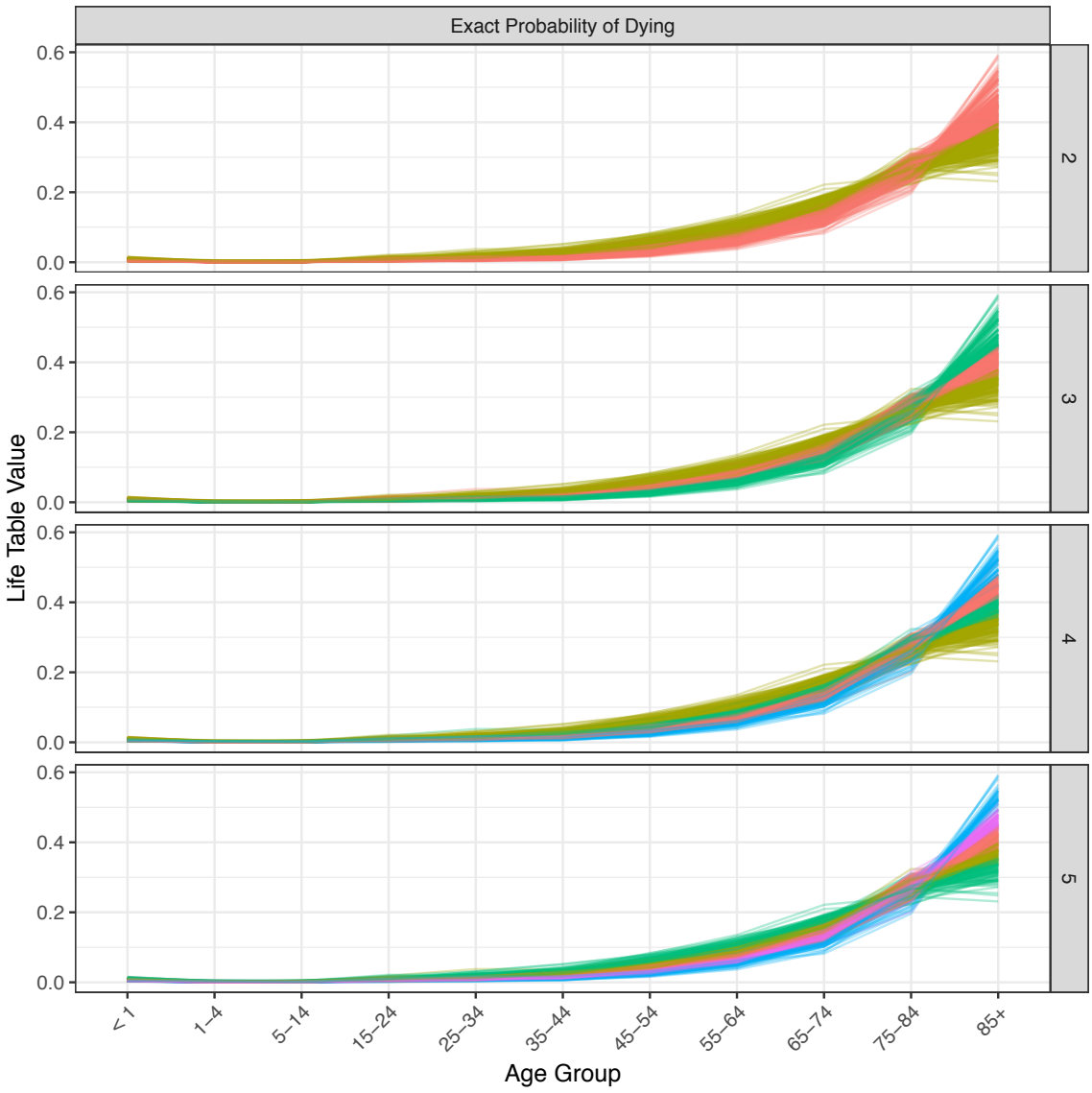

Figure 3.5: Exact Probability of Dying versus Age Groups. Results of PAM Algorithm for Hellinger dissimilarity matrix using random subset of stable counties. Rows of the grid represent the number of clusters specified in PAM.

was implemented, from both trained PAM algorithms (Euclidean and Hellinger) to classify a testing dataset of stable counties which were randomly selected from the rows of data that were not randomly selected for the training of the PAM algorithm. The counties of the testing dataset were classified to the closest medoid from the trained PAM algorithm based on the Euclidean distance. Testing was first carried out using the PAM algorithm which was trained with the Euclidean dissimilarity matrix. The results of the classification of the

testing dataset are shown in Figure 3.6.

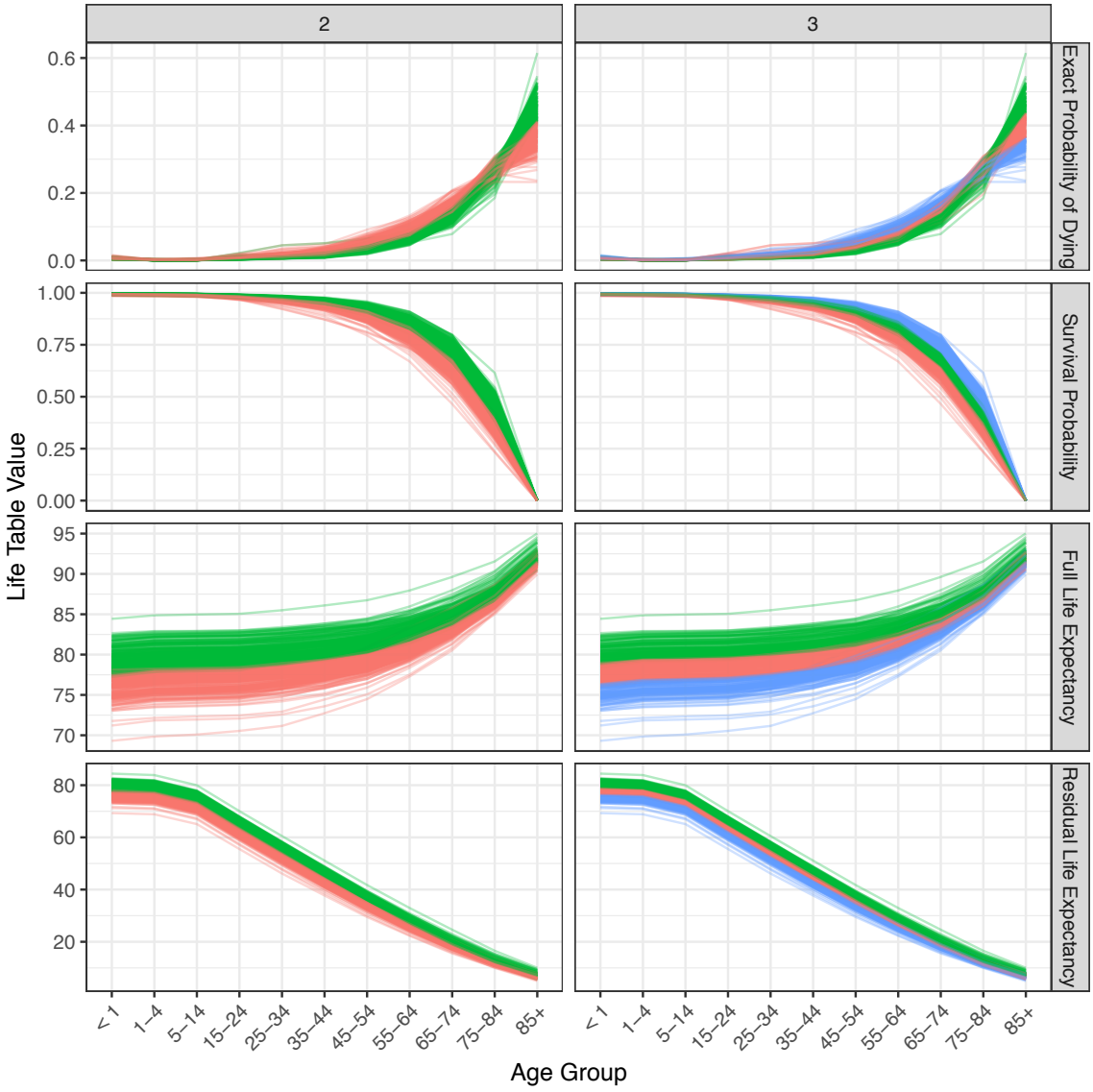

Figure 3.6: Classification of testing dataset using PAM algorithm medoids (Euclidean dissimilarity matrix) from training dataset. Columns of the grid represent the number of clusters specified in PAM.

Clearly, from Figure 3.6, we can see that the testing dataset classification is quite distinct for the stable counties within the various life table elements. Testing was then applied to the PAM algorithm which was trained with the Hellinger dissimilarity matrix. The results of the classification of the testing dataset are shown in Figure 3.7.

The results of Figure 3.7 suggest that the testing dataset classification is also very distinct

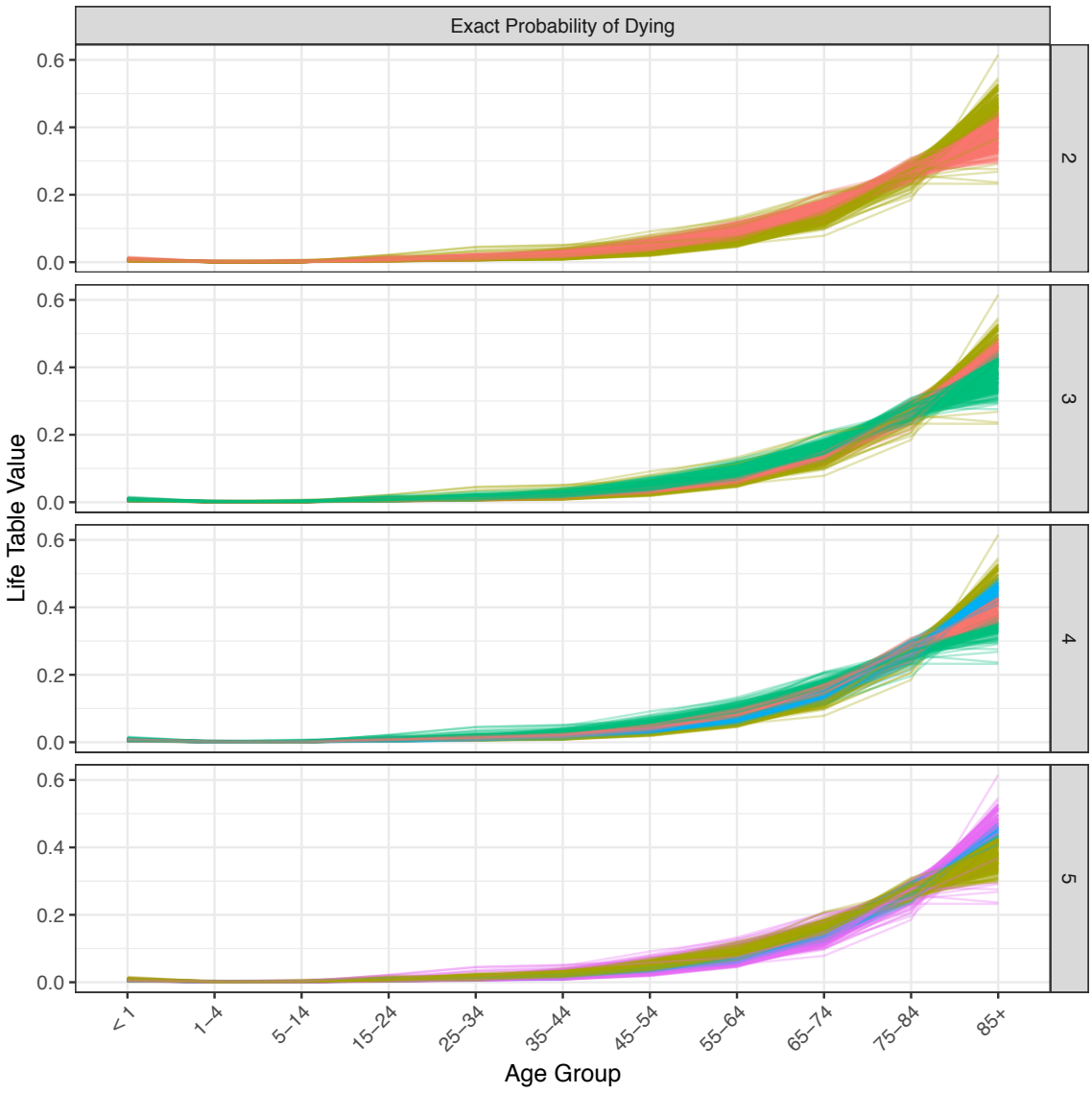

Figure 3.7: Classification of testing dataset using PAM algorithm medoids (Hellinger dissimilarity matrix) from training dataset. Rows of the grid represent the number of clusters specified in PAM.

for the stable counties whilst looking at the exact probability of death. If little information is known about a specific small geographical county, these patterns which are being identified using these particular classification methods could serve as invaluable information in determining the patterns of life table elements that these certain areas may fall into. The main interest of this research is to obtain patterns of life expectancy of the stable counties within the United States whilst implementing the PAM algorithm for classification so that

these patterns can help classify counties which have unstable information.

This research first focused on testing the various methods implemented upon two very distinct counties data. The methods worked well in distinguishing these two chosen counties. A final procedure of the research was to take all stable counties within the United States and assign the stable counties dataset as the training dataset and assign the unstable counties dataset as the testing dataset for the PAM algorithm. Both the Euclidean and Hellinger dissimilarity matrices were utilized. For the Euclidean dissimilarity matrix, the pair-wise Euclidean distances were computed for the full life expectancies of all the counties within the stable dataset. The number of clusters within this particular model was varied from three clusters to six clusters. The results of clustering all the stable counties based on the pair-wise Euclidean distances for the full life expectancy can be viewed in Figure 3.8.
We can see that there is a clear separation of all the stable counties within the United States, based upon the full life expectancies, even as the number of clusters is increased all the way to six clusters. The training of the PAM model for the pair-wise Euclidean distances of the full life expectancies for the stable counties was then applied to the testing dataset, or rather, the dataset containing all of the unstable counties. The results of the clustering of all unstable counties based upon the PAM algorithm which was trained on the stable counties can be viewed in Figure 3.9. The unstable counties were classified to the closest medoid obtained from the trained PAM algorithm on the unstable counties.
It is clear that the testing of the trained PAM algorithm is still separating the full life expectancies for the unstable counties fairly well. The distinction is not quite as clear as when the model is trained, however, there are still clear groups of clusters. One final procedure carried out in this research was to look at how well the Hellinger distance discriminates the exact probabilities of dying for the stable counties within the United States. A PAM model was trained on all of the stable counties, still utilising the definition of stable defined earlier, for the exact probabilities of dying within an age interval. The dissimilarity matrix

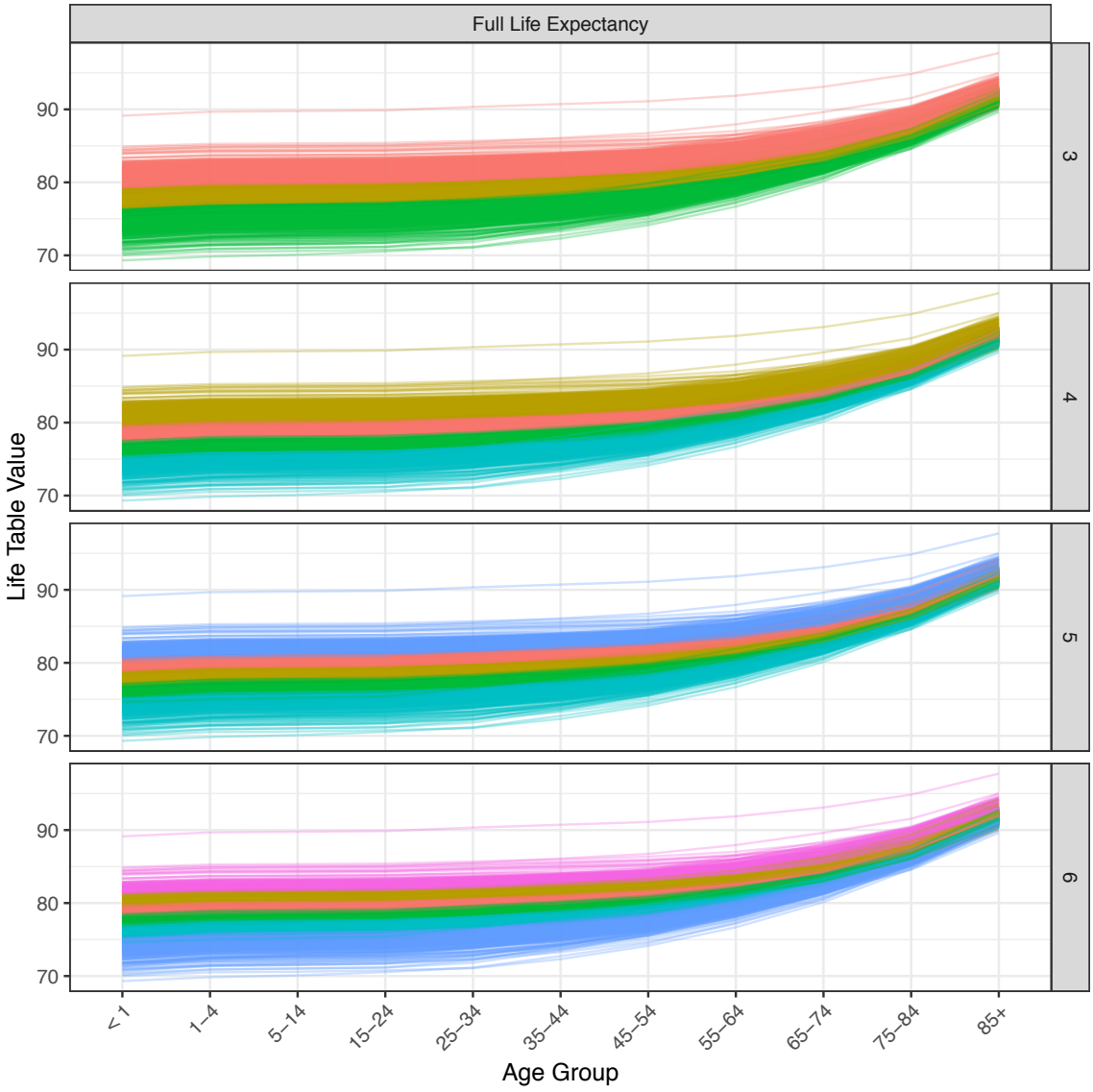

Figure 3.8: Full life expectancy versus Age Groups. Results of PAM Algorithm for Euclidean dissimilarity matrix using all stable counties. Rows of the grid represent the number of clusters specified in PAM.

is now defined as the pair-wise Hellinger distances between all counties' exact probabilities of dying. The results of the training PAM algorithm can be seen in Figure 3.10.

The results of the trained PAM algorithm for the Hellinger distances do not look as clear as the Euclidean distances, nevertheless, we can still see a clear separation of the stable counties' exact probabilities of dying. This particular trained PAM algorithm was then applied to the unstable counties dataset, in which Euclidean distance was implemented

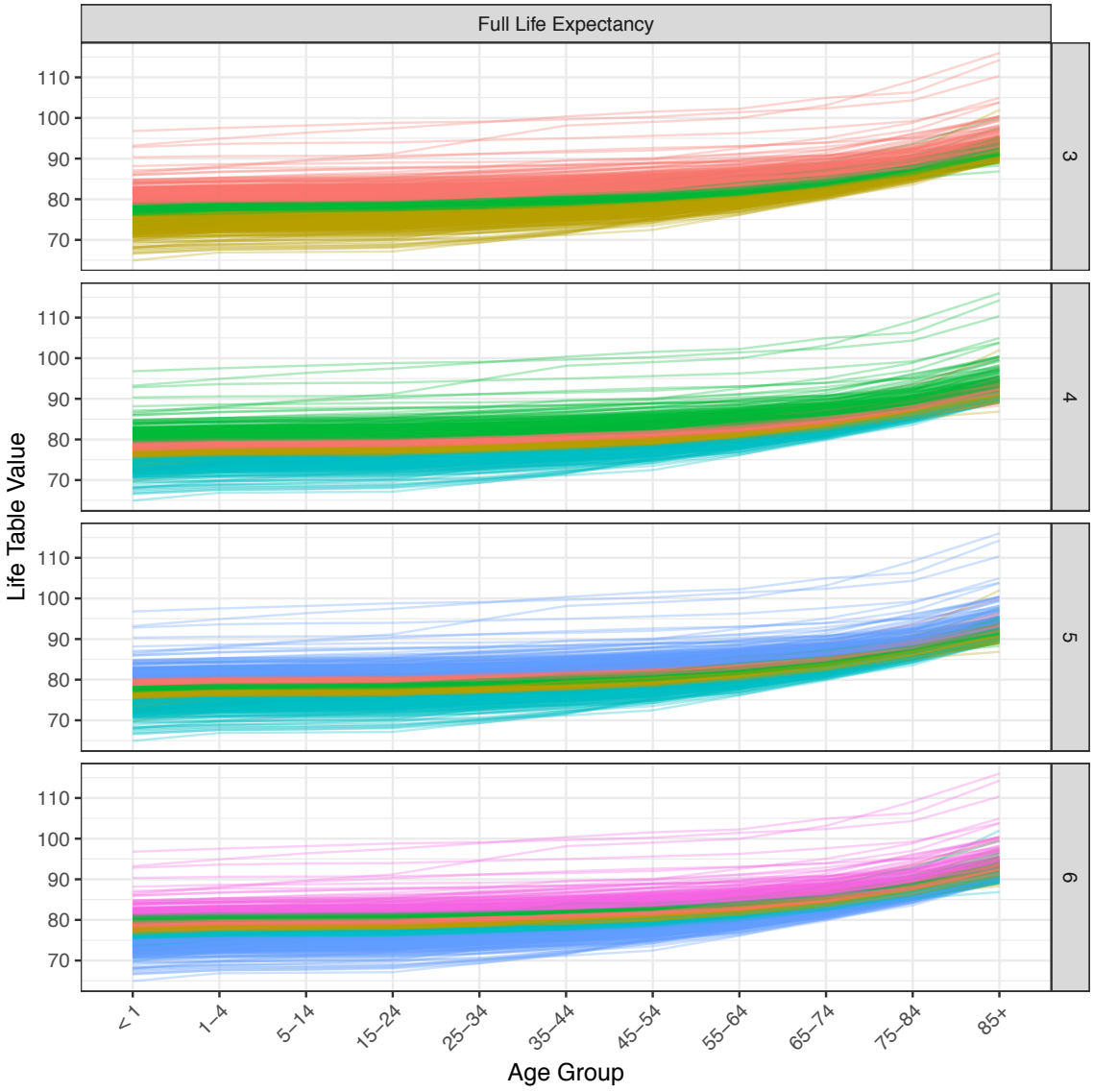

Figure 3.9: Classification of all unstable counties using PAM algorithm medoids (Euclidean dissimilarity matrix) from all stable counties. Rows of the grid represent the number of clusters specified in PAM.

to identify which unstable counties probabilities of dying is closest to the trained PAM medoids. The results of testing are displayed within Figure 3.11, which displays the exact probabilities of dying for all the unstable counties within the United States, and the colours of the lines represent the cluster with which the particular county can be identified.

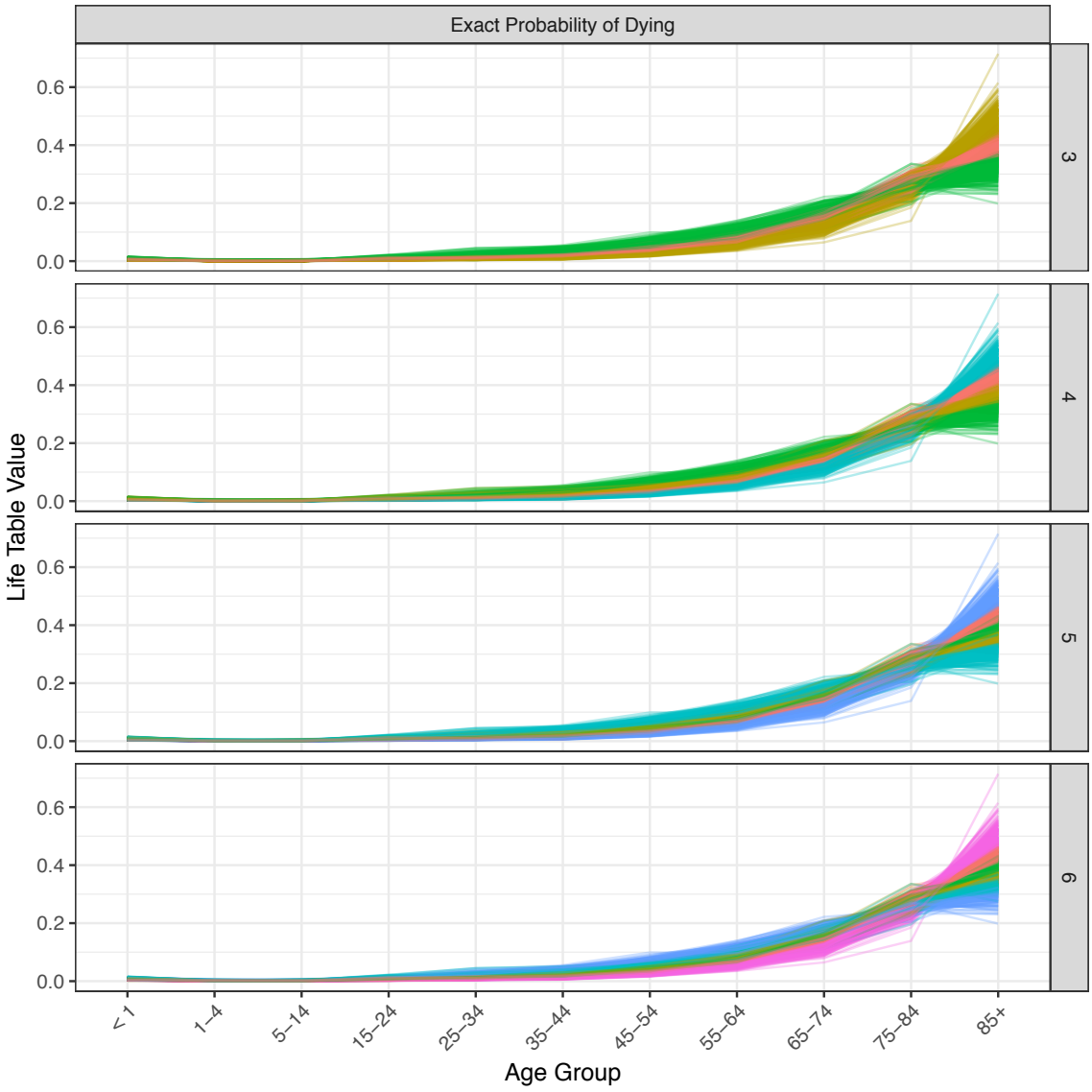

Figure 3.10: Exact Probability of Dying versus Age Groups. Results of PAM Algorithm for Hellinger dissimilarity matrix using all stable counties. Rows of the grid represent the number of clusters specified in PAM.

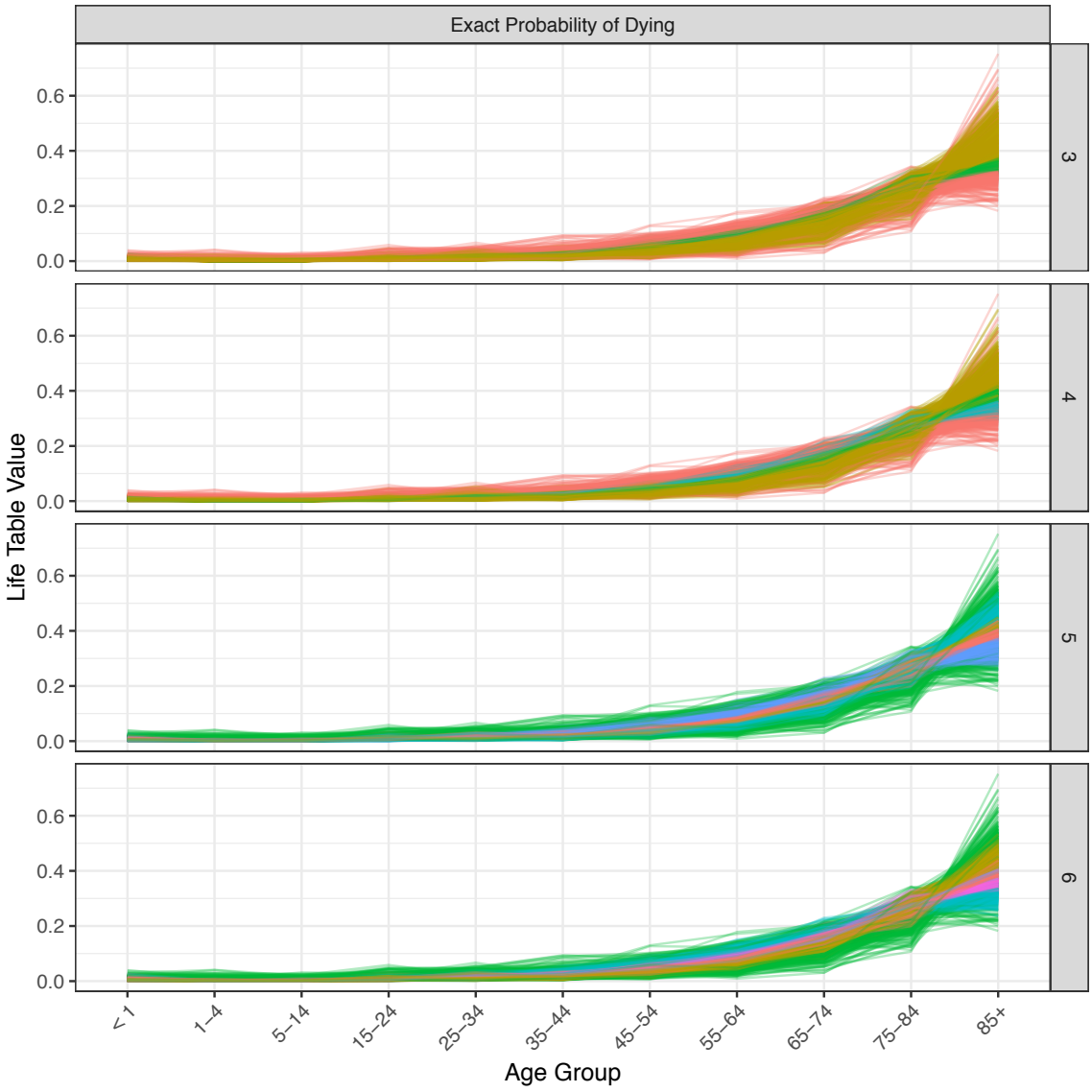

Figure 3.11: Classification of all unstable counties using PAM algorithm medoids (Hellinger dissimilarity matrix) from all stable counties. Rows of the grid represent the number of clusters specified in PAM.

CHAPTER 4

Conclusion

The CDC Wide-ranging ONline Data for Epidemiologic Research (CDC WONDER) makes many health-related datasets available to the public health community with topics amongst the 20 collections of public datasets including United States population estimates, vaccinations, Tuberculosis cases, environmental exposures, births, deaths, and cancer diagnoses. The web application provides public access to data extractions which allow researchers to perform analyses with various statistical methods. County-level national population and mortality data are available in the Multiple Cause of Death data collection. The death counts of small geographical regions tend to be a limited number or even zero, which can cause the conditional probability of death to be zero or infinitesimal. The main concern with these publicly available datasets is that the number of fatalities reported in various counties, especially smaller areas, can be unrepresentative of the area which leads analysis of these regions to become challenging. This particular research aimed at outlining methods to obtain distinct patterns of various life table elements by implementing, in particular, the Partitioning Around Medoids algorithm (PAM). It is evident from the multiple training and testing algorithms detailed in this paper that the PAM algorithm provides a very effective platform for identifying various patterns of mortality probabilities, survival probabilities, and even life expectancies. Future work could use the same methods that obtained these different patterns of life table elements to make predictions of which pattern a small geographical areas life table elements is most closely identifiable. Other ideas include using this predicted pattern for a Bayesian paradigm, in which the predicted pattern could serve as a prior for the model, and other information relating to that particular small geographical region could also be included in the model. Partitioning around medoids is the only classification model which was implemented within this particular research. This

clustering algorithm was particularly inviting due to the direct usage of the calculated dissimilarity matrices, which could utilize both Euclidean and Hellinger distances. Future work could be done to look at other distance measures which could be used within this dissimilarity matrix, or other classification methods could be investigated to ensure that patterns are being identified within new populations' abridged life tables.

REFERENCES

Chiang, Chin Long. Life table and mortality analysis. Geneva: WHO, 1978. Print.

Harsha, Prahladh. "Hellinger Distances." Communication Complexity. N.p., 23 Sept. 2011. Web. *http://www.tcs.tifr.res.in/ prahladh/teaching/2011-12/comm/lectures/l12.pdf*.

Johnson, Robert E. "Small Area Life Expectancy: Computational Challenges, Controversies, and Potential Solutions." N.p., 17 Feb. 2016.

"Multiple Cause of Death, 1999-2015 Request - CDC WONDER." *https://wonder.cdc.gov/mcd-icd10.html*. N.p., n.d. Web. 16 Apr. 2017.