14203480 Brian Buckley

STAT40750 Data Mining Project

The objective of the project is to use the Clark-Sharrow female and male model life tables from the Human Mortality Database to study if the life tables fall into natural clusters with similar mortality profiles.

Initial Analysis of Project data

A plot of the full data and the mean (Figure 1) shows that the average mortality of females and males broadly follows a similar profile. There is an increase of mean mortality for males between the ages of approximately 10 and 20 compared with females of the same age range. This increase in mean male mortality trends back towards the female mortality profile after approximately age 20. The overall range of mortality at a given age across the plot is slightly greater for females but the range for males seems to contain more outliers between the ages of approximately 10 to 40 and 80 to 110.

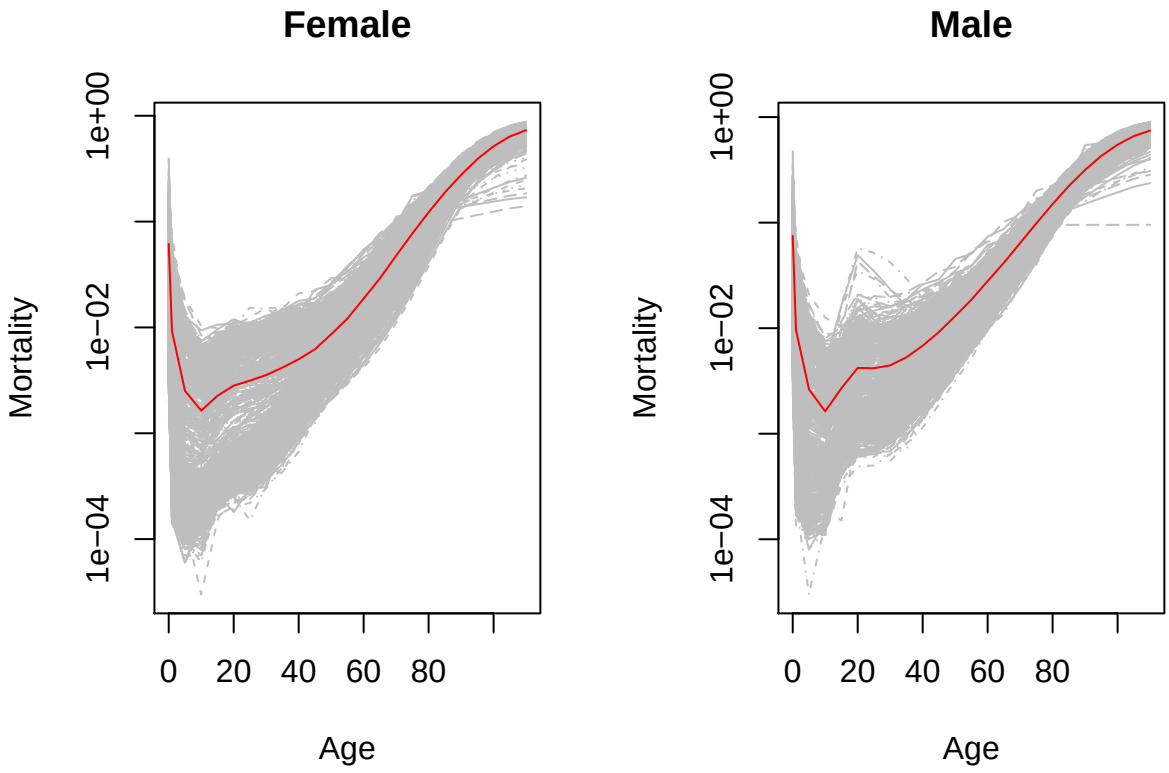

Figure 1: Age against the logarithm of Mortality for female and male life table data. The mean of the data is shown by the red line.

Clustering Analysis

The data were clustered using k-Means clustering to investigate if the life tables fall into groups, where the mortality experience is similar across ages. The objective also included studying the difference between the life tables and clustering results for the female and male tables.

Figure 2 indicates that k=5 seems the best estimate of the elbow so this was chosen. To ensure the k-Means algorithm finds the global minimum total within sum of squares rather than a local minimum the **nstart** argument of the **kmeans** R function was set to try 20 random starting points.

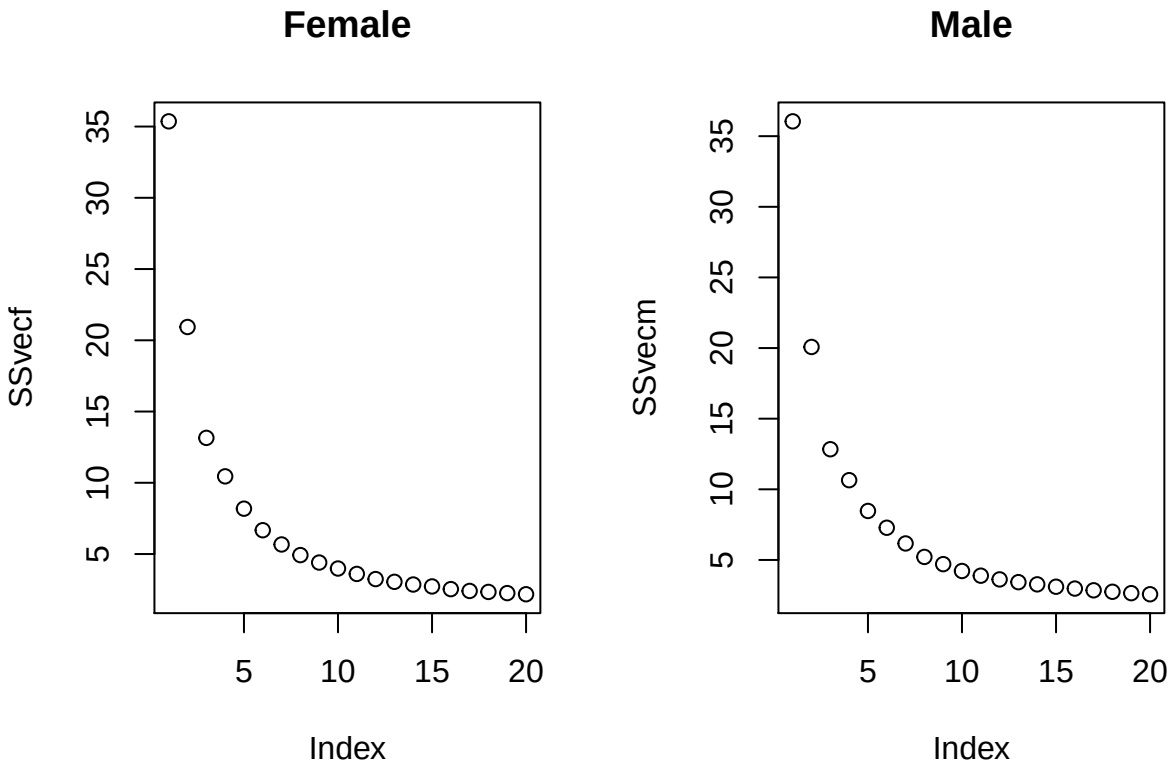

Figure 2: Total within sum of squares for k=1 to 20 for females and males.

Figure 3 shows the female and male cluster centers. Both female and male plots show similar clusters that consist of:

1. Five distinct clusters with the first group having a relatively high mortality level across all age ranges (group 1 in the plot), the second group a lower mortality level across the age ranges, similarly for groups three and four until group 5 which has a relatively low mortality rate across the age ranges.
2. The mortality rate of Group 5 in both female and male plots is the lowest of the cluster groups until approximately age 82 at which stage it trends higher than most of the other groups.

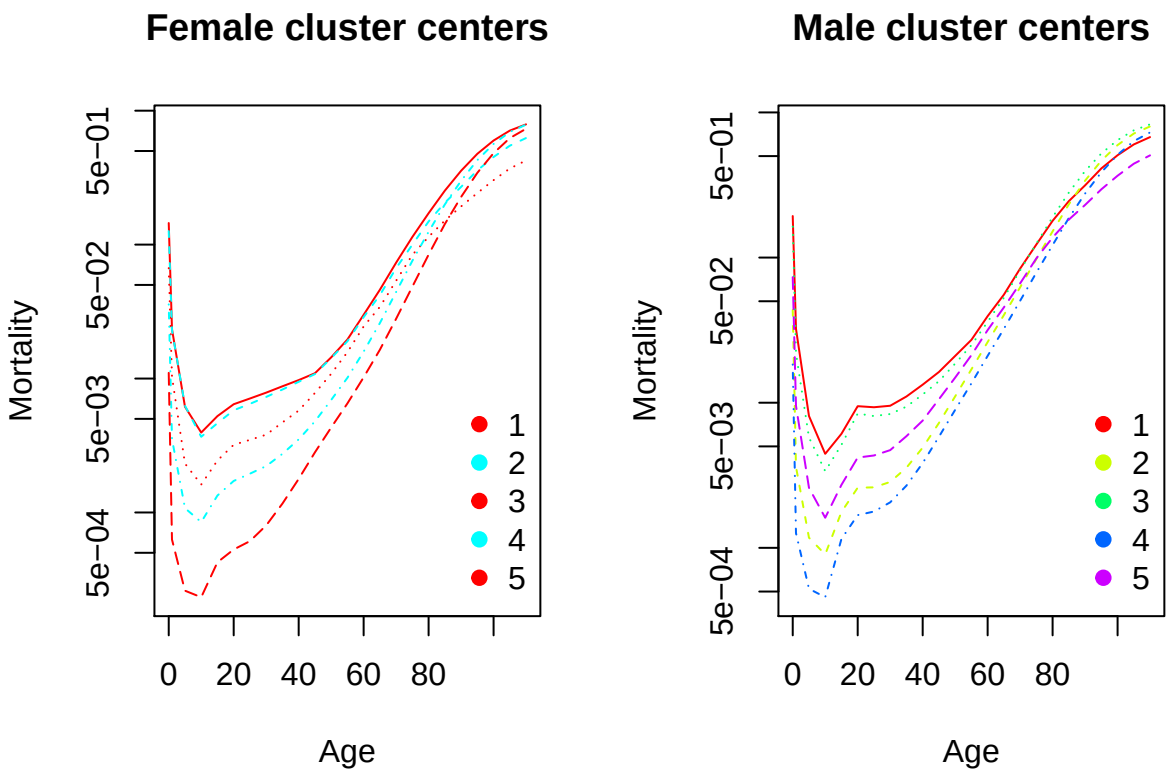

Figure 3: Age against the k-Means cluster centers for females and males. The logarithm of Mortality was used for the y-axis as in Figure 1.

Assessing Clustering Results

The silhouette plot of the data (Figure 4) shows that clusters 4 and 5 in the female data have high average silhouette which indicates strong clustering. Clusters 2 and 3 in the female data have some observations with low silhouette. The male data has high average silhouette for different clusters (2 and 4) and no clusters with silhouette as low as the two in the female data.

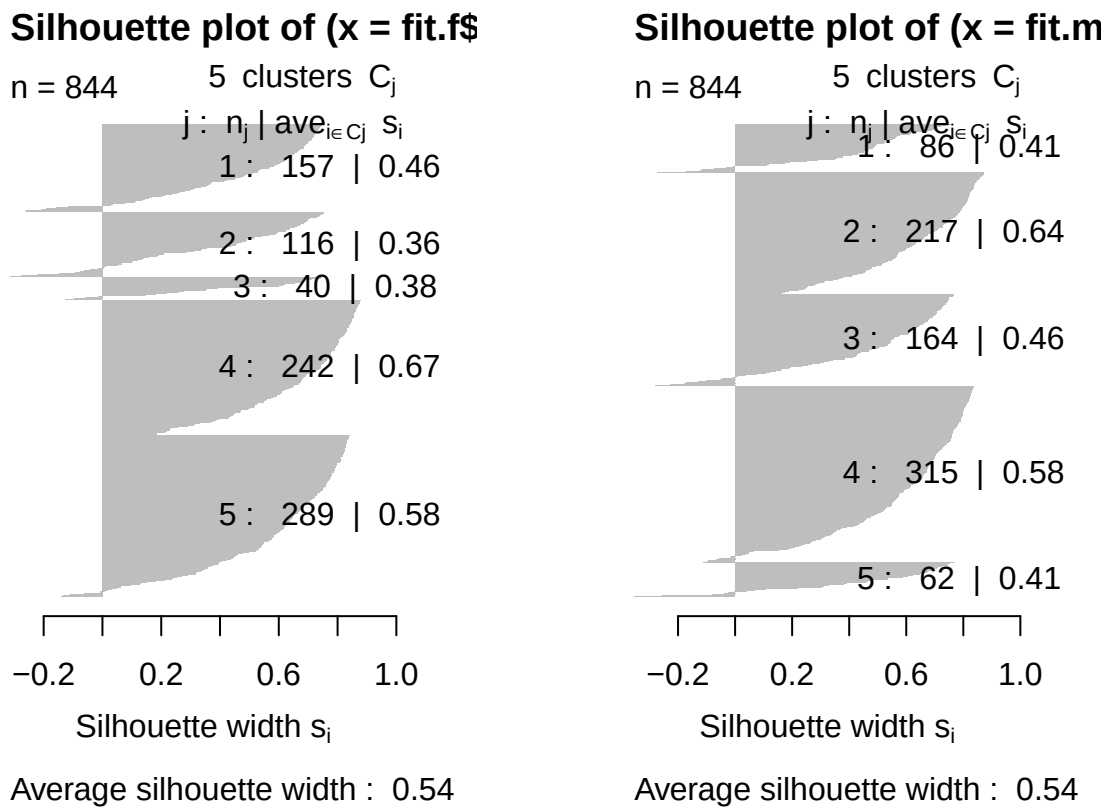

Figure 4: Silhouette plot of the clustering solution.

Interpretation of Clustering Results

A box and whisker plot of the cluster centers (Figure 5) taken against the logarithm of mortality (left-hand plot) and the actual mortality data (right-hand side) highlights the differences between the female and male clustering results. The male cluster data shows higher interquartile range (IQR) and a higher spread of mortality for the mx.0 age group. The female cluster data shows higher IQR and Range for age groups mx.95 to mx.110

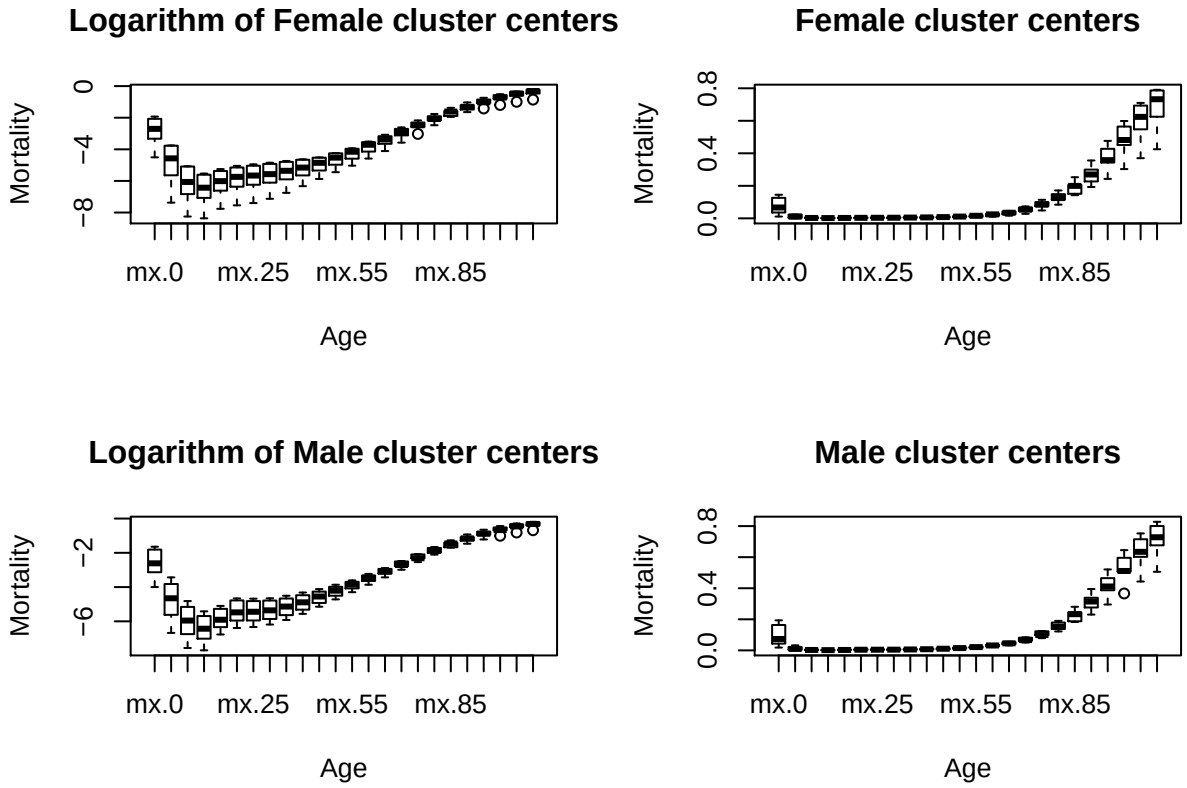

Figure 5: Box and whisker plots of the logarithm of cluster centers and the cluster centers for female and male data.

Conclusions

The clustering shows that the mortality for both females and males fall into natural groups with similar mortality profiles. The female and male groups have notable differences.

Appendix

The R code used in this analysis is shown below.

```
setwd("C:/Users/buckleyb/Documents/Personal/Courses/UCD Data Analytics/STAT40750 Data Mining/Code")
# Dataset: Clark-Sharrow model life tables The data for the project can be
# loaded from the LifeTables package:
library(LifeTables)
# The project data is loaded using the following command:
data(MLTobs)
# We are focussing on the female and the male life tables.  flt.mx.info
# contains the female table data mlt.mx.info contains the male table data
# The first four columns contain information on the table type.  The
# remaining columns give mortality values for different ages (0-110).
# Construct a matrix with the mortality data for each table only.
Yf <- (flt.mx.info[, -(1:4)])
Ym <- (mlt.mx.info[, -(1:4)])
# A matrix plot of the data (with logarithm on the y-axis) helps visualize
# the data x is a vector containing the ages in the life tables
```

```r
x <- c(0, 1, seq(5, 110, by = 5))

par(mfrow = c(1, 2))
Yfa <- colMeans(Yf)
matplot(x, t(Yf), col = "gray", log = "y", type = "l", ylab = "Mortality", main = "Female",
    xlab = "Age")
lines(x, Yfa, pch = 2, col = "red")
Yma <- colMeans(Ym)
matplot(x, t(Ym), col = "gray", log = "y", type = "l", ylab = "Mortality", main = "Male",
    xlab = "Age")
lines(x, Yma, pch = 2, col = "red")

# k-Means algorithm Choose optimum k by plotting the total within sum of
# squares up to k=20
set.seed(1000)
K <- 20
SSvecf <- rep(NA, K)
SSvecm <- rep(NA, K)

for (k in 1:20) {
    SSvecf[k] <- kmeans(Yf, centers = k, nstart = 20)$tot.withinss
    SSvecm[k] <- kmeans(Ym, centers = k, nstart = 20)$tot.withinss
}

par(mfrow = c(1, 2))
plot(SSvecf, main = "Female")
plot(SSvecm, main = "Male")

set.seed(1000)
# Look at the k = 5 results for both female and male
fit.f <- kmeans(Yf, centers = 5, nstart = 20)
fit.m <- kmeans(Ym, centers = 5, nstart = 20)
# fit.f fit.m

# plot the clusters on the data
par(mfrow = c(1, 2))
matplot(x, t(Yf), col = fit.f$cluster, log = "y", type = "l", ylab = "Mortality",
    main = "Female", xlab = "Age")
matplot(x, t(Ym), col = fit.m$cluster, log = "y", type = "l", ylab = "Mortality",
    main = "Male", xlab = "Age")

# plot the cluster centers
par(mfrow = c(1, 2))
matplot(x, t(fit.f$centers), col = rainbow(fit.f$cluster), log = "y", type = "l",
    ylab = "Mortality", main = "Female cluster centers", xlab = "Age")
legend("bottomright", legend = c("1", "2", "3", "4", "5"), pch = c(19), col = rainbow(fit.f$cluster),
    bty = "n")
matplot(x, t(fit.m$centers), col = rainbow(fit.m$cluster), log = "y", type = "l",
    ylab = "Mortality", main = "Male cluster centers", xlab = "Age")
legend("bottomright", legend = c("1", "2", "3", "4", "5"), pch = c(19), col = rainbow(fit.m$cluster),
    bty = "n")

# Inspect the results further using Silouette
```

```r
library(cluster)

df <- dist(Yf)^2
silf <- silhouette(fit.f$cluster, df)
dm <- dist(Ym)^2
silm <- silhouette(fit.m$cluster, dm)

par(mfrow = c(1, 2))
plot(silf)
plot(silm)

# Compare using k-medoids
df1 <- dist(Yf, method = "binary")
fit2.f <- pam(df1, k = 5)
Yf[fit2.f$medoids, ]

dm1 <- dist(Ym, method = "binary")
fit2.m <- pam(dm1, k = 5)
Ym[fit2.m$medoids, ]

# Compare results
par(mfrow = c(1, 2))
table(fit.f$cluster, fit2.f$clustering)
table(fit.m$cluster, fit2.m$clustering)


par(mfrow = c(2, 2))
boxplot(log(fit.f$centers), main = "Logarithm of Female cluster centers")
boxplot(fit.f$centers, main = "Female cluster centers")
boxplot(log(fit.m$centers), main = "Logarithm of Male cluster centers")
boxplot(fit.m$centers, main = "Male cluster centers")
```