Neandertal Cluster Analysis

James Bain & Libby Cowgill

9/15/2017

Data

The data are comprised of 84 different activities compiled from a network of about 100 hospitals across the United States. This data are published in the form of Excel files and made available through the NEISS.

NEISS collects information about product and activity related injuries. Details about the anatomical region, type of injury and other demographic variables are collected. Data about injuries are then transformed to be activity specific and aggregated to match the anatomical categories defined in Berger & Trinkaus (1995). Once body parts injuries are tabulated, the injury patterns are normalized into proportions for each activity. Below is an example:

```
##                        head_neck shoulder_arm      hand      pelvis
## acrobatics/gymnastics 0.01941748     0.526699 0.2079288 0.001618123
##                              leg      foot       trunk
## acrobatics/gymnastics 0.05582524 0.1820388 0.006472492
```

Partitioning Around Medoids (PAM)

Clustering algorithms can be used to uncover patterns that are often too subtle for humans to find. They do so by calculating the distance between observation and "clustering" observations that are close to one another. Partitioning algorithms involve explicitly supplying the number of clusters k. Therefore, we have to find the optimal number of groups k to supply our clustering algorithm.

Partitioning around Medoids (*PAM*) is a partitioning algorithm much like *k-means* but instead of using a centroid as the center point of each cluster, a medoid is used instead. Medoids are attibutes of the data which represent the center-most observation in a cluster of observations, whereas a centroid is an artifially created point used to represent the center of a cluster. Therefore, a centroid is more sensitive to outliers as opposed to medoids.

In order to cluster the points, *PAM* minimizes the following function

$$F(x) = minimize \sum_{i=1}^{n} \sum_{j=1}^{n} d(i,j) z_{ij}$$

$d(i,j)$ is the disimilarity between entities i and j and the term z_{ij} is a variable meant to ensure that only the dissimilarity of items with in the same cluster are computed toward the main function. Observations are then assigned to a cluster given that its presence in that cluster minimizes the main function $F(x)$ (Kaufman and Rousseeuw 1987).

Finding the Optimal Number of Clusters

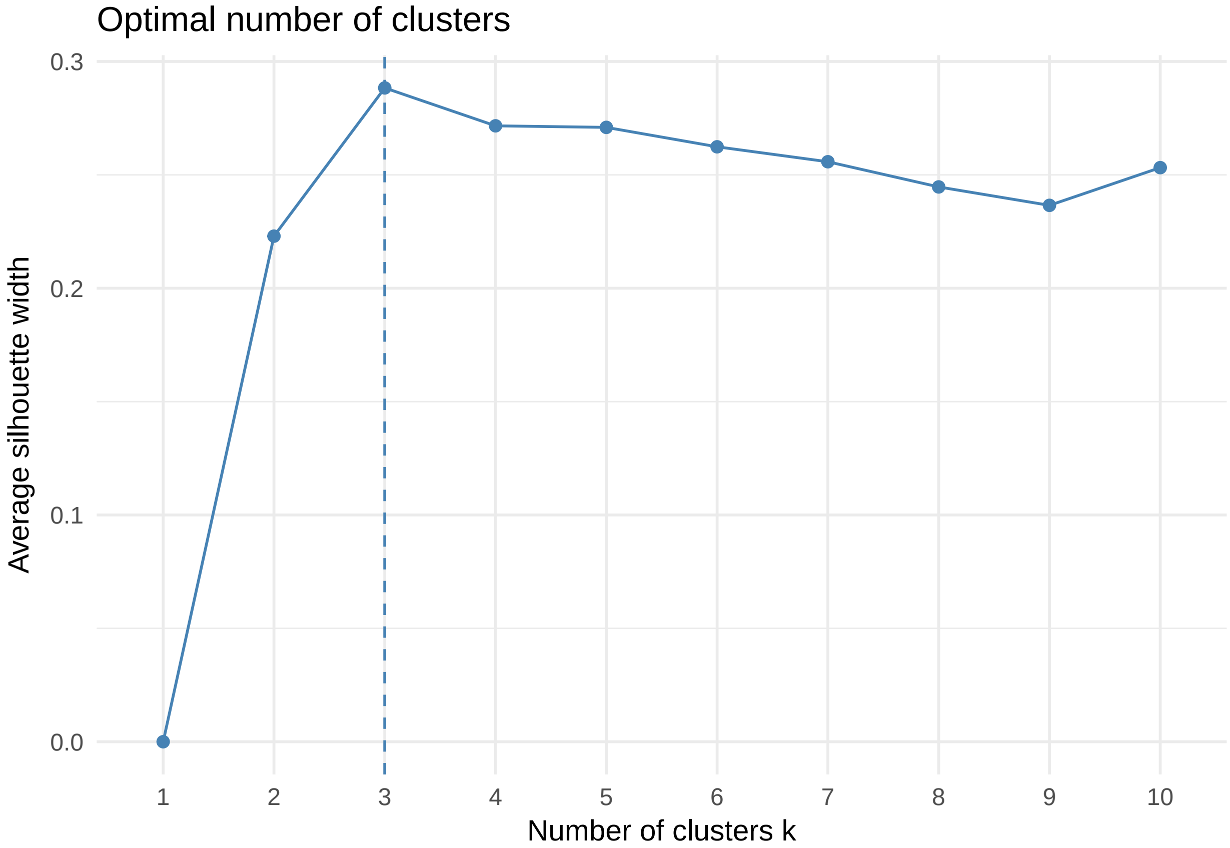
Partitioning clustering methods require that a dismilarity matrix between observations as well as the number of clusters k as inputs to the algorithm. However, it is possible to iterate over a range of ks to calculate the optimal number of clusters. For this, we employ a method known as the *average silhouette* method in which the maximum average silhouette $\bar{s}$ is the optimal number of k clusters for a range of values for k.

$$s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}}$$

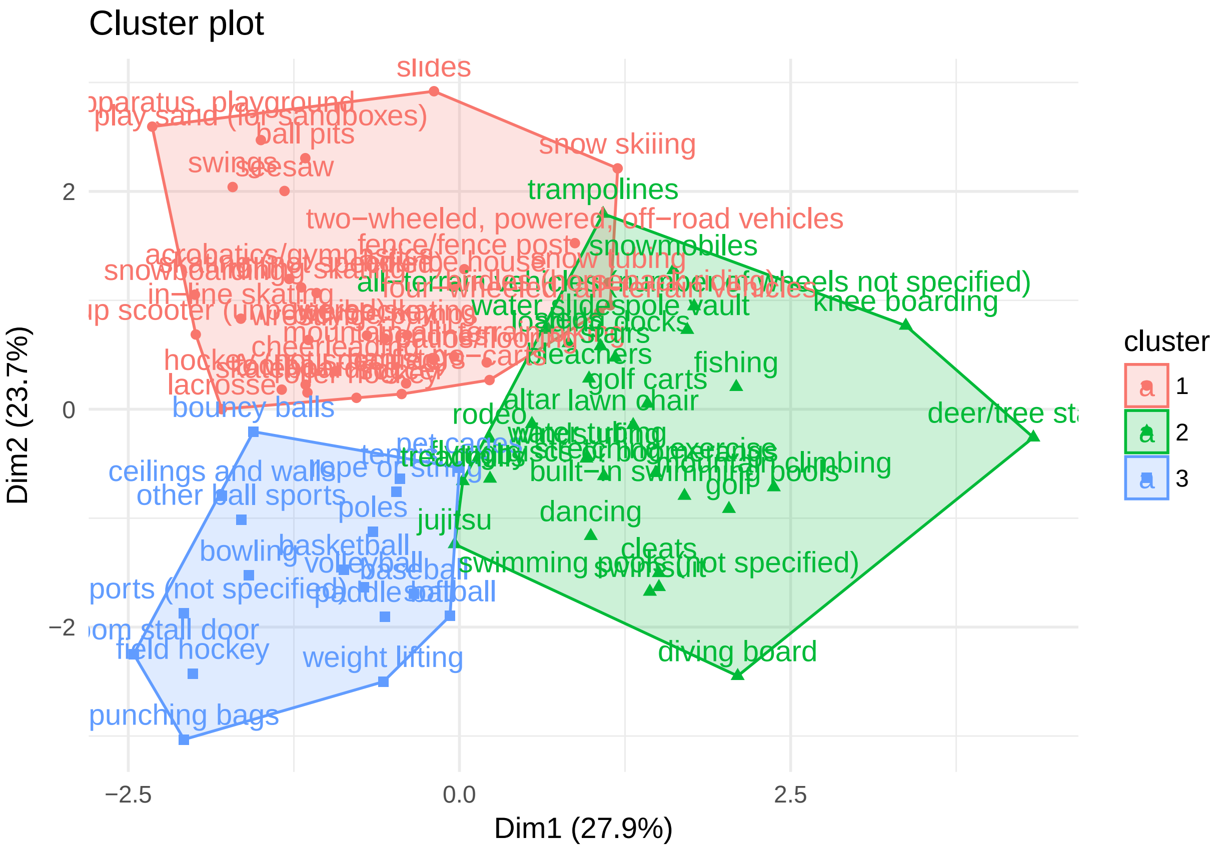
$a(i)$ here is the average dissimilarity between observation i and all other observations within its designated group. $b(i)$ the minimum average dissimalirity of i to the other clusters other than its own. $b(i)$ can thus be thought of as the disimilarity between i and its closest neighboring cluster. $s(i)$ is therefore a value between -1 and 1 with a value closer to 1 meaning that observation is appropriately clustered (Kaufman and Rousseeuw 2009).

Results

Given a dissimilarity matrix caculated using *manhattan distance*, $k = 3$ is the optimal amount number of clusters given the *silhouette method*.



With 3 clusters *PAM* outputs the following groups which can be viewed graphically by plotting them on the two top principal components.



References

Berger, Thomas D, and Erik Trinkaus. 1995. "Patterns of Trauma Among the Neandertals." *Journal of Archaeological Science* 22 (6). Elsevier: 841–52.

Kaufman, Leonard, and Peter Rousseeuw. 1987. *Clustering by Means of Medoids*. North-Holland.

Kaufman, Leonard, and Peter J Rousseeuw. 2009. *Finding Groups in Data: An Introduction to Cluster Analysis*. Vol. 344. John Wiley & Sons.