Differentiating Cellulose Tape for Trace Evidence

Teaching Fellow - Minh Thu Bui
Supervisor - Professor Masanao Yajima
Yingmai Chen, Maysen Pagan, Chang Shi, Yan Wang

1 Project Background

Annamae Joyce is a Graduate student in the Chobanian and Avedisian School of Medicine at Boston University. Annamae is an intern at the Boston Police Department crime lab and is interested in differentiating cellulose tape, or scotch tape, to determine if it can be used as trace evidence.

The sample size of this experiment is 22 tapes. Each roll of tape was divided into 3 sections based on each roll's length. Site 1 refers to the beginning section of each roll, site 2 refers to the middle section, and site 3 refers to the last or end section.

The width and thickness of each tape were measured using a caliber tool. The width measurements of the tapes were measured on one day, with three measurements taken at each site, resulting in a total of nine measurements per roll. The thickness was measured over multiple days, with an emphasis on blind measurements, ensuring the accuracy and integrity of the data. Similar to the width measurements, the thickness measurements also consisted of three measurements at each site, yielding a total of nine measurements per roll.

2 Objectives

The client seeks to determine whether various types of tape can be reliably distinguished from one another based on certain measured features. The primary aim is to investigate the potential use of tape as trace evidence in forensic science. Cellulose tape is fequently found on victims or used by suspects in the creation of explosive devices. The question arises of whether tape can be considered as trace evidence to possibly link individuals to cases.

3 Data Cleaning and Organization

In the data preparation process for four distinct datasets—width.csv, thickness.csv, stereo_phys.csv, and plm.csv—a meticulous approach was employed to ensure cleanliness and organization. The width.csv dataset was loaded and immediately cleaned by removing average and standard deviation columns, excising the term "tape" from the X column, converting the first ten columns to numeric values, and discarding any rows filled with NAs or labeled "tape 7". After renaming the columns to conform to a specific format, the data was reshaped into a longer format for easier manipulation.

Similarly, the thickness.csv dataset underwent an equivalent process of loading, pruning, cleansing, and transforming, with the addition of splitting the data into thickness_ba and thickness_backing subsets. These subsets were then individually reshaped and eventually merged with the width dataset to create a single, more comprehensive dataset named data.

For the stereo_phys.csv file, after the initial steps of loading and cleaning were replicated, the columns were renamed to more descriptive terms, and the physical data it contained was incorporated into the data dataset to further enhance its depth.

The plm.csv file was also subjected to a similar sequence of actions. Notably, the BOPP column was transformed into binary values to better reflect the nature of its data, and like the others, its refined output was merged into the data dataset.

The final step in the data preparation phase was the encoding of categorical variables across several columns into numeric form, a crucial step for subsequent analytical tasks that require numerical input. This thorough conversion from categorical to numeric across the datasets standardizes the data, preparing it for more complex analysis or machine learning applications.

4 Models

The method used to analyze the dissimilarities between each of the 22 tapes involves visualizing a heatmap with hierarchical clusterings of the tapes. These clusterings are determined by the "distances" between each tape.

4.1 Tape Distances

Defining a distance between tapes provides a measure of dissimilarity between each tape. Variables for each of the tapes included three numerical variables (width, thickness of backing, and thickness of backing and adhesive) as well as 12 categorical variables. Not all of the variables are numeric and as a result, the Gower's distance is calculated between two tapes which

accounts for both numerical and categorical variables. Before taking the Gower's distance between each tape, the categorical variables were first converted to numeric values by encoding each level of the variable 1 to the total number of levels of that variable. For example, the categorical variable `Texture` has three levels: "Rough", "Smooth", "Slightly Rough". "Rough" was coded as 1, "Smooth" was coded as 2, and "Slightly Rough" was coded as 3. Each number was then divided by the total number of levels to obtain normalized variables from 0 to 1. In the example, "Rough" now has the numerical value 0.33, "Smooth" now has the numerical value 0.67, and "Slightly Rough" now has the numerical value 1.

Once all categorical variables were coded numerically, Gower's distance calculates a matrix of dissimilarities for each of the $\frac{n(n-1)}{2} = \frac{22(22-1)}{2} = 231$ pairs of tapes. This method combines the Manhattan distance for numerical variables and Hamming loss for categorical variables to get the total distance between two observations. For the three numerical variables, a range-normalized Manhattan distance is calculated:

$$\left| \frac{T_{ik} - T_{jk}}{\text{range}_k} \right|$$

where $i, j = 1, 2, ..., 22$ and $k = 1, 2, 3$. For the remaining 12 categorical variables, the Hamming loss is calculated with the following formula:

$$I(T_{im} \neq T_{jm})$$

where $i, j = 1, 2, ..., 22$ and $m = 4, 5, ..., 15$. This indicator will equal 1 if $T_{im} \neq T_{jm}$ and 0 if $T_{im} \neq T_{jm}$. Gower's distance between two tapes is calculated by summing these variable distances and dividing by 15, the number of variables.

4.2 Multidimensional Scaling

Once we have the pairwise distances between tapes from the data, multidimensional scaling provides one method of visualizing these clusters and their distances from each other on a two dimensional plot. Although the data for each tape is multivariate, multidimensional scaling preserves the distances between pairwise observations when plotting. Multidimensional scaling takes in the calculated distances and returns a set of 22 points where the distances between each point is approximately equal to the dissimilarities between each point. These points are then plotted on a two dimensional scatter plot to visualize the "closeness" of the tapes and can be viewed below.
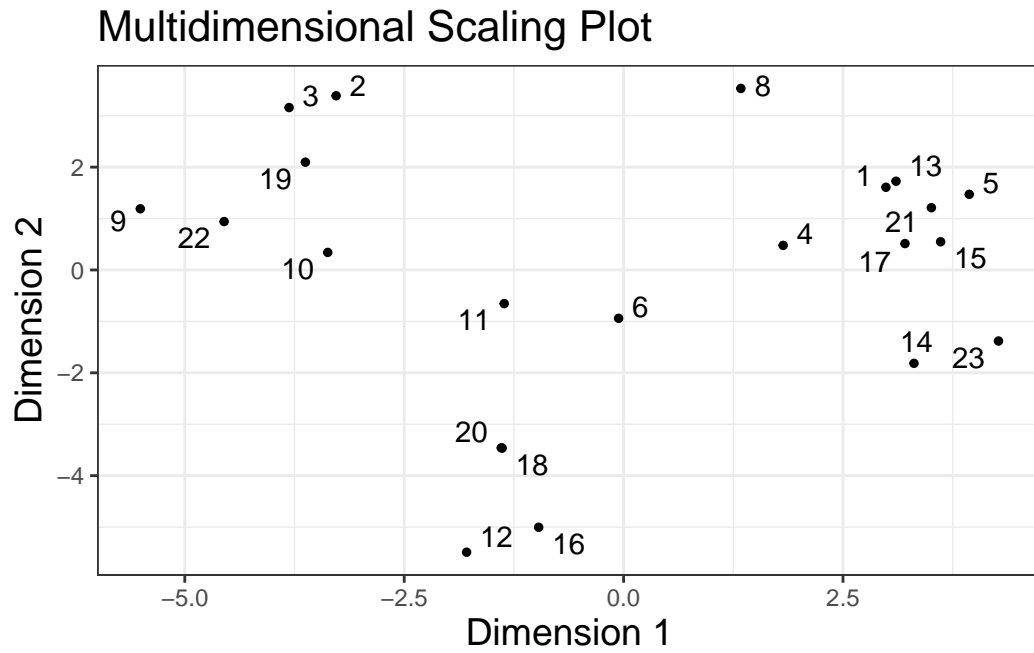
Multidimensional Scaling Plot

Figure 1: Multidimensional Scaling Plot of 22 Different Tapes. Tape 7 is omitted.

4.3 Heatmap

We can also use a heatmap to visually identify similarities and dissimilarities between tapes. In a heatmap cells are color-coded to quickly compare one row or tape to another. The heatmap also allows one to aggregate the rows in clusters based on the distances calculated from Section 4.1. Figure 2 below is the heatmap generated for the 22 tapes using Gower's distance.
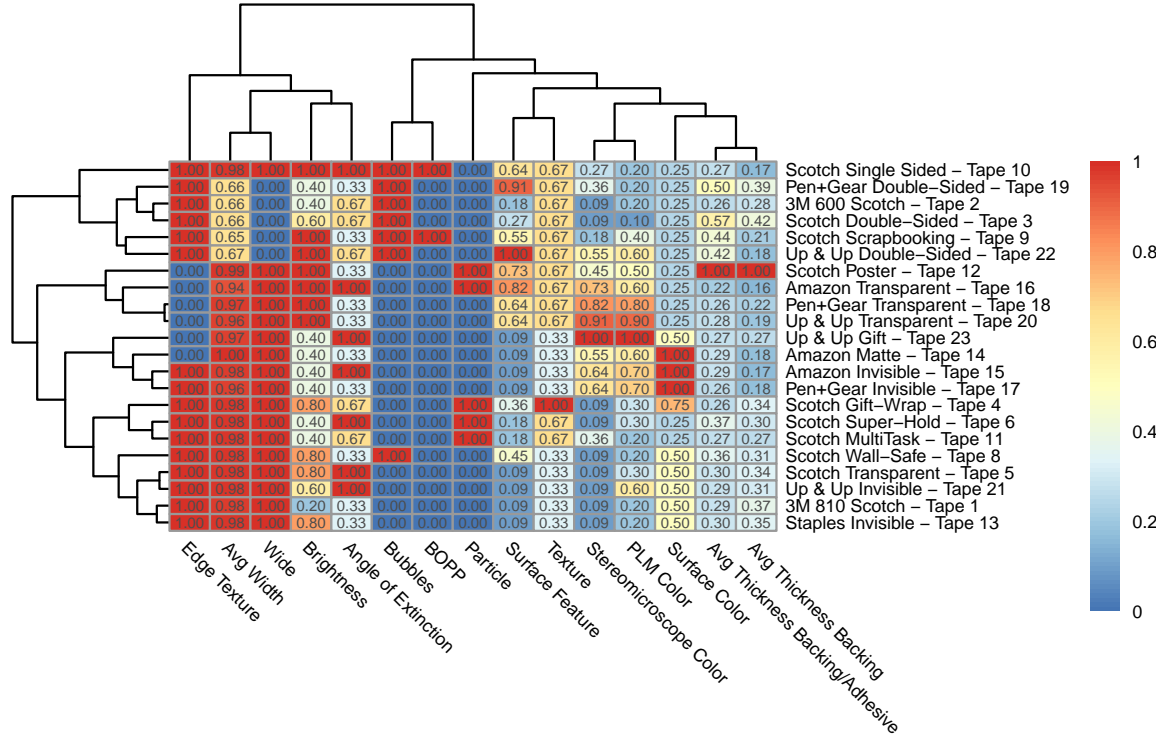

Figure 2: Clustered Heatmap of 22 Different Tapes. Tape 7 is omitted

5 Analysis and Conclusion

From Figure 1, we can begin to visualize some of the groups of tapes. In the top left corner, tapes 2, 3, 9, 10, and 22 are clustered together. Tapes 18 and 20 almost share the exact same point with 12 and 16 directly below them suggesting another cluster of four tapes. Tapes 11 and 6 float in the middle of the plot and may share similarities with tape 10 on the left or tape 4 on the right. The tapes in the top right corner could be divided into two clusters possibly with 14 and 12 in one, and the other tapes in another. From this multidimensional scaling plot, we can view the rudimentary groupings of tapes in about 5 clusters.

According to Figure 2, 22 tapes were divided into 5 clusters. Cluster 1 is composed of tapes 10, 19, 2, 3, 9, and 22. Cluster 2 is composed of tapes 12, 16, 18, and 20. Cluster 3 contains tapes 23, 14, 15, and 17. Cluster 4 contains tapes 4, 6, and 11. Lastly, cluster 5 is composed of tapes 8, 5, 21, 1, and 13.

Cluster 1 features: All tapes in this cluster have a smooth edge texture, are narrow in width, exhibit a clear surface color, possess a smooth texture, contain bubbles, and have no particles. In Surface Feature, each tape is different. For example, tapes has Minimal Surface Features, Large bubble and Irregular, Circular bubbles, Irregular large bubble(water-like), Many Grooves and bubbles, and Grooves(water-like).

Cluster 2 features: All tapes of this cluster are rough, are wide in width, exhibit High Bright in brightness, have a smooth texture, a clear surface color, don't have bubbles, and are categorized as MOPP. In Stereomicroscope Color, each tape is different. For example, tapes has Pink/Green/Yellow/Purple/Blue, Blue/Pink, Pink/Red/Orange, and Purple/Blue.

Cluster 3 features: All tapes of this cluster are wide in width, exhibit Bright in brightness, have a rough surface feature and texture, have no particles, don't have bubbles, and are categorized as MOPP.In PLM Color, tape is different. For example, tapes has Yellow, Light Green, and Yellow/Gold.

Cluster 4 features: All tapes of this cluster have a smooth edge texture, are wide in width, have particles, don't have bubbles, and are categorized as MOPP. In Brightness, tape is different. For example, tapes has Bright and Low Bright.

Cluster 5 features: All tapes of this cluster have a smooth edge texture, are wide in width, mostly exhibit a rough surface feature, have a rough texture, display a Pink color under the stereomicroscope, have a Matte surface color, contain particles, and are categorized as MOPP. In Brightness, tape is different. For example,tapes has Low Dark, Low Bright and Dark.

Therefore, according to the analysis of pheatmap, we can know that pheatmap can effectively distinguish different tapes and find out their similar rules. Each cluster has many of the same variables in its tape. At the same time, each cluster also has a different place, which can distinguish all tapes, it is easy to find the tape number and brand.

6 Appendix
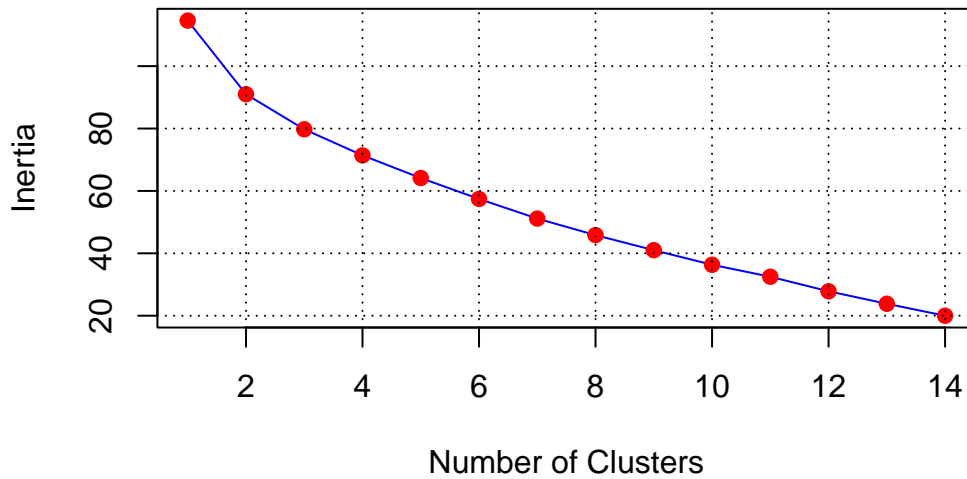
Elbow Method for Optimal Number of Clusters

Figure 3: 5 optimal number of clusters

Figure 3 demonstrates another method of confirming our conclusion of 5 clusters. Using the elbow method in k-means clustering, the optimal number of clusters occurs at the "elbow" of the curve of the line. This appears to occur between 4 and 6 clusters which further supports our previous conclusions. However, this method should be reproduced and analyzed with caution as our sample size is small which may produce unclear and poor conclusions. With larger sample sizes, the elbow method may more accurately determine the optimal number of clusters.