Visualization of Personality and Phobia Type Clustering with GMM and Spectral

Ting Tin Tin[1, *], Cheok Jia Wei[2], Ong Tzi Min[3], Lim Siew Mooi[4],
Lee Kuok Tiung[5], Ali Aitizaz[6], Chaw Jun Kit[7], Ayodeji Olalekan Salau[8]
Faculty of Data Science and Information Technology, INTI International University, Kuala Lumpur, Malaysia[1]
Faculty of Computing and Information Technology,
Tunku Abdul Rahman University of Management and Technology, Kuala Lumpur, Malaysia[2, 3, 4]
Faculty of Social Science and Humanities, Universiti Malaysia Sabah, Sabah, Malaysia[5]
School of Technology, Asia Pacific University, Kuala Lumpur, Malaysia[6]
Institute of Visual Informatics, Universiti Kebangsaan Malaysia, Bangi, Malaysia[7]
Department of Electrical / Electronics and Computer Engineering, Afe Babalola University, Ado-Ekiti, Nigeria[8 (a)]
Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Chennai, Tamil Nadu, India[8 (b)]

Abstract—**Personality traits, the unique characteristics defining individuals, have intrigued philosophers and scholars for centuries. With recent advances in machine learning, there is an opportunity to revolutionize how we understand and differentiate personality traits. This study seeks to develop a robust cluster analysis approach (unsupervised learning) to efficiently and accurately classify individuals based on their personality traits, overcoming the limitations of manual classification. The problem at hand is to create a system that can handle the subjective nature of qualitative personality data, providing insights into how people interact, collaborate, and behave in various social contexts and thus increase learning opportunities. To achieve this, various unsupervised clustering techniques, including spectral clustering and Gaussian mixture models, will be employed to identify similarities in unlabeled data collected through interview questions. The clustering approach is crucial in helping policy makers to identify suitable approaches to improve teamwork efficiency in both educational institutions and job industries.**

Keywords—Unsupervised learning; learning opportunities; clustering; personality; machine learning; Gaussian mixture model; spectral clustering

I. Introduction

Personality traits have been discussed among people and philosophers since the early ages. Personality traits are a set of characteristics and attributes that define the uniqueness of an individual. People have unique personality traits that experts agree are the result of traits mixed with learnt behaviors [1]. Throughout history, fellow philosophers such as Aristotle, Theophrastus, and many others have studied human behaviors and tried to make sense of them. For example, Aristotle used vanity, modesty, and cowardice as factors of moral and immoral behavior [2]. Until now, many efforts have been made to understand how humans function and how their personality and social behaviors correlate with each other using various methods such as psychology testing, surveys and questionnaires, social experiments, computer modeling, and simulations.

With recent advances in Machine Learning technology, there is an opportunity to improve the accuracy of differentiating people with different personality traits. Such systems enable people to understand how others communicate, collaborate, manage stress, and many more. This allows them to have a better grasp of how to resolve conflicts better and communicate more effectively. By creating a personality recognition system, we can create different styles of teaching, coaching, leading, communicating, and many more [3]. For example, where personality traits affect workplaces, by understanding more about our peers, we can build a more cohesive team as we can understand how they work and how we can work with them. A unique personality is like a double-edged sword. Sometimes, it can increase the whole team's synergy and increase efficiency, and sometimes it can be the chisel that tears the whole team apart. It is okay to have people with different personalities working together, but the major concern is the way to communicate with different people to avoid conflicts within the team. Therefore, we must consider personality traits to increase efficiency, synergy, and a more dynamic and cohesive team [4].

Based on previous related work, the researchers had attempted to classify and analyze personality traits manually. This has led to several issues where identifying personality traits requires a large amount of time, making it inefficient to analyze a large group of people [5]. Due to the problem stated above, it is necessary to combine the technology of computer science with the study of personality psychology to increase the efficiency and precision of identifying different personality traits to personalize e-learning [6]. Furthermore, personality traits are difficult to analyze as it is qualitative data that are subjective to individuals rather than quantitative data which represents numerical facts [7], [8], [9]. Today, unsupervised machine learning methods such as clustering are commonly used to group people with similar personality traits into a group of clusters. For example, a study by Feng et al. (2008) used fuzzy clustering analysis based on the common characteristics to create a personalized study strategy [10].

Clustering Analysis is where the computer tries to identify a structure or similar pattern in unlabeled data. Cluster is an important term in cluster analysis, where a cluster is a collection of similar items. There are a few commonly used clustering methods such as Hierarchical Clustering, which identifies different clusters using a tree-shaped structure known as the dendrogram, Partitional Clustering such as k-means clustering, where the algorithm reallocates items to an initially specified number of groups and density-based clustering, such as DBSCAN algorithm which is used to discover randomly shaped clusters [11]. These methods can be used to differentiate the different types of personality based on different characteristics and categories such as questions about phobias and personality traits.

Understanding and characterizing human personalities is critical in today's interconnected society, from human resources and marketing to healthcare and social sciences. However, the existing analysis of personality traits is mainly dependent on manual classification and subjective interpretation, leading to inefficiencies and possible biases, particularly for large and diverse populations. Therefore, there is an urgent need for a more efficient and accurate personality recognition system that uses machine learning technology. The major problem is to create a robust clustering analysis approach that can handle qualitative data, which is inherently subjective and complicated. The proposed clustering techniques can divide people into distinct groups based on their personality traits, providing valuable information on how people interact, collaborate, and behave in different social circumstances.

This technology can alter how we understand human behavior, improve decision-making processes, and promote personalized experiences and services by taking a data-driven and objective approach to personality assessment. However, careful consideration of ethical and privacy problems and validation with psychiatric experts are required for the appropriate development and implementation of such a system. In this context, future research should look at novel ways to integrate qualitative and quantitative data, addressing cross-cultural differences, and ensuring the system's interpretability and fairness to enable significant applications across various areas.

This study aims to identify similarities in unlabeled data related to personality traits using unsupervised clustering methods. Various clustering techniques will be utilized to group different personality traits according to interview questions. For instance, clustering techniques include Spectral Clustering, Ordering Points To Identify the Clustering Structure (OPTICS), K-modes, Gaussian Mixture Model, and Affinity Propagation. We can validate the precision and effectiveness of the personality traits for each clustering technique after training, testing and evaluating those models. Therefore, we can advance our understanding of human behavior and personality traits through data-driven approaches.

This paper is constructed with the following sections. Section II presents the latest studies and research on clustering personality using different algorithms. Section III explains the research methodology design. This is followed by Section IV which discusses the result and finally Section V concludes the study with limitations and future works.

II. Literature Review

A. *Spectral Clustering*

Spectral clustering has become more common recently as it is a simple algorithm to implement and does not require a large amount of resources to process the algorithm [12]. There are a few pros and cons regarding spectral clustering. For example, this algorithm applies to data with high dimensionality, and it also handles categorical variables well as it can calculate the similarity between data points by using an eigenvector instead of using distance to calculate the data points, which is crucial to our research. On the other hand, spectral clustering is relatively slow compared to other traditional algorithms such as k-mean clustering and we are required to determine the k-value of the spectral clustering algorithm, which may be hard if we do not have an intuition on the number of clusters a dataset should have [13], [14].

The algorithm falls into the category of clustering like k-mean clustering and uses the eigenvector of a matrix obtained from the distance between the data points. Eigenvectors and eigenvalues are relatively important concepts in spectral clustering, where the eigenvector is interpreted as a vector that undergoes pure scaling without any rotation and the eigenvalue is interpreted as the scaling factor of a vector [15].

$$A\upsilon = \lambda\upsilon \tag{1}$$

The first step of spectral clustering is to construct an affinity matrix based on the data set. The affinity matrix is used to construct and determine a matrix of similarity between the data points. The affinity matrix is then normalized and partitioned using the largest K-eigenvectors [16]. Since spectral clustering is considered partitional clustering, it has the evaluation metrics of silhouette score, Calinski-Harabasz index, and Davies-Bouldin Index. On the basis of the previous study, we can see that spectral clustering is commonly applied in various real-life situations. For example, a study conducted by McFee & Ellis (2014) used the algorithm to analyze the structure of songs that detects repeated patterns in songs [13]. Furthermore, in a study conducted by Bach & Jordan (2006), (1) is used to perform speech separation [17].

B. *OPTICS*

OPTICS (Ordering Points to Identify the Clustering Structure) is categorized as a density-based clustering algorithm. The algorithm is inspired by DBSCAN but with a few better modifications. For example, the OPTICS clustering algorithm can partition data with varying densities and shapes. It is widely used in large data sets with high dimensionality [18]. The OPTICS algorithm works by adding a random data point in the cluster to an order list and then continuing to expand the cluster iteratively based on the closest data point to the selected data points. The OPTICS algorithm also calculates the reachability distance for each data point [19]. Based on previous studies, the algorithm is also commonly used in real-life applications such as detecting outliers in data sets and clustering wireless sensor networks [20], [21], [22].

Reachability distance is a measurement that indicates how easily can a datapoint be reached by another data point can reach a data point. Euclidean distance is used to calculate the distance between two data points and then further be used to determine the reachability distance. The algorithm computes the reachability distance of each data point and the generated data is used to plot out the reachability plot which can help identify clusters and the hierarchical structure of the data [23].

Even though the OPTICS algorithm uses similar concept as DBSCAN algorithm which uses density-based clustering, the OPTICS algorithm is considered a better algorithm as the OPTICS algorithm maintains a priority queue to determine the reachability distance, whereas DBSCAN only uses radius queries. Furthermore, the OPTICS clustering technique requires less maintenance, as it has fewer parameters compared to DBSCAN which requires one to maintain the epsilon parameters but to optimize the parameters of the OPTICS algorithm, domain knowledge is required as it is very sensitive to parameters that define density [24]. In addition to that, the OPTICS algorithm is also good at handling clusters with different densities, whereas the DBSCAN clustering algorithm struggles to handle data sets with different densities of clusters, since it only depends on a single value of epsilon to help determine the cluster size for all data points [18].

C. KMODES

The K mode algorithm is a variant of one of the most popular clustering algorithms, the K mean algorithm. K-modes are designed to specifically handle categorical data instead of numerical data which the K-means algorithm is weak at. The k-mode algorithm calculates the mode instead of the mean of the clusters. This modification allows the K-modes algorithm to cluster large categorical datasets more efficiently compared to the K-means algorithm [25]. Furthermore, based on previous studies, the algorithm is used for customer segmentation in e-Commerce business [26]. In addition to that, the algorithm is also used to detect and prevent crime by combining data mining technology with the algorithm. Equations (2) and (3) show the calculation of KMODES.

$$d(x,y) = \sum_{i=1}^{f} \delta(X_j, Y_j) \tag{2}$$

$$C(Q) = \sum_{i=1}^{n} d(Z_i, Q_i) \tag{3}$$

The K-modes algorithm works by first generating a number of clusters by randomly selecting data points to act as the initial cluster centers with each data point representing the centroid of the cluster where the value k is selected manually by the user. Then, each data point in the dataset is assigned to the cluster whose cluster center is closest to it based on the second equation above. After assignment, the clusters are updated based on the allocated data points, and the cluster centers are recalculated to represent the new centroids of each cluster. This update involves calculating the latest modes for each cluster. These steps and calculations are repeated until convergence is reached or stopped based on predefined criteria [27].

D. Gaussian Mixture Model

The Gaussian mixture model is a parametric density function of probability that can be represented as the weighted sum of Gaussian component densities in many applications, such as image clustering to detect human skin color and image segmentation, identifying restaurant hotspots, weather forecasts, flight safety monitoring, and many others [28], [29], [30], [31], [32], [33]. Before getting into the Gaussian mixture model, two key components of clustering must be known which are hard clustering and soft clustering. Hard clustering is the method where models try to force a data point to one of the clusters, and this means that the data point is assigned a membership degree to either 0 or 1. On the other hand, soft clustering models tend to assign data points to their appropriate membership value. Instead of assigning a membership degree of either 0 or 1, soft clustering assigns the membership degree of a point between 0 and 1. This means that each data point can belong to two or more clusters, and this may be more natural in many situations compared to hard clustering [34].

The Gaussian mixture model falls under the category of soft clustering model as it allows data points to be assigned partially to clusters. The model works by clustering data points into different clusters and thus estimating the probability density of new data points. For example, a data point can be 30% cluster A and 70% cluster B. The Gaussian mixture model is commonly applied in machine learning and pattern recognition due to the ability of the model to handle complex data distributions. Even so, it is difficult to implement this algorithm incorporating categorical variables, as the model assumes that all the features are normally distributed. The model works by applying the EM on top of its algorithm as shown in (4). The E-step uses training data to predict a new datapoint, whereas the M-step uses the information obtained from the E-step to calculate the derivative of the log-likelihood to cluster the new datapoint according to the calculations. This process repeats until convergence and the Q calculated from the EM step can be used for clustering purposes [35].

$$Q^i\left(z_k^{(i)}\right) = p\left(z_k^{(i)} \mid x^{(i)};\ \theta\right) = \frac{\pi_k N\,(x^{(i)};\, \mu_k\,, \Sigma k)}{\sum_{k=1}^{K} \pi_k N\,(x^{(i)};\, \mu_k\,, \Sigma k)} \tag{4}$$

E. Affinity Propagation

The Affinity Propagation algorithm is a clustering algorithm that first aims to find a data point as an exemplar for each cluster. It group the data based on either their similarity or distance [36]. This algorithm is commonly used to group images such as human faces, and is also used for forecasting rainfall [37]. Affinity propagation algorithm does not require determining the number of clusters, unlike algorithms like k-mean clustering. This is made possible due to the parameters used in the Affinity Propagation algorithm, such as preference, which handles the number of exemplars being used, and the damping factor, which is responsible for preventing large changes in the Responsibility and Availability matrices and ensuring convergence in the algorithm [38]. This method requires heavy calculations as it needs to calculate the similarity value between each data and reselect data as an exemplar. This makes the algorithm computationally expensive and makes it difficult to scale with larger datasets.

The algorithm performs by first creating a similarity matrix. Given a dataset, the similarity between data points needs to be calculated by using some commonly used equations, such as the negative Euclidean distance. The higher the similarity of the data points, the closer the data points are to each other based on the similarity equation. Then, two matrices, the Responsibility matrix, which quantifies how well-suited a data point is to be the exemplar of another data point, and the Availability matrix, which quantifies how much support a data point receives from another data point as an exemplar is created. Then both matrices are updated and recalculated iteratively until convergence and new data points can be clustered on the existing clusters. This algorithm is useful when the initial number of clusters is unknown, and it can handle complex data structures, too.

III. Methodology

A. Data Collection and Preparation

The data set that this study has chosen represents a list of survey answers regarding preference, interests, habits, opinions, and fears of people ages 15 years old to 30 years old. This survey was conducted by students of the Statistics class at FSEV UK in 2013. It contains 1010 unique records with 150 columns, and was taken from Kaggle, a platform where it consistently hosts various types of forecasting competitions, and it also contains data sets provided by various data scientists. Instead of taking the whole dataset, we have decided to focus on the phobia, personality, and demographic category in this data set. The demographic will be then combined with the personality subset, and phobia subset which may provide more information gain for the clustering algorithms. The phobia subset contains 10 independent variables which are flying, storm, darkness, heights, spiders, snakes, rats, aging, dangerous dogs, and fear of public speaking where it was rated from 1-5 indicating the severity of fear. The description of each variable in each subset is shown in Tables I and II. These variables in each subset will be used in clustering modelling.

TABLE I. VARIABLES IN THE PERSONALITY AND PHOBIA SUBSET (INT64)

Personality Subset			
Daily Events Prioritizing workload Writing Notes Workaholism Thinking ahead Final judgment Reliability Keeping promises Loss of interest Funniness Fake Criminal damage Decision Making Friends versus money	Elections Self-criticism Judgment calls Hypochondria Empathy Eating to survive Giving Borrowed stuff Loneliness Cheating in school Health Change in the past God Dreams Charity Number of friends	Compassion to animals Punctuality Lying New Environment Mood swings Appearance and gestures Socializing Achievements Response to a serious letter Children Assertiveness Getting angry Knowing the right people	Public speaking Unpopularity Waiting Life struggles Happiness in life Energy levels Small – big dogs Personality Finding Lost Values Getting up Interests or hobbies Parent' advice Questionnaires or polls Internet usage
Phobia Subset			
Flying Storm Darkness	Heights Spiders Snakes	Rats Ageing	Dangerous Dogs Fear of public Speaking

TABLE II. DEMOGRAPHIC SUBSET

Column Name	Type of Data
Age	Integer
Height	Integer
Weight	Integer
Number of siblings	Integer
Gender	String
Left – right-handed	String
Education	String
Only child	String
Village - town	String
House – block of flats	String

B. Feature Selection

Fig. 1 shows the method we have used to extract subsets from the dataset obtained from Kaggle. The phobia, personality, phobia + demographic and personality + demographic subset will then be used to fit the clustering algorithm. The dataset is first preprocessed by using methods like feature mapping to convert object data type into numeric datatype. Then the missing values are replaced by using Mean, Mode and Median depending on the distribution of the columns. Fig. 2 indicates that more than 80% of the number of friends' columns in the personality data set are null values. Therefore, the column is disputed due to lack of information in the column.

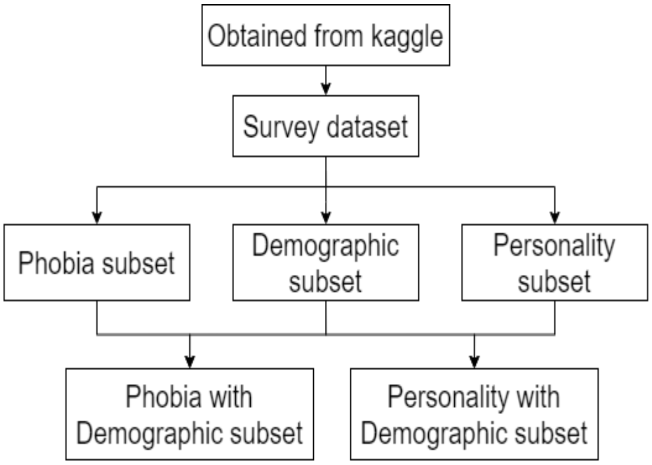


Fig. 1. Data pipeline.

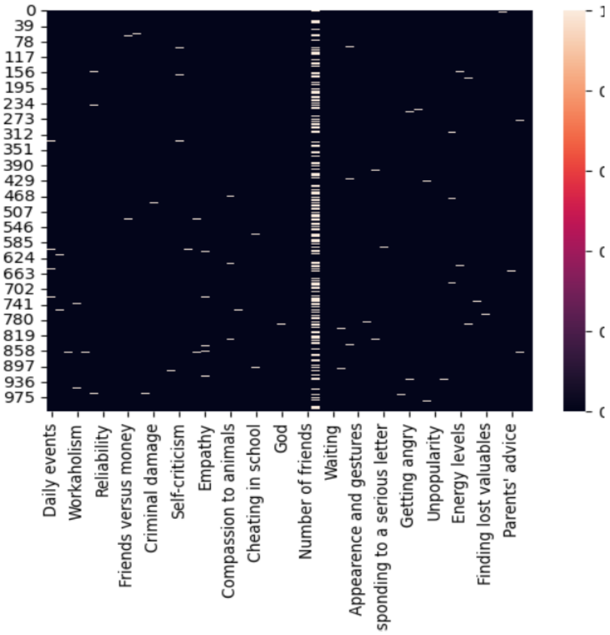


Fig. 2. Heatmap of missing value in the personality data set.

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \quad (5)$$

Eq. (5) shows the formula of variance which will be used for feature selection in our case. Selecting features using variance threshold is a common and efficient way to dispute columns that do not provide information to the model. Feature with low variance indicates that the data in the feature do not vary much and thus not providing good information to the clustering model. 20% is used in our dataset and the variables are removed from the data set. This indicates that both variables do not provide enough information for the model as these two columns only consist of similar values, which decreases the variance of the variables. Features with low variance can sometimes be redundant, meaning they carry information like other features. Removing redundant features can help simplify your model and potentially improve its performance.

C. Algorithms Used

1) Spectral clustering: Spectral clustering has become more common recently as it is a simple algorithm to implement and does not require a large amount of resources to process the algorithm [12]. There are a few pros and cons regarding spectral clustering. For example, this algorithm applies to data with high dimensionality, and it also handles categorical variables well as it can calculate the similarity between data points by using an eigenvector instead of using distance to calculate the data points, which is crucial to our research. The differences between K-means and Spectral Clustering are shown in Fig. 3 where Spectral Clustering can perform on high-dimensionality data whereas K-means perform poorly. On the other hand, spectral clustering is relatively slow compared to other traditional algorithms such as k-mean clustering, and we are required to determine the k-value of the spectral clustering algorithm, which may be hard if we do not have an intuition on the number of clusters a dataset should have [14].

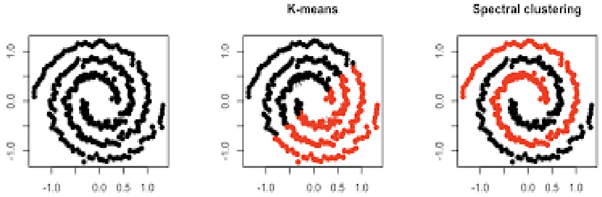


Fig. 3. Difference between K-means and spectral clustering.

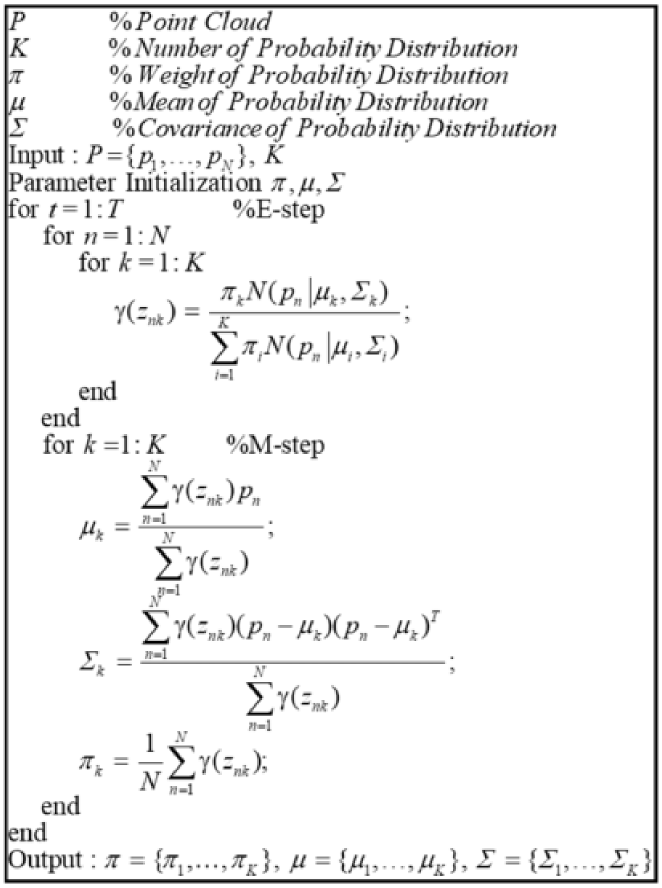
The algorithm falls into the category of clustering like k-mean clustering and uses the eigenvector of a matrix obtained from the distance between the data points. Eigenvectors and eigenvalues are relatively important concepts in spectral clustering, where the eigenvector is interpreted as a vector that undergoes pure scaling without any rotation and the eigenvalue is interpreted as the scaling factor of a vector [15].

$$Au = \lambda u \quad (6)$$

The first step of spectral clustering is to construct an affinity matrix based on the data set. The affinity matrix is used to construct and determine a matrix of similarity between the data points. The affinity matrix (6) is then normalized and partitioned using the largest K-eigenvectors [16]. Since spectral clustering is considered partitional clustering, it has the evaluation metrics of silhouette score, Calinski-Harabasz index, and Davies-Bouldin Index.

2) Gaussian Mixture Model (GMM): The Gaussian mixture model is a parametric density function of probability which can be represented as the weighted sum of the densities of the Gaussian components [28]. Based on the study conducted by other researchers, we can see that the algorithm was applied in image clustering where it detects human skin color and image segmentation [33]. Before getting into the Gaussian mixture model, two key components of clustering must be known which are hard clustering and soft clustering. Hard clustering is the method where models try to force a data point to one of the clusters, and this means that the data point is assigned a membership degree to either 0 or 1. This means that each data point can belong to two or more clusters, and this may be more natural in many situations compared to hard clustering. The Gaussian mixture model falls under the category of soft clustering model as it allows data points to be assigned partially to clusters. This process (Fig. 4) repeats until convergence and the Q calculated from the EM step can be used for clustering purposes [35]. The algorithm used is shown in (6).

$$Q^i\left(z_k^{(i)}\right) = p\left(z_k^{(i)} \,\middle|\, x^{(i)} ;\, \theta\right) = \frac{\pi_k N(x^{(i)}; \mu_k, \sum k)}{\sum_{k=1}^{K} \pi_k N(x^{(i)}; \pi_k, \sum k)} \quad (6)$$



```
$P$           %Point Cloud
$K$           %Number of Probability Distribution
$\pi$         %Weight of Probability Distribution
$\mu$         %Mean of Probability Distribution
$\Sigma$      %Covariance of Probability Distribution
Input : $P = \{p_1, \ldots, p_N\}$, $K$
Parameter Initialization $\pi, \mu, \Sigma$
for $t = 1 : T$            %E-step
    for $n = 1 : N$
        for $k = 1 : K$
            $\gamma(z_{nk}) = \frac{\pi_k N(p_n | \mu_k, \Sigma_k)}{\sum_{i=1}^{K} \pi_i N(p_n | \mu_i, \Sigma_i)}$;
        end
    end
    for $k = 1 : K$        %M-step
        $\mu_k = \frac{\sum_{n=1}^{N} \gamma(z_{nk}) p_n}{\sum_{n=1}^{N} \gamma(z_{nk})}$;
        $\Sigma_k = \frac{\sum_{n=1}^{N} \gamma(z_{nk})(p_n - \mu_k)(p_n - \mu_k)^T}{\sum_{n=1}^{N} \gamma(z_{nk})}$;
        $\pi_k = \frac{1}{N} \sum_{n=1}^{N} \gamma(z_{nk})$;
    end
end
Output : $\pi = \{\pi_1, \ldots, \pi_K\}$, $\mu = \{\mu_1, \ldots, \mu_K\}$, $\Sigma = \{\Sigma_1, \ldots, \Sigma_K\}$
```

Fig. 4. Pseudocode of the E-M step in GMM.

D. Model Pipeline

Fig. 5 indicates the clustering method used in this study. The four different subsets will be inputted into the spectral clustering model and the GMM model. Then we will determine whether the subset with demographic content or without demographic content will be used based on its performance in the model. After determining the subset to be used, the model will then be fine-tuned by using tools like GridSearchCV, silhouette score, and elbow graph to determine the best parameter combination. Lastly, each model will generate two clusters which are Phobia cluster and Personality cluster.

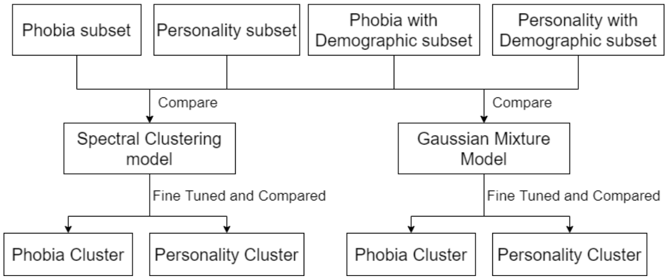


Fig. 5. Model pipeline flow chart.

IV. Result and Discussion

A. Exploratory Data Analysis (EDA)

This section explores the data set using heatmap and line graph, bar chart, violin plot, and cluster map. A heatmap is used to identify potential association between 1) phobias and demographic categories (age, gender, education) as shown in Fig. 6; and 2) personality traits and demographic categories. Based on the phobia vs. Age line graph in Fig. 7, we can see that the level of phobia roughly remains the same throughout the graph. Based on this, we can conclude that the things we fear are the same no matter the age. Fear of height can be the same for a 16-year-old young adult or even a 30-year-old adult.

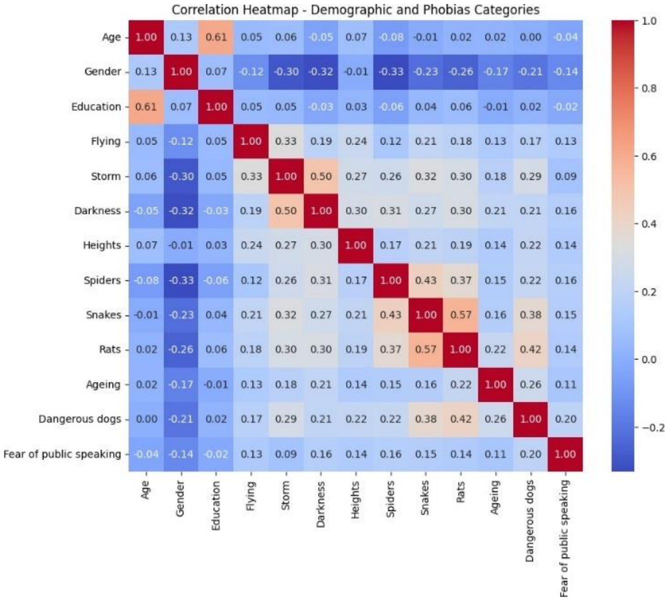


Fig. 6. Heatmap of demographic and phobia categories.

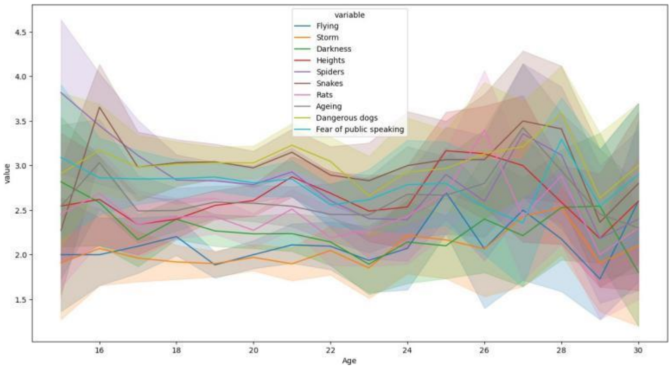


Fig. 7. Phobia vs. Age line graph.

From the bar graph in Fig. 8, 1 represents women and 2 represent males. Due to the large amount of data and bar graphs produced, only a snippet of the bar graph is presented in Fig. 8. The bar graph clearly indicates that female participants gave a neutral value in most views on life and opinion.

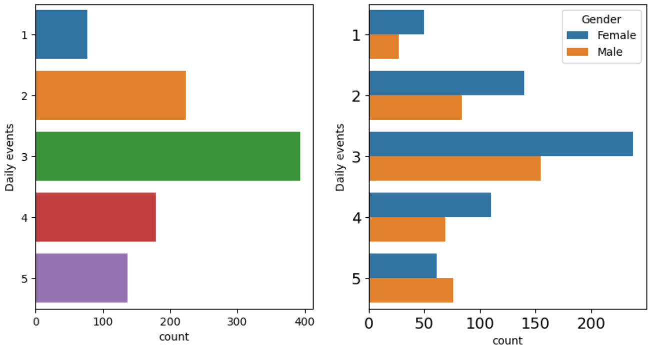


Fig. 8. Snippet of bar plots produced in exploratory data analysis using Python.

The shape of the violin represents the distribution of the data. The wider sections indicate higher density, while narrower sections indicate lower density. The width at a specific point on the violin corresponds to the frequency of data values at that point. Fig. 9 shows the distribution for level of phobia for each phobia based on the questionnaire.

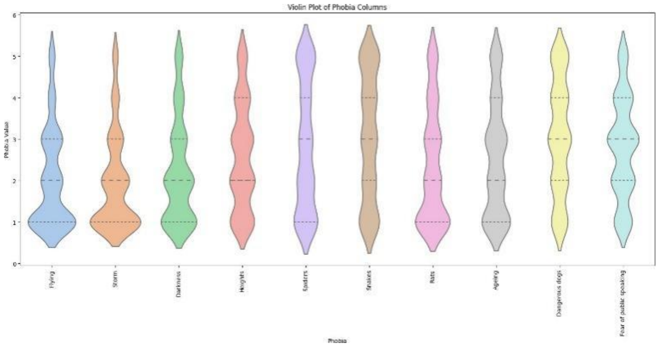


Fig. 9. Violin plot to represent data distribution.

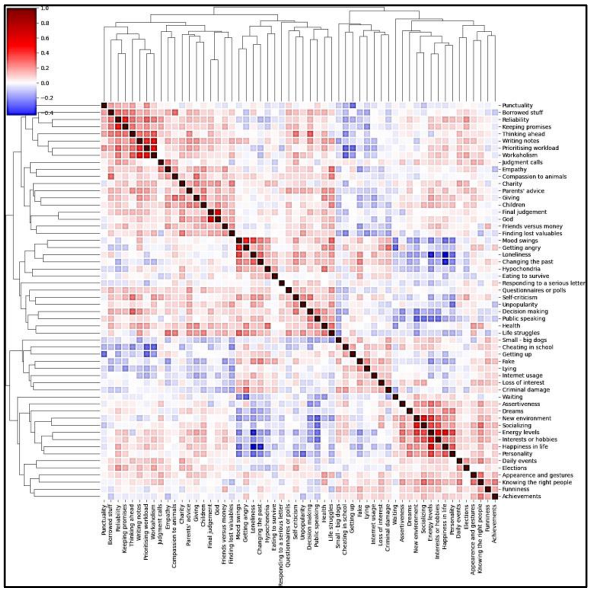


Fig. 10. Cluster map with associated clusters.

The cluster map is useful for identifying patterns not only in individual values but also in the arrangement of rows and columns. Group similar rows and/or columns together to form clusters, which can reveal underlying structures or relationships in the data. The cluster map in Fig. 10 indicates that the data can be grouped into 4, 5 or 6 clusters. Cluster map showing the correlation between each attribute in the personality test. Meanwhile, Fig. 11 cluster map shows that weight, height, and gender are highly correlated, followed by age and education and followed by number of siblings and only child. Therefore, the data can be grouped into 3 clusters. According to Fig. 12 cluster map, there may have 2 clusters for phobia data frame which are storm and darkness, as well as spiders, snakes, and rats.

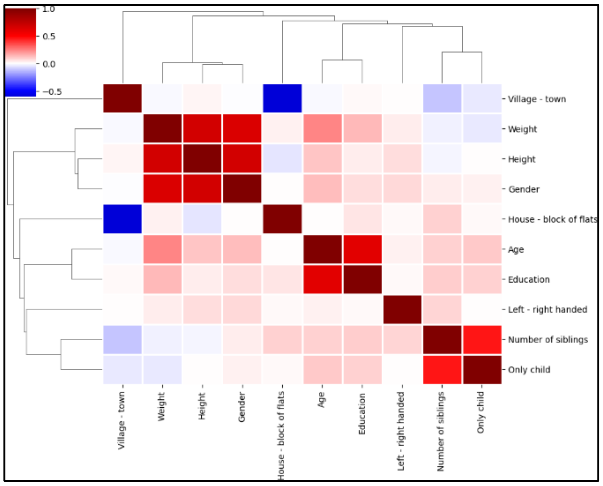


Fig. 11. Cluster map for displaying the correlation between demographics.

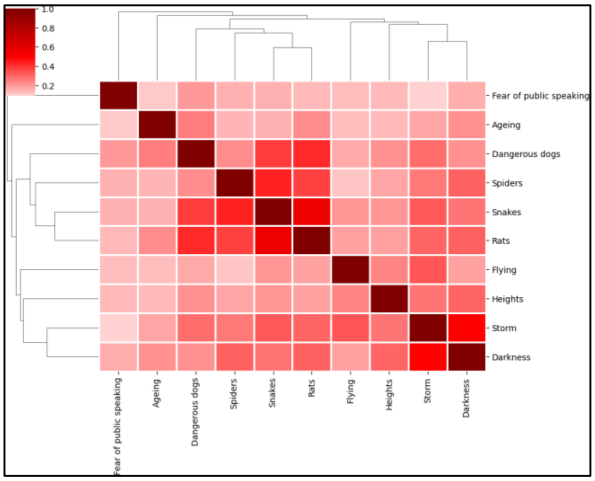


Fig. 12. Cluster map to display the correlation between each types of phobia.

B. Gaussian Mixture Model

By combining the result distortion elbow method and silhouette score graph in different numbers of clusters, we can determine that the optimal dataset and number of clusters are a subset without demographic content, 2 clusters for phobia subset and 4 clusters for personality subset. Based on Fig. 13, even though the graph indicates that 2 clusters is optimal, we have decided to use 4 clusters because it represents four main dominant traits of a person.

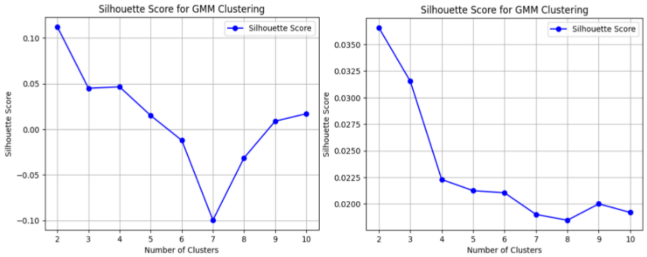


Fig. 13. GMM Silhouette score graph for the phobia subset and the personality subset.

The graphs on Fig. 14 and 15 are obtained by letting the model fit and predict the dimensional reduced data. The dimensionality reduction phase is carried out by using Principal Component Analysis (PCA). Based on Fig. 7, we can see that the yellow groups represent Level 1 phobia while the purple clusters represent Level 2 phobia. Furthermore, the clusters in Fig. 8 also show 4 different clusters of dominant traits. But in the case of GMM, the clusters are not well defined and are all grouped together. Therefore, we can determine that the GMM clustering model is not suitable for clustering personalities as it cannot identify the hidden relationship in the data set.

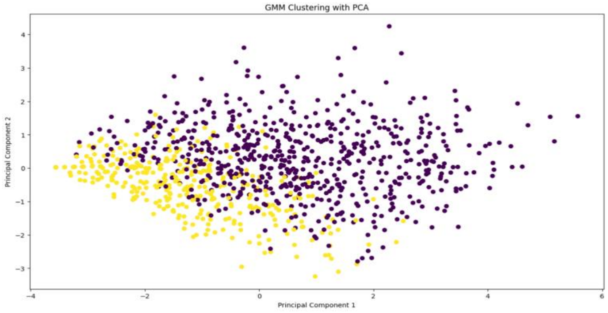


Fig. 14. Groups of GMM phobia subsets.

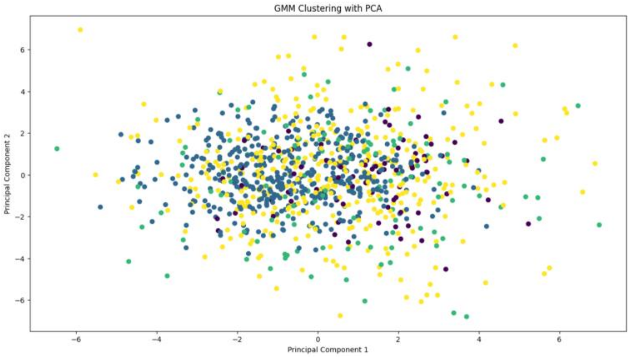


Fig. 15. Groups of personality subsets of GMM.

C. Spectral Clustering

The process of selecting parameters for spectral clustering is the same as GMM. The subset with demographic contents is selected, as the result has shown that subset with demographic contents provide better information gain and higher silhouette score. By observing both the distortion elbow method and silhouette analysis in Fig. 16 and 17 we can see that the optimal cluster is 2 clusters for the phobia subset and 4 clusters for the personality subset. Based on Fig. 9, 4 clusters since it represents 4 main dominant traits of a person. Not only that, we have also used GridSearchCV to adjust parameters like “affinity”, “gamma”, and “eigen_solver”. But the best combination found by GridSearchCV does not cluster the data well, therefore, we have decided to use the default parameters.

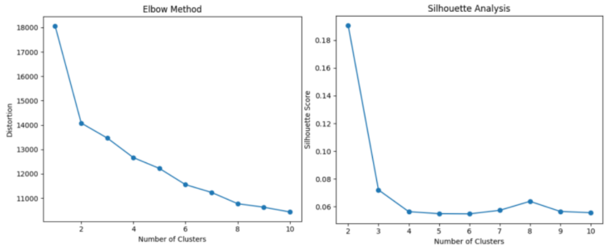


Fig. 16. Spectral clustering distortion and silhouette graph for phobia with demographic subset.

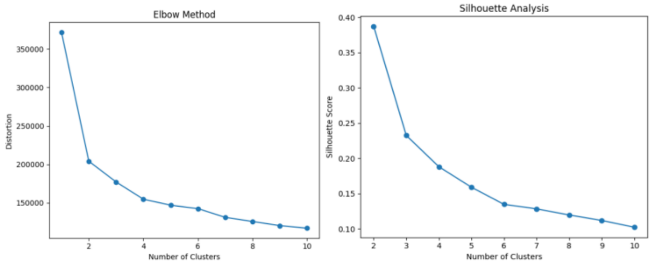


Fig. 17. Spectral clustering distortion and silhouette graph for personality with demographic subset.

The graphs in Fig. 18 and 19 are obtained from fitting dimensionality reduced data by using PCA into the spectral clustering model. Based on Fig. 18, we can see that the yellow clusters represent Level 1 phobia, while the purple clusters represent Level 2 phobia. Furthermore, the clusters in Fig. 19 also show four different clusters of dominant traits where the yellow clusters (Cluster 0) represent Steady personality type, purple clusters (Cluster 1) represent Influential personality type, blue clusters (Cluster 2) represent Compliant Personality Type, and green cluster (Cluster 3) represents dominant personality type.

Compared to previous research of clustering phobias and personality categories, one very relevant research by Anik et al. (2024) clustered the different types of phobia types using BERT-based classification model on Tweet data set. This present study presents different approaches in classification which provides information to alternative methods to analyze different data sets and discover different results or visualization.

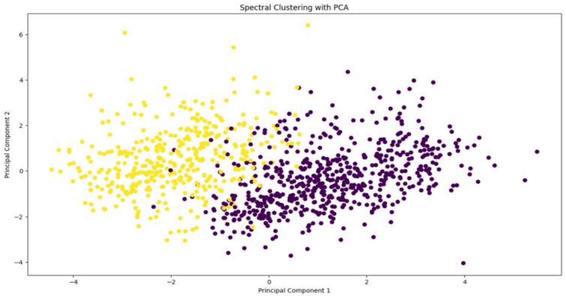


Fig. 18. Spectral clustering phobia with demographic subset clusters.

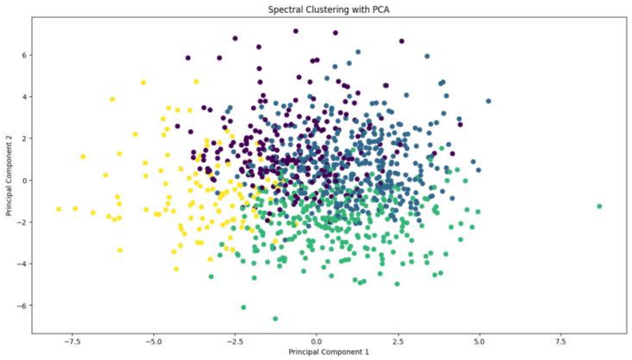


Fig. 19. Spectral clustering personality with demographic subset clusters.

V. Conclusion

As we can see, we have used two different models for two different datasets (phobias and personality categories) (one with demographic content and one without it). The list of models used for GMM are 1) GMM for phobia dataset (with demographic content); 2) GMM for phobia dataset (Without demographic content); 3) GMM for personality dataset (With demographic content); 4) GMM for personality dataset (Without demographic content). Meanwhile, for spectral clustering, the list of models used are: 1) SC for phobia dataset (with demographic content); 2) SC for phobia dataset (Without demographic content); 3) SC for personality dataset (With demographic content); 4) SC for personality dataset (Without demographic content).

For the phobia dataset, the best-performing model is Spectral Clustering, as it has a higher average silhouette score and thus makes well-defined clusters compared to GMM. For the personality dataset, the best-performing model is also spectral clustering, as it can identify the hidden relationship between the questions and thus cluster a more distinct cluster from each other. This can be shown by different clusters having more distinct characteristics, and the clusters are also less intersected with other clusters. Therefore, we can use this model to identify the dominant traits of an individual, to understand more about them. This helps identify strengths and weaknesses of the person, and thus reduces conflicts and makes coordination and cooperation much easier in many domains, especially group work in the school, university, or job. The models can be further tested using different data sets from different countries to ensure generalization of the result.

References

[1] H. Gillette, "Are you born with personality or does it develop later on? Psych Central." Accessed: Jul. 19, 2024. [Online]. Available: https://psychcentral.com/health/personality-development

[2] G. Matthews, I. J. Deary, and M. C. Whiteman, Personality traits, 2nd ed. Cambridge University Press, 2005.

[3] L. Mosley, "The importance of understanding personality type in the workplace," LinkedIn. Accessed: Jul. 19, 2024. [Online]. Available: https://www.linkedin.com/pulse/importance-understanding-personality-type-workplace-lauren-copeland/

[4] A. Cabrera, "Personalities in the workplace: Why is it important?," Peopledynamics. Accessed: Jul. 19, 2024. [Online]. Available: https://peopledynamics.co/personalities-workplace-importance/#:~:text=Understanding%20your%20people%27s%20personalities%20can,chisel%20that%20tears%20it%20apart

[5] A. Talasbek, A. Serek, M. Zhaparov, S.-M. Yoo, Y.-K. Kim, and G.-H. Jeong, "Personality Classification Experiment by Applying k-Means Clustering," International Journal of Emerging Technologies in Learning (iJET), vol. 15, no. 16, p. 162, Aug. 2020, doi: 10.3991/ijet.v15i16.15049.

[6] R. Karthika, V. E. Jesi, M. S. Christo, L. J. Deborah, A. Sivaraman, and S. Kumar, "Intelligent personalised learning system based on emotions in e-learning," Pers Ubiquitous Comput, vol. 27, no. 6, pp. 2211–2223, Dec. 2023, doi: 10.1007/s00779-023-01764-7.

[7] S. M. Aslam, A. K. Jilani, J. Sultana, and L. Almutairi, "Feature Evaluation of Emerging E-Learning Systems Using Machine Learning: An Extensive Survey," IEEE Access, vol. 9, pp. 69573–69587, 2021, doi: 10.1109/ACCESS.2021.3077663.

[8] M. Rahman, H. Sarwar, MD. A. Kader, T. Gonçalves, and T. T. Tin, "Review and Empirical Analysis of Machine Learning-Based Software Effort Estimation," IEEE Access, vol. 12, pp. 85661–85680, 2024, doi: 10.1109/ACCESS.2024.3404879.

[9] C. W. Puah, W. L. Eng, C. H. Tan, S. C. Tan, and T. T. Ting, "Digital Culture: Online shopping adoption among college students in Malaysia," in International Conference on Digital Transformation and Applications, 2021, pp. 137–143.

[10] Feng Tian, Shibin Wang, Cheng Zheng, and Qinghua Zheng, "Research on E-learner Personality Grouping Based on Fuzzy Clustering Analysis," in 2008 12th International Conference on Computer Supported Cooperative Work in Design, IEEE, Apr. 2008, pp. 1035–1040. doi: 10.1109/CSCWD.2008.4537122.

[11] T. S. Madhulatha, "AN OVERVIEW ON CLUSTERING METHODS," IOSR Journal of Engineering, vol. 02, no. 04, pp. 719–725, Apr. 2012, doi: 10.9790/3021-0204719725.

[12] U. von Luxburg, "A tutorial on spectral clustering," Stat Comput, vol. 17, no. 4, pp. 395–416, Dec. 2007, doi: 10.1007/s11222-007-9033-z.

[13] B. McFee and D. P. W. Ellis, "Analyzing Song Structure with Spectral Clustering," in 15th International Society for Music Information Retrieval Conference, 2014.

[14] C. Ellis, "When to use spectral clustering." Accessed: Jul. 20, 2024. [Online]. Available: https://crunchingthedata.com/when-to-use-spectral-clustering/

[15] H. Abdi, "The Eigen-Decomposition: Eigenvalues and Eigenvectors." Accessed: Jul. 19, 2024. [Online]. Available: https://personal.utdallas.edu/~herve/Abdi-EVD2007-pretty.pdf

[16] X.-Y. Li and L. Guo, "Constructing affinity matrix in spectral clustering based on neighbor propagation," Neurocomputing, vol. 97, pp. 125–130, Nov. 2012, doi: 10.1016/j.neucom.2012.06.023.

[17] F. R. Bach and M. I. Jordan, "Learning spectral clustering, with application to speech separation," Journal of Machine Learning Research, vol. 7, pp. 1963–2001, 2006, Accessed: Jul. 19, 2024. [Online]. Available: https://jmlr.csail.mit.edu/papers/volume7/bach06b/bach06b.pdf

[18] A. Gupta, "ML | OPTICS Clustering Explanation," geeksforgeeks. Accessed: Jul. 19, 2024. [Online]. Available: https://www.geeksforgeeks.org/ml-optics-clustering-explanation/

[19] Z. Deng, Y. Hu, M. Zhu, X. Huang, and B. Du, "A scalable and fast OPTICS for clustering trajectory big data," Cluster Comput, vol. 18, no. 2, pp. 549–562, Jun. 2015, doi: 10.1007/s10586-014-0413-9.

[20] H. Hassanpour, A. H. Hamedi, P. Mhaskar, J. M. House, and T. I. Salsbury, "A hybrid clustering approach integrating first-principles knowledge with data for fault detection in HVAC systems," Comput Chem Eng, vol. 187, p. 108717, Aug. 2024, doi: 10.1016/j.compchemeng.2024.108717.

[21] M. Hajihosseinlou, A. Maghsoudi, and R. Ghezelbash, "A comprehensive evaluation of OPTICS, GMM and K-means clustering methodologies for geochemical anomaly detection connected with sample catchment basins," Geochemistry, vol. 84, no. 2, p. 126094, May 2024, doi: 10.1016/j.chemer.2024.126094.

[22] P. Lalwani, H. Banka, and C. Kumar, "CRWO: Clustering and routing in wireless sensor networks using optics inspired optimization," Peer Peer Netw Appl, vol. 10, no. 3, pp. 453–471, May 2017, doi: 10.1007/s12083-016-0531-7.

[23] G. Iván and V. Grolmusz, "On dimension reduction of clustering results in structural bioinformatics," Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics, vol. 1844, no. 12, pp. 2277–2283, Dec. 2014, doi: 10.1016/j.bbapap.2014.08.015.

[24] R. Roy, "Optics clustering (intro)," Medium. Accessed: Jul. 19, 2024. [Online]. Available: https://bobrupakroy.medium.com/optics-clustering-intro-76dcdaf94bde#:~:text=Disadvantages%3A&text=It%20fails%20if%20there%20are%20no%20density%20drops%20between%20clusters.&text=It%20is%20also%20sensitive%20to,parameter%20settings%20require%20domain%20knowledge

[25] Z. Huang and M. K. Ng, "A Note on K-modes Clustering," J Classif, vol. 20, no. 2, pp. 257–261, Sep. 2003, doi: 10.1007/s00357-003-0014-4.

[26] D. Kamthania, A. Pawa, and S. Madhavan, "Market Segmentation Analysis and Visualization using K-Mode Clustering Algorithm for E-Commerce Business," Journal of Computing and Information Technology, vol. 26, no. 1, pp. 57–68, 2018, doi: 10.20532/cit.2018.1003863.

[27] N. Sharma and N. Gaud, "K-modes Clustering Algorithm for Categorical Data," Int J Comput Appl, vol. 127, no. 17, pp. 1–6, Oct. 2015, doi: 10.5120/ijca2015906708.

[28] L. Li, R. J. Hansman, R. Palacios, and R. Welsch, "Anomaly detection via a Gaussian Mixture Model for flight operation and safety monitoring," Transp Res Part C Emerg Technol, vol. 64, pp. 45–57, Mar. 2016, doi: 10.1016/j.trc.2016.01.007.

[29] Y. Balakrishna, S. Manda, H. Mwambi, and A. van Graan, "Determining classes of food items for health requirements and nutrition guidelines using Gaussian mixture models," Front Nutr, vol. 10, Oct. 2023, doi: 10.3389/fnut.2023.1186221.

[30] G. Jouan, A. Cuzol, V. Monbet, and G. Monnier, "Gaussian mixture models for clustering and calibration of ensemble weather forecasts," Discrete and Continuous Dynamical Systems - S, vol. 16, no. 2, pp. 309–328, 2023, doi: 10.3934/dcdss.2022037.

[31] C. O'Sullivan, "Identifying Restaurant Hotspots with a Gaussian Mixture Model," Towards Data Science. Accessed: Jul. 08, 2024. [Online]. Available: https://towardsdatascience.com/identifying-restaurant-hotspots-with-a-gaussian-mixture-model-2a840ab0c782

[32] Amy, "Gaussian Mixture Model (GMM) for Anomaly Detection," GrabNGoInfo. Accessed: Jul. 08, 2024. [Online]. Available: https://medium.com/grabngoinfo/gaussian-mixture-model-gmm-for-anomaly-detection-e8360e6f4009

[33] M.-H. Yang and N. Ahuja, "Gaussian mixture model for human skin color and its applications in image and video," M. M. Yeung, B.-L. Yeo, and C. A. Bouman, Eds., Dec. 1998, pp. 458–466. doi: 10.1117/12.333865.

[34] D. J. Bora and A. K. Gupta, "A Comparative study Between Fuzzy Clustering Algorithm and Hard Clustering Algorithm," International Journal of Computer Trends and Technology, vol. 10, no. 2, pp. 108–113, 2014, Accessed: Jul. 19, 2024. [Online]. Available: https://arxiv.org/ftp/arxiv/papers/1404/1404.6059.pdf

[35] Q. Cai, Z. Xue, D. Mao, H. Li, and J. Cao, "Bike-Sharing Prediction System," 2016, pp. 301–317. doi: 10.1007/978-3-319-40259-8_27.

[36] L. Wang, K. Zheng, X. Tao, and X. Han, "Affinity propagation clustering algorithm based on large-scale data-set," International Journal of Computers and Applications, vol. 40, no. 3, pp. 1–6, Jul. 2018, doi: 10.1080/1206212X.2018.1425184.

[37] W. Huang and Y. Li, "Application of Affinity Propagation Clustering Method in Medium and Extended Range Forecasting of Heavy Rainfall Processes in China," Atmosphere (Basel), vol. 13, no. 5, p. 768, May 2022, doi: 10.3390/atmos13050768.

[38] D. Dey, "Affinity Propagation in ML | To find the number of clusters. GeeksforGeeks." Accessed: Jul. 19, 2024. [Online]. Available: https://www.geeksforgeeks.org/affinity-propagation-in-ml-to-find-the-number-of-clusters/