




Please cite the Published Version

Farrukh, Muhammed Umar, Wainwright, Richard, Crockett, Keeley , McLean, David  and Dagnall, Neil  (2023) Building Actionable Personas Using Machine Learning Techniques. In: 2022 IEEE Symposium Series on Computational Intelligence (IEEE SSCI), 04 December 2022 - 08 December 2022, Singapore.

DOI: <https://doi.org/10.1109/SSCI51031.2022.10022180>

Publisher: IEEE

Version: Accepted Version

Downloaded from: <https://e-space.mmu.ac.uk/630932/>

Additional Information: © 2023 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

Enquiries:

If you have questions about this document, contact openresearch@mmu.ac.uk. Please include the URL of the record in e-space. If you believe that your, or a third party's rights have been compromised through this document please see our Take Down policy (available from <https://www.mmu.ac.uk/library/using-the-library/policies-and-guidelines>)

Building Actionable Personas Using Machine Learning Techniques

1st Muhammed Umar Farrukh
The Insights Family
Manchester, GB
0000-0002-9882-8214

2nd Richard Wainwright
The Insights Family
Manchester, GB
0000-0001-8359-8199

3rd Dr. Keeley Crockett
Department of Computing and Mathematics
Manchester Metropolitan University
Manchester, GB
0000-0003-1941-6201

4th Dr. David McLean
Department of Computing and Mathematics
Manchester Metropolitan University
Manchester, GB
0000-0001-7894-5176

5th Dr. Neil Dagnall
Department of Psychology
Manchester Metropolitan University
Manchester, GB
0000-0003-0657-7604

Abstract—Personas are quantifiable and describable ways of grouping people based on their behaviours. They are valuable to businesses as it enables them to better understand their customer base. The creation of personas from survey data requires establishing the client requirements, building a quantifiable personality scale, developing personality questions for a survey, and human subjective analysis. In this work, we have utilised clustering to automate the persona development process.

We have developed a real-world survey for children (from 17 countries) which included 25 personality-based questions (based on the OCEAN model), 22 questions that captured purchase behaviour, and other general features from the children's landscape. There were 63,969 completed questionnaires with a high proportion of categorical features, which were preprocessed to allow different segmentation methods to be tested. Preliminary results with simple K-means and a Euclidean distance function demonstrated that this was inappropriate for the survey data set. A novel distance function for K-means clustering has been developed, which can handle a mixture of feature types and to allow the importance of each feature to be varied, using a linearly weighted distance method. The function also incorporates the haversine distance function to provide a distance between two locations, enabling potential cultural differences to be examined. We have also implemented Gaussian Mixture Model on the same feature set to compare the results and see the limitations of Gaussian Models

Our novel approach generated clusters based on a combination of features including personality, consumer behaviour and location which has demonstrated key cultural differences across the globe. Results from our novel approach show that location distance is one of the key features when constructing personas as culture (location) has a significant effect on the way children answer survey questions.

Index Terms—Clustering, Personas, K-means, Persona Segmentation, Survey Feature Clustering, Gaussian Mixture Models

I. INTRODUCTION

Personas are often used as a mechanism for understanding user's needs, attitudes and behaviours [1]. The concept of personas was introduced by Cooper [2] to facilitate designing

user-based profiles. Personas help to introduce imaginative profiles and to develop scenarios about the way products will be used by a consumer or how an advert or marketing campaign will be received. Personas are segmented profile features that represent a group of people (e.g. customers). They provide a broad range of qualitative and quantitative insights that other approaches do not typically offer [3].

There have been many research studies on creating personas [4]–[6] using user-based requirement. One of the key components in building personas is the OCEAN model [7], [8] which is widely used in psychology for the assessment of human personality. The acronym OCEAN comprises of the following five personality factors: Openness, Conscientiousness, Extroversion, Agreeableness and Neuroticism. Clustering has been around for many years and in recent years it has been used to segment survey based features [9] and is quickly becoming the key building block of the personification methodology [10]. There has been some interest in utilising clustering with the big 5 personality model with varying degrees of success [11]–[14]. In this paper we will introduce the methodology of persona development using clustering techniques and analyse how different features (mapped to survey questions) can be weighted depending on their perceived importance. Typically in clustering, expert knowledge is required to understand the resulting model [1]. In-depth analysis and expert knowledge is used to pick out the nuances from the cluster centroids to identify areas for persona labelling. The creation of personas comes as a result of the segmentation of profile based clusters which provide a richer understanding of the consumers for improved target marketing.

The first step in our persona development methodology requires the development of essential survey questions that can help us to better understand a person's psychological profile from the response made. For this we utilised the OCEAN model, based on 25 personality questions already used in [15], 22 questions that showed purchase behaviour (taken

from company survey), 5 value added features (taken from company survey which captured family related information and favorite subjects, hobbies) and what city and country the respondent was from, using digital survey technology [16]. We then collected data from 17 different countries (63969 data points) and tested the reliability of respondent responses for the OCEAN model questions using the psychometric based standard reliability technique known as Cronbach's alpha testing. These 25 questions consisted of 5 groups of 5 sub-factors where each group was related to one OCEAN dimension. All sub-factor question responses had a Cronbach's alpha score [17] of more than 0.70 which proved the reliability of the responses made.

The complete set of 53 question responses were then pre-processed, where they were converted into factors (e.g. 25 OCEAN model questions were converted into 5 main factors by taking the average among each trait of 5 questions). Our methodology successfully segmented these profile features using an adapted K-means clustering with a novel distance function. We then compared the results for standard K-means (Euclidean distance), K-means with our novel distance function and Gaussian mixture models (GMM), which is considered as a more advanced segmentation technique [18].

The main contributions of the research presented in this paper are:

- A new method which generates actionable personas [19] through segmentation of personality-based features combined with location and consumer based features extracted from global survey data. This method includes the generation of personality based statistical features used within Standard Tens (STEN) score methodology to compare the cultural differences amongst the survey population. The methodology is applied to a complex real-world dataset captured from 63969 survey responses over 17 countries. From this data set 5 personas were discovered through clustering using evidence based data set and these were then labelled by experts. The main benefit of these personas will be to the Digital industry (Entertainment industry which includes movie and television show creation industry) as these personas will reflect the top trends of particular regions. These labelled personas can be used by digital technology companies to study the audience before designing or creating any content. These personas will provide the audience (survey respondents) with a voice that can be reached by the digital content creation industries, ensuring that they create user-specific content.
- An adaptive distance function, built purposely for the identification of distinctive personality profiles is applied within the K-means clustering algorithm. A linear weighted distance function is employed to combine all features, whilst giving the client the ability to adjust the weights to focus on different features based upon marketing requirements. Usability of weights between range of [0.1-1.0] on the features (multiplying of weights with feature variable distance) also provided us the ca-

pability to increase or decrease importance of feature in segmentation.

The rest of the paper is organised as follows: Section II considers related work including the ocean model and a brief introduction in to clustering.. Section III introduces the new method which generates actionable personas through a series of empirical experiments. The data set used and a description of the features is provided. Section IV presents some of our initial results. Section V provides an in-depth discussion and what it means for the development of actionable personas. VI concludes the paper and presents ideas for further work.

II. BACKGROUND WORK

A. The OCEAN Model

The OCEAN model is a widely accepted methodology and construct that can be used to describe the personality variations across five predefined dimensions. Often commonly known as the five-factor model, the big 5 model, CANOE or, the OCEAN model the methodology was originally developed in 1949 by D.W. Fiske [20]. Since its inception, the model has gone through many iterations including work done by [21]–[24] however the work continues to strive for the collective goal of being able to effectively organise personality traits and segment these traits as an alternative to the comprehensive theory of personality. The OCEAN model comprises of the following five fundamental personality traits: Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism

B. Clustering

Commonly used within industry, clustering is used to segment data. There are four different groupings of clustering. These different methods are defined as: Centroid Models, Distribution Models, Density Models and Connectivity Models. Whether the groupings that are produced are good enough or not can be tested using different evaluation metrics including: The Silhouette Scores which measure the separation of clusters [25]. The Calinski-Harabasz Index which looks at the ratio within cluster distribution [26] as well as the the Davies-Bouldin Index This measure the spread and how dense clusters are [27]. To truly understand clusters and interpret what the clusters are showing, domain knowledge is required.

III. A METHOD FOR DEVELOPING ACTIONABLE PERSONAS

This section describes a generalized method for developing actionable personas through the creation of cluster profiles from survey data. Figure 1 provides an overview of the methodology of personas development.

We started by first designing a survey which required obtaining a number of questions that can help to analyze the common personality factors existing in the population, so we used the OCEAN model with 25 questions (5 questions associated with each of the 5-dimensions). In this study, we also wanted to look for common purchase behaviours existing in a population, so we used the 22 purchase consumer based questions designed by company research team. Finally, we

added 5 value added features along with location. Of these, 3 questions were related to family income and, 2 questions were related to likes and dislikes (favourite subject and favourite hobby). In location we asked about the current city and country of the respondent (taken as one combined location). A total of 53 question responses were captured per participant and these were then pre-processed to convert them into a set of features (see III, E and F) for segmentation techniques (clustering).

Overall the steps of methodology are as follows:

- Survey question design which is comprised of Personality questions (see subsection A), purchase consumer questions (see subsection B) and Value Added Features (see subsection C). Features were then extracted from survey questions.
- Survey Data Collection and pre-processing have been defined in subsections D and E.
- Feature Engineering and Extraction have been defined in subsection F which includes feature normalization, standardization, encoding and conversion of location-based features into co-ordinates.
- Data Analysis and Visualization is defined in section G
- Methods and Techniques implemented in three different phases have all been defined in section H

Figure 1 shows the complete process of personas development which is given all detailed in Section 3.

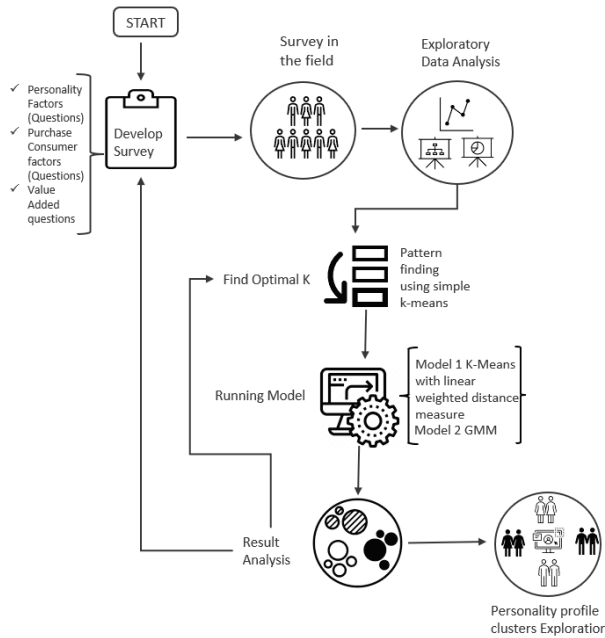


Fig. 1. Methodology Diagram

A. Personality Trait Feature Selection using 25 items questionnaire

To associate the common behavioural attributes with the Big 5 (OCEAN model) personality traits we followed the 25

items survey questionnaire methodology using a three point Likert scale from [15]. All 25 items (questions) consisted of simple behavioural question statements with a 3 point Likert scale, consisting of Not True-0(NT), Somewhat True-1(ST) and Certainly True-2(CT). The 25 Items were grouped into the Big 5 personality traits based on the results of a pilot study [15] in which psychometric analysis techniques, exploratory and confirmatory factor analyses and item response theory analysis was undertaken for a sample of data collected in 2014 from 642 children with a mean age of 11.7 years old [15]. In our work, the 25 question items were grouped into personality trait factors(OCEAN model) in the initial feature engineering phase to be used in segmentation of personality profile clusters.

B. Purchase Consumer Scale Development into Factors using 22 items questionnaire

Purchase consumer behaviour directly links towards the consumer behaviour of the particular respondents. A set of 22 items were developed by the company which were aimed to capture the purchase consumer behaviour of the kids as part of the core business. Prior to development of this study these 22 questions were already part of established company surveys and operational with existing clients. A sample of 40364 kids aged from 6-18 was taken from the survey data base. A 5 point Likert scale was used to capture the responses on a scale of 1-5 where 1 represented strongly disagree and 5 represented strongly agree. Using the Exploratory Factor Analysis methodology [28], the principal axis was applied on the 22 questions responses to group them into 3 different consumer factors:

- 1) New product/ Novelty
- 2) Convincing/ About me (self-conscious)
- 3) Hedonistically motivated consumer innovation (hMCI) (self-oriented/focused/more concerned personal preferences)

Consumer factors developed from the factor analysis methodology were grouped based on having a strong inter-correlation. The internal reliability of the factors was checked using Cronbach's alpha score [29]. For the three new factors generated the internal-reliability score was 0.692 (close to the acceptable threshold of 0.7)

C. Value Added Features

The survey included 5 closed questions which captured user-based attributes (family income, favourite hobby, family members, number of family members, favourite subject) related to user likeliness and background. These questions were designed to capture the relationship of different cultural and family related concepts.

D. Study Participants

Data was collected from 17 different countries which included Australia, Brazil, Canada, China, France, Germany, India, Indonesia, Italy, Japan, Mexico, Philippines, Poland, Russia, Spain, United Kingdom, and the USA. 63,969 data points were collected in total, 31,964 data points in total were

collected for boys estimating up-to 49.97% and 32005 data points were collected for girls estimating up-to 50.03%. This ensured that all the data points collected from each region consisted of equal percentage of male and female respondents for each age group. The age of participants involved in this study was from 6-18 years old. The survey was conducted digitally.

E. Pre-Processing and Analyzing of Data

In this step we started with grouping the 25 personality questions into their 5 main factors(e.g 5 questions asked to capture openness factor of respondent) by using the average technique (taking the average of the 5 responses), similarly Purchase consumer questions were grouped into 3 factors by using the average technique. All the responses were imputed for the missing values where no missing values were found for the personality scale items and purchase consumer scale items. Additional interest based questions responses were not given by some respondents, so all the blank answers were converted into Not Answered. Respondents personal information was stripped off in the initial stages of data pre-processing and not used as part of the research.

F. Feature Engineering and Extraction

Feature Engineering step comprises of transforming and pre-processing of all the features, in this step we started with behavioural based features grouping according to the the Personality OCEAN model in which average of the sub-trait features were calculated, total 5 average personality traits features were generated. 22 Purchase consumer based questions were grouped by Exploratory Factor Analysis using the principal axis [30](this technique is similar to grouping the items having very strong co-relation) which showed strong correlations among three groups, these features belonging to three different groups were grouped using average mean technique generating three purchase consumer factor features. Location feature when collected in the survey consisted of the central town or city name along with the country name (this also helped in maintaining the privacy of location for the respondent), was transformed into longitude and latitude. Age feature was re-scaled between value of 1 and 0 using min-max re-scaling technique whereas Gender feature was labelled as 0 and 1, where 0 was allotted for Male respondents and 1 for Female respondents. Features Family income was categorized into High, Medium and Low using the average binning technique where first the income level was converted into dollars and average income for that country was taken out, and then average income were assigned Medium, less than that were assigned Low and more than average was given High category (this step was done by manual research using different search engines such as Google). Family Members feature which had responses ranging between 1-4 or more than 5 was one hot encoded. Currently living with had 5 responses which were Living with Dad, Mum, Grandparents, younger Brother and Sister, older brother and sister, all responses for Living with features were one hot encoded as well. Favourite

Subject had a total of 19 different subjects as response from the 17 different countries, favourite hobby had a total of 10 different responses, both favourite subject and favourite hobby features were one hot encoded [31]. All the features which went through pre-processing stage are shown through diagram in figure 2

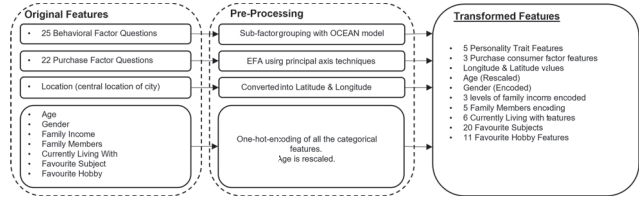


Fig. 2. Feature Pre-Processing

G. Data Analysis and Visualization

Personality scale Factors averages for each country were studied using the Standard Tens (STEN) score methodology. Standard Tens is a common methodology used for ranking the responses of personality questionnaire [32] items on a scale of 1-10 based on the calculation of the Z-score. It is also referred as a tool that divides the scale into 10 units where each unit is used for indication of an individual's ranked position in the population (or data set collected sample). Extreme ranges of the STEN score normally fall into the 1-2% of bell curve whereas the majority of the population falls into the average range between 5-6 [33]. The STEN score is based on the transformation of the Z-score, along with the standard deviation of 2 and mean of 5.5 [34]. All the personality sub-traits were grouped for each region and the average STEN score was calculated for each personality trait using the sub-trait, after calculation of Z-score based on the average of personality trait (personality factor e.g Openness) calculated (e.g average of the 5 questions for openness factor was taken and then average value was used for calculating the z score), the STEN score for each personality trait using the formula 1

$$STEN = (Z * 2) + 5.5 \quad (1)$$

The STEN score for each country can be studied using the bar chart plot shown in figure 3. A country-wise comparison for each personality trait gives a high level personality trait comparison and reflects the effect of cultural differences on personality traits. Location based differences around the world can be studied in greater depth using grouping of personality traits. In figure 3 distribution of personality factors STEN score is shown for each country for e.g in Australia high score for Agreeableness and Neuroticism are captured, medium score ranks are seen for Openness and Conscientiousness and low score rank is seen for Extroversion, which gives an overall sample reflection of personality attributes in Australia.

H. Methods and Techniques

In the following experiments the full 63969 data points were used. Clustering techniques were used for the purpose

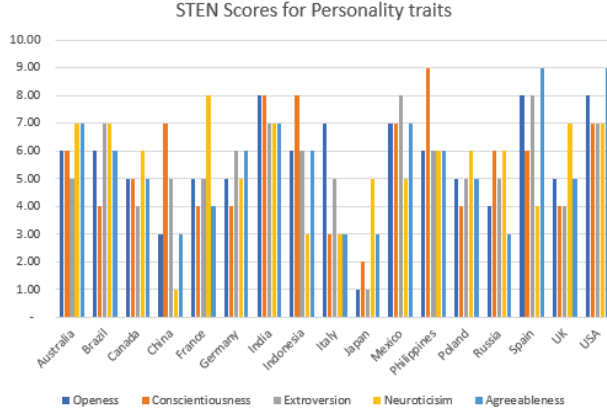


Fig. 3. STEN scores for each personality trait (X - Country wise comparison, Y - STEN scale values 1-10)

of segmenting data points and extracting centroid information to be explored. Model exploration and building of distance functions that can help to segment a combination of user attributes in a distinct manner is described in this section. A total of 17 different user related features were used with three different clustering techniques to explore the compactness and distinctiveness of clusters. The optimal value for finding the number of clusters was found using elbow graphs, Akaike Information Criteria (AIC), and Bayesian Information Criteria (BIC) scores. Cluster validity was tested using different measures such as silhouette score and completeness score. The overall methodology was split into two phases: in phase one, personality trait features, purchase consumer factors, age, gender features were tested with simple K-means clustering. After testing with simple K-means, all the limitations from the algorithm were removed by building a distance function comprising of a weighted linear sum of features which included a location-based distance function and the ability to handle the one hot encoded variables. In phase two, an advanced feature weight handling technique known as Gaussian Mixture Models was implemented to compare the results with K-means. Each phase is described as follows:

1) *Phase 1:* In III-H1 of our proposed approach we utilise the K-means clustering algorithm [35] but with a novel distance function which comprises of a linearly weighted sum of various feature dis-similarities. Clustering on Geo-spatial points is popular in many navigational, traveling and social media platforms. In [36] (where it was referred to as multi-reference clustering) it was stated that the simple Euclidean distance function cannot be used for location points as it does not take into consideration, distance across the curvature of the earth's surface. Studies using K nearest neighbour [37] [38] have also been linked with location-based clustering. The surface-based distance between two locations (described via longitude and latitude) is calculated using the haversine distance formula, which takes into consideration the earth's radius III-H1. Study [39] provides a detailed comparison

between K-means and DBSCAN clustering techniques using distance measures like Euclidean, haversine and Hausdorff on data sets including distance points. Studies [39] [40] state that haversine distance used with K-means clustering outperforms both Euclidean and Hausdorff distance measures. We used the haversine distance formula for the location-based attributes and combined it with linear distance function which calculates the distance between two points using equation 3. Weight values from [0.1 - 1.0] were applied to all the calculated distance value parameters. In phase two data points normalization functions were implemented from scratch, one which can just normalize the data points between 0 and 1, second a re-scaling function was applied to the feature location distance which used 12756 as the maximum(longest distance in km between two points on earth). 0 was the minimum distance between two points so that the distance maximum and minimum value can be between 0 and 1. Experiments were conducted by balancing weights and varying weights on different features to see the effects on cluster compactness and diversion in terms of location, personality, consumer factors. This equation 3 enables us to combine all the personality based features(including all features in survey) from different locations of the world and also to balance the feature importance while segmenting at the same time.

$$D = 2r \sin^{-1} \left(\sqrt{\sin^2 \left(\frac{\phi_2 - \phi_1}{2} \right) + \cos(\phi_1) \cos(\phi_2) \sin^2 \left(\frac{\lambda_2 - \lambda_1}{2} \right)} \right) \quad (2)$$

Haversine Formula Equation, ϕ_1, ϕ_2 is longitude of two points λ_1, λ_2 is the latitude of two points

$$\begin{aligned} \text{Linear Weighted Distance} = & \sum_{j=1}^{n=5} W_1(Px_n - Py_n) + \\ & \sum_{j=1}^{n=3} W_2(Fx_n - Fy_n) + \\ & \sum_{j=3}^{m=9} W_k(Vx_m - Vy_m) \\ & + W_{10}(\text{re-scaled location distance}) \end{aligned} \quad (3)$$

Equation 4: W1= Weights(0.1-1.0), P = Personality Trait Factor items , F = Purchase Consumer Factors, V = Value Added Features with all other feature, re-scaled location distance = distance between two points using haversine distance , after being re-scaled with separate location distance re-scaling function

2) *Phase 2:* This distinctive approach introduced in phase 2 for creating location-based profiling clusters using customized distance function, was compared with the clusters produced using the Gaussian Mixture Model technique in III-H2. The Gaussian mixture model (GMM) works on the concept of

learning definable mixture models from data in an unsupervised learning way [41]. GMM tries to find clusters with the same technique as simple K-means but using a probabilistic model with probabilistic cluster assignments using a weighting technique and normalization of input data [42]. GMM provides the ability to control the degree of freedom for assigning the cluster shape as it allows us to cluster in spherical and diagonal shapes and provides the capability for density estimation which can help to understand the distribution of the data.

IV. EXPERIMENTATION AND RESULTS

This section describes a series of experiments conducted in two different phases for building actionable personas. First, preliminary experiments (IV-A) were conducted on a limited set of features using simple K-means clustering (Euclidean distance) to explore the feature space. We also conduct k-means clustering experiments using the linearly weighted distance function (equation 3) and investigate different feature weights. In IV-B, we apply the Gaussian Mixture Models and provide a comparison with results generated in IV-A. In these experiments, feature variables were converted into numerical features as discussed in Section III-E and III-F. Clusters resulting from IV-A and IV-B results were explored using simple data analysis and visualization techniques in python, and visualization graphs from Microsoft Power BI. The Silhouette score and Calinski Harabasz index were used to check the quality of the clusters generated with each particular technique. The Completeness score for the overall clusters was not used as there were no labels assigned to data points in the data set. Cluster results produced were used to label the particular data point profile. The resulting clusters were developed in to distinct groupings with different personality traits and behaviours. These groupings will enable companies to understand specific ways of effectively marketing to their client base.

A. Phase 1

Empirical experiments were first conducted using K-means clustering algorithm using the Euclidean distance formula. A feature set of 15 variables, were supplied by The Insights Family [43] The preliminary experiments using K-means using Euclidean distance was to find the optimal k value (number of clusters) in the data set. This k value will be used in generation of results for both IV-A and IV-B.

An elbow graph using Sum of square errors (SSE) on the y axis and number of clusters on the x axis was established the optimal value of k clusters. The average Silhouette score for clusters was 0.1456 whereas Calinski Harabasz index was 16653.321. The average silhouette score and Calinski Harabasz index were explored with a k value of 4 as well but was disregarded due to more optimal results with using the value k=5 clusters.

Further experimentation, with a linearly weighted distance (3) function with a k value of 5 using all features from the pre-processing stage was conducted. Different weight combinations were applied to features and the impact on the cluster

compactness (silhouette score) evaluated. Weights applied had ranges between [0.1 .. 1.0], and the top 10 iterations which had silhouette scores of more than 0.3 are shown in the table I, A maximum compactness score (silhouette score) was recorded at 0.61 for various iterations. The most prominent effects on silhouette score scale and geographical spreading of location data points was found when relatively high weights were associated with personality and purchase consumer features. The weights used were described in I. Location distance was allocated the lowest weight, as when higher weights were assigned clusters became dominated by the location feature (weighting the scaled haversine distance, see section IV.). A total of 100 iterations were performed with sets of different weights, however in our results table I section we will show only the 10 iterations with the most varied (different silhouette scores) results. When higher weights were applied to age and gender an increase in the silhouette score was noted. Whereas, if a higher weight is applied to age only, it negatively affected the silhouette score of the overall clustering. An average affect was seen when a combination of weights was applied with value added features (i.e. user-based features such as number of family members, family income, currently living with etc.). Throughout the experiments K was consistently set to 5. Results for the clusters formed were explored in two different approaches which consisted of measuring the silhouette score on a scale of 0 to 1, when different weight combinations are applied. we plotted the data points, for each cluster, on a geographic map to visualise cultural influences on personas. As well as an optimal silhouette score a contribution from all features in the data set is also desirable (allowing customers of a future product to explore all dimensions of the data according to their particular requirements). As an illustration for this paper, iteration 9 was selected for analysis. In this example, a mixture of weights from 0.4 to 0.9 to all features were applied to all the features of categorical distribution and high weights on personality and purchase consumer feature as we wanted our clusters to be more personality and purchase consumer focused. Analysis of iteration 9 clusters was carried out by checking each of the variable dimensions of the features involved in clustering, we started analysing our clusters by plotting the location of the profiles on a geographical map in figure 5, Figure 5b shows a second geographical plot displaying the clusters using k-means with linearly weighted distance whereas Fig. 5c shows the clusters for simple k-means with euclidean distance. Different colors were used to demonstrate the grouping of clusters in these figures. For k-means (linearly weighted distance) the clusters can be seen to be well segmented geographically, for e.g cluster 0 covers all the data points from America and Canada, Mexico and parts of France, cluster 2 covers South America and parts of France, cluster 3 covers UK, Europe, and cluster 4 covers Russia and Asia. Cluster 1 covers parts of Australia, Indonesia and surrounding areas. This geographical (cultural) effect on clustering is clearly due to the inclusion of the haversine distance within our distance formula (see equation 2). Simple k-means (Euclidean) is shown in the third plot of 5 here,

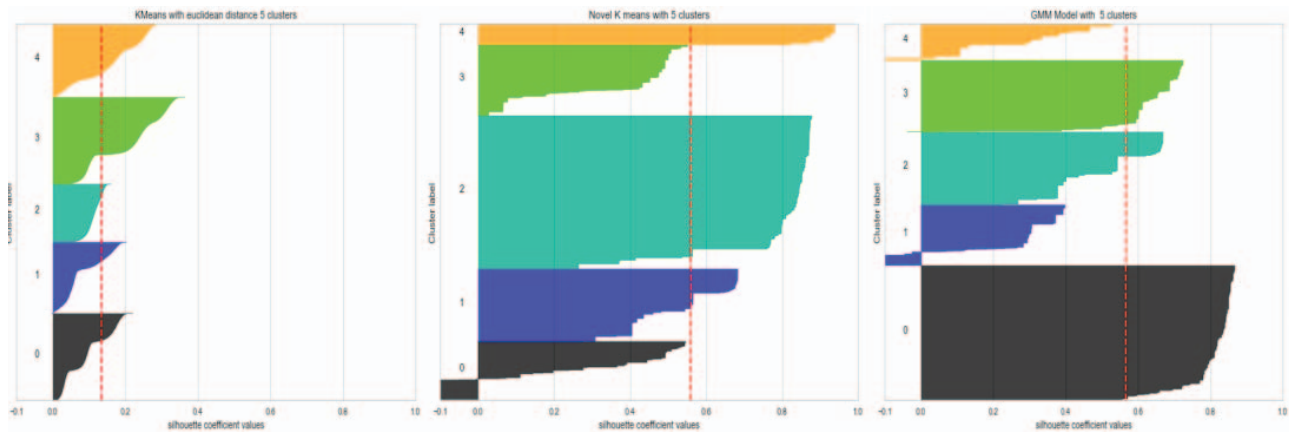


Fig. 4. Silhouette Score scale comparison for different methods applied

we can see very geographically (culturally) mixed clusters. Distribution of data points for cluster 0 was 18.13% , 26.05% for cluster 1, 6.10% for cluster 2, 33.70% for cluster 3 and 16.02% for cluster 4. Overall distribution of data points for clusters was dependent on the number of locations covered, age and gender was equally distributed for all the clusters according to the number of respondents covered in each cluster. Utilising location distance as a point of reference for the clustering segmentation we started exploring the second dimension of Personality features (the 5 averaged personality OCEAN traits). For each cluster (geographically, culturally distinct) we can now examine the associated personality traits. STEN scores were calculated from the 5 averaged personality features for each cluster. This ranks the clusters based on personality traits (OCEAN model). Results for cluster 0 showed a high average STEN score value for neuroticism (in the range 5.5-6.2) for France and Canada regions whereas agreeableness average STEN score value was recorded highest for US and Mexico regions, and openness was recorded average for all the regions between 4.4-5.5. Cluster 1 showed highest neurotic average STEN scores between range of 5.2-5.9 was seen for Indonesia, Japan, Philippines, Australia and lowest for China with average STEN score value of 4.5. Agreeableness average STEN score highest value was seen for Australia, Indonesia and Philippines with range between 5.2-5.9. Similarly extroversion highest values were recorded between 5.9-6.1 for the same regions as of agreeableness, Openess and Consciousness least value was recorded for Japan whereas for all other regions it was average between 5.4-6.1. Cluster 2 consisted of some particular regions of France and Brazil where neuroticisim average score was seen to have the highest average STEN score, France region had somewhat similar personality trait STEN score as represented in the other clusters whereas for Brazil average scores between 5.1-5.6 was seen for all the personality traits. In cluster 3 France personality traits showed somewhat similar behaviour having high neuroticisim score and average all other scores, Agreeableness highest average STEN score was seen for Spain

and Germany regions whereas for openness highest average STEN scores were seen for Italy and Spain. In cluster 4 France represented similar behaviour for personality like it did in other clusters having high Neuroticisim average score whereas in this cluster extroversion average STEN score for France was high as compared to other clusters, Japan along with France showed similar behaviour by having neuroticism and openness STEN scores.

Clusters representing the different location segments were then checked for the three Purchase consumer factors discussed in the pre-processing stage (III-E). Cluster 0 covering locations of France, Canada, US, Mexico showed that US and Canada respondents are more interested in products that are New and Novel, whereas France region respondents showed an average interest for all the purchase consumer factors. Cluster 1 showed that Australia, Indonesia and China respondents are more interested in New/ Novel products along with having the self-conscious factor "convincing/ about me". Japan respondents in this cluster showed that an average response for purchase consumer factors whilst Philippines respondents in this cluster showed more interest in the Convincing factor. Cluster 2 covering locations of France and Brazil consisted of average responses for all the three factors. Cluster 3 respondents showed that kids in Poland, UK and Italy are more interested in New and Novel products along with the Convincing factor, whereas kids respondents from Spain, France, and Germany are more interested in products which have convincing factor in them (for e.g products with reasonable price and good amount of features in them). In cluster 5 respondents from China showed a varied response from cluster 1. For example, Chinese respondents were found to be more interested in the Convincing factor involved products. Respondents from Russia showed that they are equally interested in products that are New/ Novel and have convincing factor in them, similarly respondents from India and Indonesia showed highest values for the New/ Novel factor across the entire dataset

Each of the region based clusters represented a mixture of common favourite hobbies, educational subjects and other

features used in the development of the clustering model. Using this analysis, we built a small use case, to explore how the personas or clusters which we have built could be used by clients by looking at and interpreting the visualisations. For our use case a Digital Industry client (movie and television show producers) wants to build a Digital Education show and where they want to have a target audience from many different regions of the world on one single platform. To provide them with a solution we started by exploring the audience for each cluster to see what are their hobbies, their income, their favourite subject, family income etc. Each persona, represented some highlights of a region on a whole for e.g Persona (cluster 0) 1 showed that people in USA, Mexico, Canada, parts of France (near that region) had Medium income family where the respondents mostly lived with their Mum and Younger brother or sister, their top three favourite hobbies were: Gaming, Indoor/Outdoor activities, Arts and Crafts, their top three favourite subjects were Computing, Creative Arts and Science. Similarly Persona 4 which had regions of Asia, Russia, India had mostly Medium and High Income Families whereas respondents three most favourite subjects were Maths, English, Science, their three most favourite hobbies were same as Persona 1 (cluster 0) but their favourite subjects were different. The provided information can be utilised by the Digital Education shows producers to understand which areas or topics will resonate most effectively with their audience as well as enabling them to understand the type psychological behaviours and motivations of the audience to make sure that any marketing campaigns, language used, and tone of the show is most effective with a large emphasis on how there are similarities and difference between the regions.

TABLE I
RESULTS FOR LINEAR WEIGHTED DISTANCE MEASURE USED WITH
K-MEANS.

Iteration	1	2	3	4	5	6	7	8	9	10
Personality Features	1.0	0.8	0.5	0.4	0.3	0.6	0.9	0.7	0.8	0.9
Purchase Consumer	0.9	0.7	0.5	0.4	0.3	0.6	0.8	0.7	0.8	0.9
Family Income	0.6	0.5	0.4	0.6	0.4	0.4	0.4	0.5	0.6	0.4
Living With	0.6	0.5	0.4	0.6	0.4	0.4	0.4	0.5	0.6	0.4
Family Members	0.6	0.5	0.4	0.6	0.4	0.4	0.4	0.5	0.6	0.4
Age	0.6	0.4	0.3	0.4	0.6	0.4	0.5	0.6	0.4	0.4
Gender	0.9	0.4	0.3	0.4	0.7	0.4	0.6	0.7	0.4	0.3
Favourite Hobby	0.8	0.8	0.6	0.5	0.4	1.0	0.6	0.8	0.9	0.8
Favourite Subject	0.8	0.8	0.6	0.5	0.4	1.0	0.6	0.8	0.9	0.8
Location Distance	0.3	0.4	0.3	0.3	0.2	0.3	0.4	0.5	0.4	0.5
Silhouette Score	0.36	0.56	0.61	0.59	0.56	0.61	0.53	0.59	0.57	0.61

B. Phase 2

In Phase 2, clusters were generated using Gaussian Mixture Models (GMM) on the same pre-processed features as in the K-means using Linear Weighted Distance (LWD). This technique checked the similarity of clusters generated from the K-means with LWD technique. Experiments conducted in phase one (K-means with LWD) provided the mechanism to weight the features for the clustering technique whereas in GMM all the weights were balanced by the algorithm automatically, where the covariance and mean are used to

define the shape of each cluster. Clusters generated from GMM had a silhouette score of 0.54 which showed that GMM clusters were not separated as well as the K-means with Linear Weighted Distance clusters. This was one of the limitations of GMM that we were not able to increase or decrease the silhouette score as we could in Phase 1. The density of data points according to the location of the respondents was also considered an important factor. Personality factor attributes and Purchase consumer attributes helped in analysing the clusters for account of geographic differences as it did in Phase 1 results but were just used for comparison purpose.

As can be seen through comparison of fig 5 A and fig 5 B, cluster results for North America, Western Europe and South America were very similar for both K-means with LWD (phase 1) and GMM (phase 2). However major differences can be seen in the clusters found for Asia and other regions of the world. K-means with LWD is heavily location dependent so the results are expected to reflect geographical cultural differences, whereas GMM clusters are density based so cultural differences within a cluster may be more diverse with this approach. For example see cluster 4 in GMM (5) which encompasses a slice of the map from Northern Russia though India and into SE Asia and Australia, grouping all these respondents into a single cluster. Cluster 0 and 4 results from GMM showed some real highlights as it considered density of data points from a location prospective.

Personality factors and all other attributes were dependent on the density of the clusters. Clusters covering a greater number of data points had variable averages of both personality and purchase consumer features, along with a greater distribution of favourite hobbies and subjects. GMM clusters showed us the proof that clusters created were mostly location dependent as both approaches results for location aspect was the starting point for exploration.

V. DISCUSSION

Experiments conducted in the two different phases all contribute towards segmentation of personality profiles from different regions of the world. Comparison of all the three technique (first Gaussian Mixture Models Clusters silhouette, second K-means with Linear Weighted Distance, third K-means with euclidean distance(basic)) applied in two different Phases is shown using a silhouette measure plot in figure 4. Clustering on these personality profiles with different likes and interests around the globe can help us to understand the groups of people with common likes, dislikes and personality attributes. An optimal grouping technique to segment the profile features has been a challenge (simple K-means was not sufficient) as we wanted to cluster the data points based on relative importance of features. Results from phase 1 in which we used the Linear Weighted Distance function with K-means helped us to segment profiles of people from different locations around the world by considering and changing the weights on different features based on there importance. Usage of these weights helped us to prove that the segmentation can be done in a unique way in which we can have the

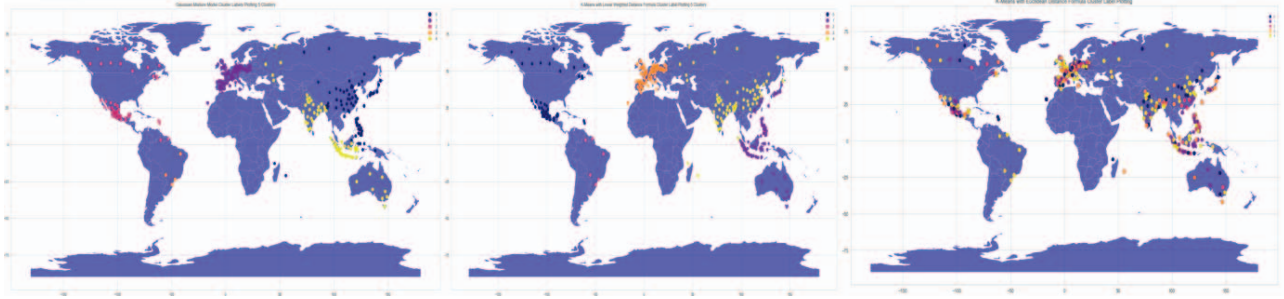


Fig. 5. Map based comparison of all the three techniques, A: Gaussian Mixture Models clusters on geographical map, B: K-means with Linear Weighted Distance on geographical map, C: K-means with euclidean distance

ability to increase and decrease the importance of features in segmentation. GMM results were very promising but had the limitation of the internal weighting technique (performed by the algorithm itself) as it assumes that all the features are in the gaussian space and tries to balance with same set of mixture weights [44]. K-means clustering with linear weighted distance results showed more promising results (where we can empirically increase and decrease the importance of feature in cluster) as it gave us the functionality to balance the weights of the features and optimise the silhouette score (can be seen in section B of Experiments and Results Table I). In Phase 1, cluster results when all the features were given weights closer to 1.0, the silhouette score decreases whereas when weights of more than 0.4 were applied to all features, the silhouette score was observed to be more than 0.50. Similarly when weights were changed for personality, purchase consumer and favourite hobby and subject, optimally tends to increase and decrease by 0.2-0.4. Overall results associated with the Linear Weighted Distance function and K-means were somehow similar to Gaussian Mixture Model results but had the advantage of tuning the weights according to requirements (a user may want to focus on particular features). Results generated from K-means with the Linear Weighted Distance method seemed more feasible to build the personas as in some personas we can increase the weight of some features whereas in some we can just decrease them, along with ability to introduce more features. The user case of personas defined in IV-A it demonstrates that the results could be used to increase the understanding of a target audience. The results from the GMM experiments can be used to build personas. However, when fewer features are available with fewer respondents, cluster results from GMM have been shown to be biased towards particular features. In K-means with Euclidean distance, the formula results can only be used to check the number of patterns in that data set and cannot be used to produce personas as the Euclidean distance prevents the proper inclusion of location data which was found to be a valuable feature in the results (culturally similar clusters). In phase 2 results we had balanced the weights and defined the weights of location-based distance features which make the features less important in the segmentation. Favourite hobbies and subjects with respect to

each cluster were more associated with the location data point for all results shown in Phase 1 and Phase 2. We concluded from the results of Phase 2 that location, favourite hobbies and subject were the most important variables for Personality profile clustering whereas when we drop these variables from the Phase 2 methodology, the silhouette score obtained is below 0.20 for the clusters formed. In our segmentation method we tried to develop a novel distance function that can help segment personality profiles by using K-means clustering. We further tested the same features we developed with the Gaussian Mixture Model clustering technique which helped us to prove that our novel distance function can help to do segmentation with a combination of feature importance. Our technique for the development of personas is a unique way of giving the voice to respondents from around the world as it can help to highlight the upcoming trends among people from different regions. These personas which we have developed in this study are just in the pilot phase whereas we aim to include many more open text based answers in these segmented profiles in future work which will help to uncover different dimensions of segmentation.

VI. CONCLUSION AND FURTHER WORK

This paper has presented a new methodological way for developing actionable personas through segmentation of personality-based features combined with location and consumer based features extracted from global survey data. A novel distance function to be used with K-means clustering is proposed to handle real time survey based data to provide digital media clients a mechanism to explore the upcoming trends and patterns in communities around the world. Usability of these dynamic and robust personas will help them to generate the content according to people likes and dislikes, e.g popular trends for different hobbies among regions can be used by digital media companies to create content for these regions which will be of direct interest for the respondents and will create high Return on Investment. The results generated from the techniques in IV-A and IV-B provided the generic skeleton of actionable personas for kids. Work is currently ongoing to improve the reliability and robustness of the developed personas by adding in additional survey responses from different questions, with further analysis in to how the

tool can be deployed in real time. This process of ensuring the tool is real-time would involve updating the centroids of the clusters depending on new data provided to the pre-existing model.

VII. ACKNOWLEDGEMENTS

This work has been funded by Innovate UK Knowledge Transfer Partnerships under Grant number 12135.

REFERENCES

- [1] M. Pröbster, M. E. Haque, and N. Marsden, "Perceptions of personas: the role of instructions," in *2018 IEEE International Conference on Engineering, Technology and Innovation (ICE/ITMC)*. IEEE, 2018, pp. 1–8.
- [2] A. Cooper, "The inmates are running the asylum. indianapolis, IA: SAMS."
- [3] J. Pruitt and J. Grudin, "Personas: Practice and theory. proceedings of the 2003 conference on designing for user experiences," *New York*, pp. 1–15, 2003.
- [4] T. Adlin, J. Pruitt, K. Goodwin, C. Hynes, K. McGrane, A. Rosenstein, and M. J. Muller, "Putting personas to work," in *CHI'06 Extended Abstracts on Human Factors in Computing Systems*, 2006, pp. 13–16.
- [5] K. Goodwin, "Perfecting your personas," *Cooper Interaction Design Newsletter*, vol. 19, pp. 295–313, 2001.
- [6] S. Mulder and Z. Yaar, *The user is always right: A practical guide to creating and using personas for the web*. New Riders, 2006.
- [7] F. Durupinar, N. Pelechano, J. Allbeck, U. Güdükbay, and N. I. Badler, "How the ocean personality model affects the perception of crowds," *IEEE Computer Graphics and Applications*, vol. 31, no. 3, pp. 22–31, 2009.
- [8] J. S. Wiggins, *The five-factor model of personality: Theoretical perspectives*. Guilford Press, 1996.
- [9] S. Mullainathan and J. Spiess, "Machine learning: an applied econometric approach," *Journal of Economic Perspectives*, vol. 31, no. 2, pp. 87–106, 2017.
- [10] D. Dzyabura and J. R. Hauser, "Active machine learning for consideration heuristics," *Marketing Science*, vol. 30, no. 5, pp. 801–819, 2011.
- [11] D. G. Smith, Big data gives the "big 5" personality traits a makeover. [Online]. Available: <https://www.scientificamerican.com/article/big-data-gives-the-big-5-personality-traits-a-makeover/>
- [12] A. Goddard, P. Hasking, L. Claes, and P. McEvoy, "Big five personality clusters in relation to nonsuicidal self-injury," vol. 25, no. 3, pp. 390–405, publisher: Routledge _eprint: <https://doi.org/10.1080/13811118.2019.1691099>. [Online]. Available: <https://doi.org/10.1080/13811118.2019.1691099>
- [13] T. J. Reece, "Personality as a gestalt: A cluster analytic approach to the big five," p. 39.
- [14] A. Kerber, M. Roth, and P. Y. Herzberg, "Personality types revisited—a literature-informed and data-driven approach to an integration of prototypical and dimensional constructs of personality description," vol. 16, no. 1, p. e0244849. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7790254/>
- [15] M. Bore, K. R. Laurens, M. J. Hobbs, M. J. Green, S. Tzoumakis, F. Harris, and V. J. Carr, "Item response theory analysis of the big five questionnaire for children—short form (bfc-sf): a self-report measure of personality in children aged 11–12 years," *Journal of personality disorders*, vol. 34, no. 1, pp. 40–63, 2020.
- [16] M. P. Couper, "Technology trends in survey data collection," *Social Science Computer Review*, vol. 23, no. 4, pp. 486–501, 2005.
- [17] K. S. Taber, "The use of cronbach's alpha when developing and reporting research instruments in science education," *Research in science education*, vol. 48, no. 6, pp. 1273–1296, 2018.
- [18] C. Biernacki, G. Celeux, and G. Govaert, "Assessing a mixture model for clustering with the integrated completed likelihood," *IEEE transactions on pattern analysis and machine intelligence*, vol. 22, no. 7, pp. 719–725, 2000.
- [19] A. Revella, *Buyer personas: how to gain insight into your customer's expectations, align your marketing strategies, and win more business*. John Wiley & Sons, 2015.
- [20] D. W. Fiske, "Consistency of the factorial structures of personality ratings from different sources," vol. 44, no. 3, pp. 329–344. [Online]. Available: <http://doi.apa.org/getdoi.cfm?doi=10.1037/h0057198>
- [21] W. T. Norman, "Toward an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personality ratings," vol. 66, no. 6, pp. 574–583, place: US Publisher: American Psychological Association.
- [22] G. M. Smith, "Usefulness of peer ratings of personality in educational research," *Educational and Psychological Measurement*, vol. 27, no. 4, pp. 967–984, 1967. [Online]. Available: <https://doi.org/10.1177/001316446702700445>
- [23] L. R. Goldberg, "An alternative "description of personality": The big-five factor structure," vol. 59, no. 6, pp. 1216–1229, place: US Publisher: American Psychological Association.
- [24] R. R. McCrae, "Cross-cultural research on the five-factor model of personality," vol. 4, no. 4. [Online]. Available: <https://scholarworks.gvsu.edu/orpc/vol4/iss4/1>
- [25] K. R. Shahapure and C. Nicholas, "Cluster quality analysis using silhouette score," in *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 747–748.
- [26] X. Wang and Y. Xu, "An improved index for clustering validation based on silhouette index and calinski-harabasz index," vol. 569, p. 052024.
- [27] B. Jumadi Dehotman Sitompul, O. Salim Sitompul, and P. Sihombing, "Enhancement clustering evaluation result of davis-bouldin index with determining initial centroid of k-means algorithm," vol. 1235, no. 1, p. 012015. [Online]. Available: <https://iopscience.iop.org/article/10.1088/1742-6596/1235/1/012015>
- [28] B. Williams, A. Onsmann, and T. Brown, "Exploratory factor analysis: A five-step guide for novices," *Australasian journal of paramedicine*, vol. 8, no. 3, 2010.
- [29] J. M. Bland and D. G. Altman, "Statistics notes: Cronbach's alpha," *Bmj*, vol. 314, no. 7080, p. 572, 1997.
- [30] J. W. Osborne, "What is rotating in exploratory factor analysis?" *Practical Assessment, Research, and Evaluation*, vol. 20, no. 1, p. 2, 2015.
- [31] J. T. Hancock and T. M. Khoshgohar, "Survey on categorical data for neural networks," *Journal of Big Data*, vol. 7, no. 1, pp. 1–41, 2020.
- [32] J. Eatwell, "Using norms in the interpretations of test results," *Practice issues for clinical and applied psychologists in new Zealand*, pp. 268–276, 1997.
- [33] K. Coaley, *An introduction to psychological assessment and psychometrics*. Sage, 2014.
- [34] M. HOLBROOK and C. E. SKILBECK, "An activities index for use with stroke patients," *Age and ageing*, vol. 12, no. 2, pp. 166–170, 1983.
- [35] D. Arthur and S. Vassilvitskii, "k-means++: The advantages of careful seeding," Stanford, Tech. Rep., 2006.
- [36] Y. Zhong, J. Li, and S. Zhu, "Clustering geospatial data for multiple reference points," *IEEE Access*, vol. 7, pp. 132 423–132 429, 2019.
- [37] N. Du, J. Zhan, M. Zhao, D. Xiao, and Y. Xie, "Spatio-temporal data index model of moving objects on fixed networks using hbase," in *2015 IEEE International Conference on Computational Intelligence & Communication Technology*. IEEE, 2015, pp. 247–251.
- [38] A.-L. Uribe-Hurtado, M. Orozco-Alzate, N. Lopes, and B. Ribeiro, "Gpu-based fast clustering via k-centres and k-nn mode seeking for geospatial industry applications," *Computers in Industry*, vol. 122, p. 103260, 2020.
- [39] S. Sharmila and B. Sabarish, "Analysis of distance measures in spatial trajectory data clustering," in *IOP Conference Series: Materials Science and Engineering*, vol. 1085, no. 1. IOP Publishing, 2021, p. 012021.
- [40] E. Maria, E. Budiman, M. Taruk *et al.*, "Measure distance locating nearest public facilities using haversine and euclidean methods," in *Journal of Physics: Conference Series*, vol. 1450, no. 1. IOP Publishing, 2020, p. 012080.
- [41] M. A. T. Figueiredo and A. K. Jain, "Unsupervised learning of finite mixture models," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 24, no. 3, pp. 381–396, 2002.
- [42] S. R. Bond, A. Hoefler, and J. R. Temple, "Gmm estimation of empirical growth models," *Available at SSRN 290522*, 2001.
- [43] "The Insights Family." [Online]. Available: <https://theinsightsfamily.com>
- [44] L. Si and R. Jin, "Adjusting mixture weights of gaussian mixture model via regularized probabilistic latent semantic analysis," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 2005, pp. 622–631.