Exploring Thematic Travel Preferences of Global Cities through Agglomerative Hierarchical Clustering for Enhanced Travel Recommendations

Soeltan Abdul Ghaffar[1,*], Wilbert Clarence Setiawan[2]

[1]*Department of Marine Information Systems, School of Postgraduate Studies, Universitas Pendidikan Indonesia, Bandung, Indonesia*
[2]*Faculty of Informatics Engineering, Universitas Taruma Negara, Jakarta, Indonesia*

(Received: June 1, 2025; Revised: July 5, 2025; Accepted: October 21, 2025; Available online: December 3, 2025)

Abstract

This study explores the application of Agglomerative Hierarchical Clustering (AHC) to categorize global cities based on thematic travel preferences, aiming to enhance personalized travel recommendations. The dataset used contains travel information for 560 cities worldwide, including thematic ratings across nine categories: culture, adventure, nature, beaches, nightlife, cuisine, wellness, urban, and seclusion, along with climate data and city descriptions. Feature engineering was performed to calculate an overall rating for each city by averaging its thematic scores, and to compute an average annual temperature from monthly climate data. The primary objective of this research was to use AHC to group cities into distinct clusters based on these thematic ratings. The analysis revealed six clusters, each representing different types of travel experiences. Cluster 1 consists of urban cultural hubs with high ratings for culture, cuisine, and urban experiences, while Cluster 2 features cities with a balance of cultural and culinary experiences alongside moderate natural and nightlife attractions. Cluster 3 represents remote, nature-focused cities with high ratings for seclusion and nature. Cluster 4 includes cities renowned for their beaches, nature, and cuisine, while Cluster 5 groups cities that emphasize adventure, nature, and seclusion. Cluster 6 is made up of destinations with a focus on nature, adventure, and seclusion, offering a balance between outdoor activities and tranquility. These findings offer a deeper understanding of the diversity in global city offerings and can significantly improve the effectiveness of travel recommendation systems by aligning cities with users' thematic preferences. By categorizing cities into meaningful clusters, personalized travel suggestions can be made based on users' specific interests, such as cultural exploration, adventure, or nature. This research lays the groundwork for future studies to incorporate additional data sources and explore alternative clustering techniques for even more refined travel recommendations. The practical applications of this research can enhance real-world travel recommendation platforms, making them more tailored and relevant to individual user preferences.

Keywords: Travel Recommendation, Agglomerative Hierarchical Clustering, Thematic Preferences, Clustering Analysis, Personalized Travel

1. Introduction

The growing demand for personalized travel recommendations reflects a broader trend in consumer behavior towards individualized experiences. As travel dynamics evolve, travelers increasingly seek tailored itineraries that resonate with their specific preferences and circumstances, a shift propelled by enhanced living standards and the proliferation of accessible technology [1]. Traditional travel advisement approaches, often reliant on expert opinions, are outpaced by the complexity of modern data sources, which include vast multimedia content from images and videos to geotagged social media posts. This complexity necessitates innovative solutions that effectively harness data mining techniques to filter and interpret user interests, enabling the creation of customized travel recommendations [2]. Personalized tourism has become a prominent focus due to its ability to merge individual traveler preferences with contextual attributes of destinations. A personalized travel recommendation system operates on the principle that understanding an individual's historical preferences and behaviors enables a more accurate portrayal of future travel choices [3]. The recommendation process often incorporates various attributes of destinations, with recent studies highlighting the significance of machine learning in validating user preferences against current trends and historical data [4]. Such

*Corresponding author: Soeltan Abdul Ghaffar (soeltan027ghaffar@gmail.com)
DOI: https://doi.org/10.47738/ijaim.v5i4.111

systems leverage algorithms that adapt dynamically to context, thereby improving the accuracy and relevance of recommendations made to travelers.

A critical component of advancing personalized travel recommendations relies on the categorization of cities based on thematic attributes, which can range from cultural offerings to culinary experiences. The classification of cities via thematic attributes enhances the effectiveness of travel recommendations by allowing for nuanced preferences rather than broad categorizations. Theoretically, this reflects a shifting paradigm where recommendations evolve from mere location suggestions to intricate narratives that encompass integrated experiences tailored to various interests [5]. As cities embody diverse cultural, economic, and geographic elements, a thematic classification would facilitate targeting specific tourist demographic segments, further honing the recommendation process [6]. The need for such categorization underscores the limitations of earlier models that often fail to consider the complexity of city dynamics and the multitude of factors influencing traveler choices. For instance, cities can significantly differ in aspects such as infrastructure, accessibility, and attractions, which all play a role in how travelers perceive and interact with these locations [7]. Hence, robust frameworks employing temporal and spatial factors are essential to accurately represent and adapt travel recommendations, resulting in efficient and personalized travel planning systems [8].

A growing body of research asserts that the integration of personality dimensions into recommendation systems can also address group travel dynamics, providing insights into collective preferences that align with distinct travel motivations [9]. This is particularly relevant given that travel experiences often arise from group decisions, wherein multiple individuals' preferences can conflict. Personality-aware recommendation systems not only enhance individual user experiences but also navigate the complexities of suggesting suitable locations that accommodate diverse preferences within a group context [10]. The rise of advanced algorithms such as neural networks and genetic algorithms evidence that the field is rapidly evolving towards utilizing artificial intelligence to process complex datasets more effectively. Such techniques allow for an adaptive approach to itinerary planning that balances multiple objectives, including cost, duration, and interest diversity, ultimately producing itineraries that align closely with user preferences and expectations [8]. This approach offers a significant advancement over past methodologies, which lacked the necessary capacity to process large-scale data inputs effectively.

Clustering techniques play a pivotal role in enhancing the effectiveness of travel recommendation systems, particularly in organizing and analyzing vast datasets related to user preferences, destinations, and Points of Interest (POIs). As the landscape of travel increasingly shifts towards personalized experiences, categorizing data into meaningful clusters becomes essential for developing insightful and relevant recommendations. The dynamic nature of various factors influencing travel decisions including personal interests, geographical characteristics, and social influences necessitates a structured approach to understanding traveler behaviors and preferences, which can be effectively facilitated through clustering methods [11], [12]. In travel recommendation systems, clustering can serve several key purposes. Firstly, it allows for the identification of similar users or destinations, enabling the generation of relevant suggestions that resonate with a target audience. By employing algorithms such as K-means and Density-Based Spatial Clustering of Applications with Noise (DBSCAN), systems can group users based on their travel behaviors and interests, thus providing more tailored recommendations. This enables platforms to form user segments that share analogous preferences, improving the precision of the recommendations generated [13], [14]. For example, as Sabet et al. [13] explain, their hybrid recommender system utilizes clustering to aggregate users into groups with similar profiles, allowing the system to predict future preferences more accurately, even for new users lacking historical data.

Additionally, clustering facilitates the efficient organization of POIs within destination contexts. By categorizing attractions based on various thematic attributes, such as cultural significance, activity type, or geographic location, recommendation systems can suggest itineraries that align closely with a user's interests. Sun et al. [15] highlight the efficacy of a city-adaptive clustering framework for discovering POIs, emphasizing that clustering techniques can lead to a more nuanced understanding of user engagement with certain attractions and activities. This approach greatly enhances recommendation capacities by identifying patterns within the User-Generated Content (UGC), such as geotagged photos or social media interactions, thus constructing more dynamic and appealing tourism offerings. Moreover, the application of clustering in analyzing user transition patterns underscores its importance in understanding how travelers navigate between different attractions or activities. Sun et al. [16] explore user transition behaviors to create optimized travel route recommendations. Their study illuminates how clusters can encapsulate

common travel routes or experiences shared among groups of users, ultimately refining the sequence in which attractions are suggested. By analyzing how collective trends manifest, these clustering techniques help to align recommendations with the natural flow of tourist behavior, thereby enhancing user satisfaction.

The primary goal of this study is to apply Agglomerative Hierarchical Clustering (AHC) to explore thematic travel preferences of global cities, providing a nuanced understanding of how various cities align with the distinct interests and preferences of travelers. As the tourism sector becomes increasingly competitive and personalized, there is a pressing need to categorize cities in a manner that reflects their appeal to different demographics based on their thematic attributes, such as cultural heritage, outdoor activities, gastronomy, and nightlife. By leveraging AHC, the study aims to identify meaningful clusters of cities that share similar characteristics, which can subsequently inform targeted travel recommendations tailored to specific traveler preferences [17], [18]. AHC is particularly relevant in this context as it enables the creation of a dendrogram, a tree-like diagram that visualizes the hierarchy of clusters, thereby illustrating the relationships between various cities based on their thematic similarities. This method excels in handling the complexity that characterizes global travel preferences, allowing for the integration of varied data sources to reflect a comprehensive picture of what different urban environments offer to potential visitors [19], [20]. Through this process, cities can be compared not just based on quantitative metrics (like visitor numbers or total facilities) but also on qualitative attributes that enhance travelers' decision-making processes.

The scope of this study involves analyzing a dataset containing travel information for 560 cities worldwide, with a focus on thematic ratings such as culture, adventure, nature, beaches, nightlife, cuisine, wellness, and urban aspects. The dataset also includes key features like city descriptions, which provide a brief summary of each city's appeal, as well as climate information, including monthly average, maximum, and minimum temperatures. Additionally, suggested trip durations (e.g., weekend, short trip, long trip) and budget classifications (e.g., budget, mid-range, luxury) are considered to provide a comprehensive understanding of the travel characteristics associated with each city. These features are analyzed to explore and categorize cities based on thematic preferences using hierarchical clustering.

2. Literature Review

2.1. Travel Recommendation Systems

Travel recommendation systems have evolved significantly over the years in response to growing demands for personalized travel experiences. The methodologies employed in these systems can be categorized broadly into Collaborative Filtering (CF), content-based recommendations, and hybrid approaches that leverage both strategies. This review discusses various existing methods and approaches, encompassing their underlying techniques and implementations, to enhance the personalization of travel recommendations for users. One of the primary techniques utilized in travel recommendation systems is collaborative filtering, which leverages the collective preferences of users to suggest destinations or activities. CF methods can be divided into user-based and item-based techniques, enabling systems to predict user preferences by analyzing similarities in past behaviors and ratings from similar users [21], [22]. However, traditional CF approaches often face challenges associated with data sparsity, especially for new users for whom no historic data exists, a problem often referred to as the "cold start" issue [22]. To counteract this, advanced collaborative filtering techniques have been implemented using various algorithms, such as matrix factorization and deep learning frameworks [23], allowing these systems to enhance their predictive accuracy through continuous learning from user interactions and experiences.

On the other hand, content-based recommendation approaches focus on the attributes and features of destinations or activities that resonate with the preferences expressed by users. These approaches filter and recommend destinations based on previously enjoyed trips and inherent characteristics of attractions, considering factors such as location, services provided, and thematic relevance [2], [24], [25]. Such strategies enhance user engagement by ensuring recommendations are closely aligned with specific preferences and interests. Hybrid recommendation systems, which combine CF and content-based methods, have emerged as a potent solution to the inherent limitations each approach possesses. For instance, Fang et al. [21] discuss the development of a deep travel conversational recommender system that integrates knowledge graphs to enhance recommendation accuracy, thereby capitalizing on the strengths of both

CF and content-based strategies. This fusion allows for a more nuanced understanding of user preferences, as it utilizes both user data and the intrinsic attributes of travel options to deliver superior recommendation outcomes.

2.2. Clustering Techniques in Tourism

Clustering techniques are increasingly recognized as powerful tools in the tourism sector, especially in the development and refinement of travel recommendation systems. Among the various clustering methods, K-means and hierarchical clustering have gained significant attention for their applications in understanding user preferences, segmenting markets, and enhancing destination management. K-means clustering is a non-hierarchical approach widely utilized in tourism studies, primarily because of its simplicity and efficiency. It partitions the dataset into K distinct clusters based on the similarity of data points, minimizing the variance within each cluster while maximizing the variance between clusters. This method allows operators to categorize tourists according to shared attributes such as preferences, demographics, and behaviors, thereby facilitating targeted marketing strategies and personalized recommendations [13], [26]. For instance, Sabet et al. [13] demonstrate how K-means clustering can effectively create user profiles to enhance personalized travel recommendations, proving beneficial for both users and service providers. However, the requirement to predefine the number of clusters poses a challenge, as this often hinges on subjective judgment or prior hierarchical insights [27].

Hierarchical clustering, in contrast, either builds a hierarchy of clusters (agglomerative) or divides a dataset into successively smaller clusters (divisive). This method is particularly useful for its ability to provide a comprehensive view of data relations through dendrogram representations, which can be instrumental in tourism for visualizing groupings of attractions, hotels, or user preferences in a spatial context [28]. Furthermore, as Ernst and Dolničar [29] highlight, hierarchical methods can help mitigate the randomness often present in market segmentation studies by ensuring a more systematic approach to clustering, which is crucial for reliable recommendations. The flexibility of hierarchical clustering allows users to decide on an appropriate cut-off point in the hierarchy to identify significant groupings relevant to their specific marketing or strategic endeavors. The synergy between K-means and hierarchical clustering can produce even more robust insights. A common approach involves initially employing hierarchical clustering to establish an optimal number of clusters and subsequently applying K-means clustering for refined segmentation within those established clusters [30]. This hybrid application improves the accuracy of group definitions and can yield substantial benefits in understanding traveler behaviors and preferences across different contexts.

2.3. Agglomerative Hierarchical Clustering

AHC is a widely used clustering algorithm in data analysis that follows a bottom-up approach to form clusters. It begins with each data point as an individual cluster and iteratively merges them based on their similarities until a single cluster containing all data points is formed or until a predefined number of clusters is reached. The key steps in AHC include the initialization of data points as separate clusters, the calculation of the distance (or similarity) between clusters, and the subsequent merging of the closest clusters until the desired clustering structure is achieved [31], [32]. One of the major advantages of AHC is its intuitive and easily interpretable output, which is often visualized through a dendrogram a tree-like diagram that illustrates the hierarchical relationships among the clusters. This visualization allows users to determine the optimal number of clusters and view how data points are grouped based on levels of similarity [33], [34]. Additionally, the dendrogram provides insights into the clustering process, enabling researchers or analysts to make informed decisions regarding the underlying relationships within the data.

Another advantage of AHC is its flexibility regarding distance metrics. Users can employ various linkage criteria, such as single linkage, complete linkage, average linkage, or Ward's method, depending on the context of the data and the nature of the analysis being performed [32], [35]. This adaptability allows AHC to cater to a wide range of applications across different domains, such as environmental science, market segmentation, and social network analysis. Moreover, AHC is well-suited for small to moderately sized datasets; it retains computational stability and can produce high-quality clusters even in the presence of noise or outliers [36], [37]. Since it does not require the number of clusters to be predetermined, it can be especially advantageous in exploratory data analysis contexts where the ideal number of clusters is unknown [31], [38].

 ISSN 2776-8007

2.4. Relevant Formula

In the context of AHC, various distance metrics can be employed to measure the closeness of data points, enabling effective clustering. One of the most commonly used metrics is the Euclidean distance, which quantifies the straight-line distance between two points in a multi-dimensional space. The formula for calculating the Euclidean distance D(i, j) between two data points (i) and (j), each with (n) dimensions, is expressed as follows:

$$D(i,j) = \sqrt{\sum k = 1^{n} \left(x_{ik} - x_{jk}\right)^{2}} \quad (1)$$

Where D(i, j) is the Euclidean distance between the two points, x_{ik} and x_{jk} are the k-th components of the vectors (i) and (j) respectively and n is the total number of dimensions (or features) of the data points. The use of Euclidean distance in AHC has several advantages, one being its straightforward geometric interpretation, which often resonates well with intuitive clustering methods. It allows for easy identification of clusters based on proximity in the feature space, making it a suitable choice for many clustering applications, particularly in contexts where spatial relationships are paramount [39], [40].

In addition to Euclidean distance, other distance metrics may also be employed within AHC frameworks to enhance clustering effectiveness depending on the data characteristics and contextual requirements. For example, the Mahalanobis distance, which accounts for the correlations between variables, can be particularly beneficial in scenarios with multi-collinearity among features. This metric is effective for identifying outliers and is often used to measure distances effectively in high-dimensional spaces [41], [42]. As noted by Sammour and Othman [43], the choice of distance metric significantly affects the performance of clustering algorithms, underscoring the importance of selecting the most appropriate metric according to the nature of the dataset.

2.5. Related Work

Recent research has demonstrated various applications of clustering techniques to analyze cities and regions based on thematic, geographical, and tourist-related data. This body of work highlights how such methodologies can inform better tourism strategies, regional development, and traveler insights. One notable study by Wang et al. [44] explores tourist distributions and sentiment variations through the analysis of social media data from scenic areas in China. They found that tourist sentiment and distribution patterns are significantly influenced by seasonal factors, suggesting that clustering cities based on temporal data can highlight seasonal tourist behaviors and improve the management of tourism activities in different seasons [44]. This finding aligns with the general trend in tourism research, where understanding the nuances of seasonal variation helps stakeholders optimize marketing strategies and resource allocation.

In an economic context, Gan et al. [45] examined the intensities of tourism economic linkages within Chinese land border cities, employing clustering to establish tourism economic cooperation circles. Their research emphasizes how geographical proximity and collaborative linkages among border cities can enhance tourism economic benefits [45]. By clustering these cities, the authors provide a framework for developing spatial cooperation models that encourage regional economic integration, thereby optimizing tourism-related resource utilization. Similarly, the work of Zhang et al. [46] discusses the spatiotemporal distribution of tourism across various Chinese cities, identifying socioeconomic and environmental impacts on local tourism. The use of clustering methodologies in their research helped produce maps indicating the spatiotemporal trends in tourism revenue from 2008 to 2017, allowing local authorities to strategize tourism developments based on influential local factors [46]. This finding underscores the utility of clustering for visualizing complex datasets related to tourism dynamics, enabling policymakers to develop more focused tourism strategies.

3. Methodology

This section outlines the methodology used to perform Exploratory Data Analysis (EDA) and hierarchical clustering on a dataset of worldwide travel cities.

3.1. Dataset Description and Feature Engineering

The dataset used in this analysis contains information about 560 cities worldwide, including metadata such as city names, countries, and regions, as well as subjective ratings across various thematic categories like culture, adventure, nature, and cuisine. The dataset also includes climate data, specifically monthly average temperatures (in JSON format) for each city. The first step was to load the dataset into a Pandas DataFrame using the pd.read_csv() function. This allows for efficient manipulation and analysis of the data. For feature engineering, two new features were created. The Overall Rating for each city was calculated by averaging the ratings from the thematic features: culture, adventure, nature, beaches, nightlife, cuisine, wellness, urban, and seclusion. This provides an overall indicator of the city's appeal across various categories. Additionally, the Average Temperature (°C) was derived from the monthly average temperatures in the avg_temp_monthly column, which contains the temperature data in JSON format. This was achieved using a custom calculate_avg_temp() function that parses the JSON string and computes the mean of the average monthly temperatures.

3.2. Exploratory Data Analysis (EDA)

EDA was performed to better understand the dataset before proceeding to clustering. The describe() method in Pandas was used to generate statistical summaries for key numerical features, including the Overall Rating, Cuisine, and Culture ratings, as well as the newly engineered Average Temperature feature. This helped in identifying any potential outliers or patterns in the data. Additionally, the first few rows of the dataset with the newly created features were displayed using head(), providing an initial view of the data and allowing us to inspect the values. Furthermore, visualizations were generated using Matplotlib and Seaborn to better understand the distribution of key numerical features. This included histograms or boxplots to visually assess the distribution and relationships among ratings, allowing for a better understanding of the thematic categories before applying the clustering algorithm.

3.3. Clustering and Data Preprocessing

For the clustering analysis, the thematic features selected for the clustering process were: culture, adventure, nature, beaches, nightlife, cuisine, wellness, urban, and seclusion. These features were chosen because they provide a comprehensive view of a city's appeal across different types of travel interests. The first step was data preprocessing, which involved handling any missing values using the dropna() method. After removing missing data, the selected features were scaled using StandardScaler from the sklearn.preprocessing module. This is a crucial step in clustering, as it ensures that each feature contributes equally to the clustering process by standardizing them to have a mean of 0 and a standard deviation of 1. The scaling process was achieved using the scaler.fit_transform(features) method, which scales the data while retaining the relationships between the cities. The preprocessed data, now standardized, was ready for clustering. AHC, a form of bottom-up clustering, was applied using the linkage() function from the scipy.cluster.hierarchy module. The Ward's method was used as the linkage criterion, which minimizes the variance within each cluster by merging clusters with the smallest increase in the sum of squared errors. This method is particularly effective when dealing with continuous data like thematic ratings. The result of the linkage() function is a hierarchical cluster tree, which forms the basis of the next analysis step.

3.4. Agglomerative Hierarchical Clustering

AHC was performed using the linkage() function with the Ward’s method, which is an efficient and popular linkage method for minimizing within-cluster variance. The hierarchical clustering process was visualized using a dendrogram, generated with the dendrogram() function from the scipy.cluster.hierarchy module. A dendrogram is a tree-like diagram that shows the arrangement of clusters based on their distances. The distance_sort='descending' parameter ensured that the dendrogram displayed clusters starting from the most distinct, and the leaf_font_size=8 ensured the city names were readable despite the large number of cities in the dataset. The cut-off line for forming clusters was set at a height of 12 on the dendrogram, which was visually selected based on the tree structure. This cut-off corresponds to six

 ISSN 2776-8007

clusters, determined by setting num_clusters = 6 in the fcluster() function, which assigns cluster labels to each city based on the cut-off. The maxclust criterion ensures that exactly 6 clusters are formed, and these labels are added to the original dataset for further analysis.

3.5. Cluster Analysis

Once the cities were assigned to clusters, an in-depth analysis was performed to understand the characteristics of each cluster. The groupby() method in Pandas was used to calculate the mean thematic ratings for each cluster. These means were then displayed in a summary table, showing the average rating for each thematic feature in each cluster. Additionally, the top three characteristics (the highest ratings) of each cluster were identified by selecting the largest values using the nlargest(3) method. This analysis allowed for a deeper understanding of what each cluster represents in terms of thematic travel preferences (e.g., cities with a high focus on adventure, wellness, or culture). A sample of the cities in each cluster was displayed to give a tangible sense of the cities grouped together. In addition to the statistical analysis, a detailed breakdown of each cluster was provided, showing the cities within each group and identifying the dominant themes. This helped in interpreting the clusters, such as labeling them as "Adventure-oriented cities," "Cultural destinations," or "Beach resorts," depending on the most significant thematic scores.

4. Results

This section presents the results from the EDA, clustering, and cluster analysis performed on the dataset of worldwide travel cities.

4.1. Feature Engineering & EDA Results

The feature engineering process resulted in the creation of two new features: the Overall Rating and Average Temperature (°C) for each city. The Overall Rating was computed by averaging the individual thematic ratings, while the Average Temperature was derived from the monthly temperature data provided in JSON format. The first few rows of the dataset showed the newly added features for each city, such as Milan (Italy), Yasawa Islands (Fiji), and Whistler (Canada), with respective Overall Ratings and Average Temperatures. The statistical summary of key numerical features showed that the Overall Rating had a mean of 3.24, indicating that most cities have moderate appeal across the different thematic categories. The Average Temperature ranged from -4.1°C to 29.4°C, with an average of 17.97°C, reflecting the diversity in the climates of the cities. This provided a good foundation for understanding the dataset's spread and variability in the thematic and climate characteristics of cities.

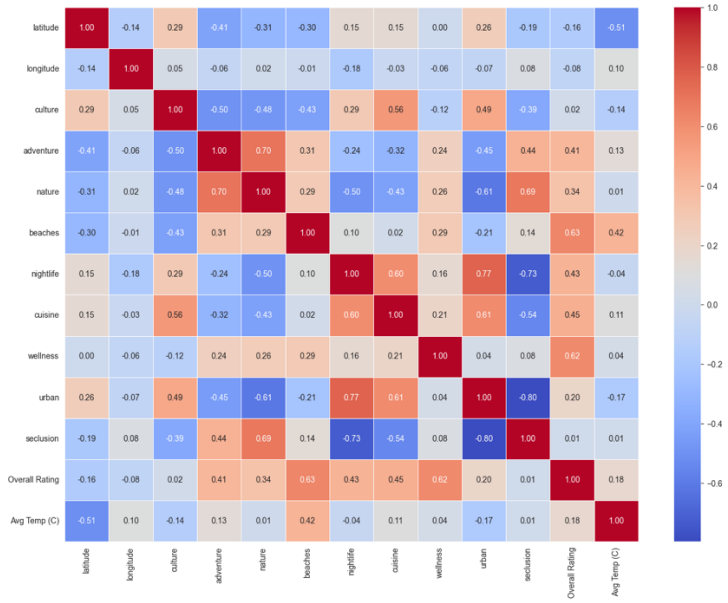

Figure 1. Correlation Matrix of Numerical Features

 ISSN 2776-8007

Figure 1 provides a comprehensive overview of the linear relationships between the different thematic travel ratings in the dataset. The color scale indicates the strength and direction of the correlation, with warm colors (like red) representing a positive correlation and cool colors (like blue) representing a negative correlation. A value of 1.00 signifies a perfect positive relationship, while -1.00 signifies a perfect negative relationship. For instance, the strong positive correlation of 0.70 between 'adventure' and 'nature' suggests that cities rated highly for adventure also tend to be rated highly for nature. Conversely, the strong negative correlations between 'seclusion' and 'urban' (-0.80) and 'seclusion' and 'nightlife' (-0.73) indicate that cities offering a high degree of seclusion are very unlikely to be major urban centers or have vibrant nightlife. This analysis is crucial for understanding the inherent trade-offs and associations between different travel preferences.

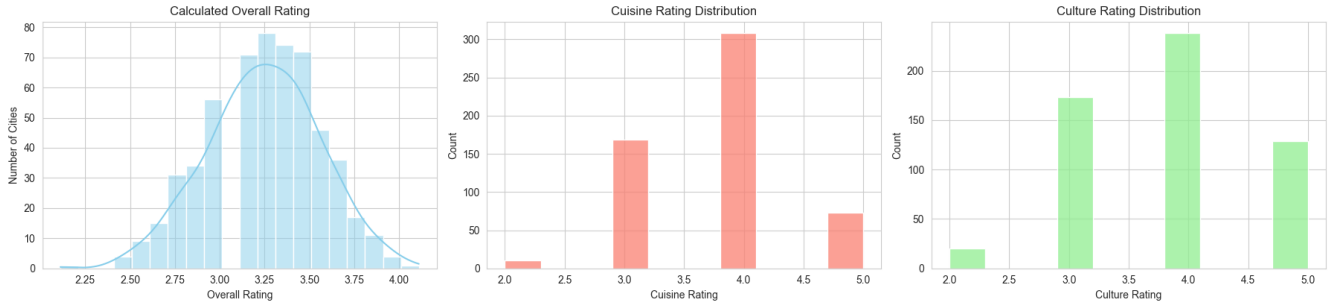

Figure 2. Distribution of Key Ratings

Figure 2 illustrate the frequency distribution of ratings across all cities for three key metrics: overall rating, cuisine, and culture. The 'Calculated Overall Rating' follows an approximately normal distribution, indicating that most cities have an average rating, with fewer cities at the extreme high or low ends. In contrast, the distributions for both 'Cuisine Rating' and 'Culture Rating' are left-skewed, with a large number of cities receiving high ratings (4.0 and above). This suggests that, in general, the cities within this dataset are perceived as having excellent cultural and culinary offerings, which are common and highly-rated attributes for travel destinations.

4.2. Clustering & Visualization

For the clustering analysis, nine thematic rating features were selected: culture, adventure, nature, beaches, nightlife, cuisine, wellness, urban, and seclusion. These features were standardized using the StandardScaler to ensure equal contribution from each feature in the clustering process. AHC was performed using Ward's method, and the resulting hierarchical tree structure was visualized using a dendrogram. The dendrogram was saved and showed the relationship between cities based on their thematic ratings. By "cutting" the dendrogram at a specific height, the cities were grouped into 6 distinct clusters, providing a clear division of cities based on their travel preferences.

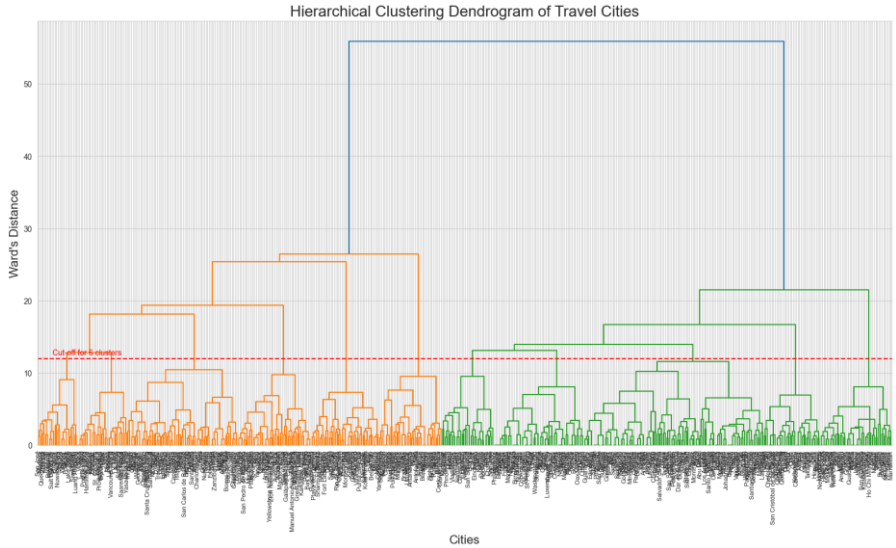

Figure 3. Hierarchical Clustering Dendrogram of Travel Cities

Figure 3 is the primary visualization of the AHC results. It illustrates how individual cities are progressively merged into larger clusters based on the similarity of their thematic travel ratings. The y-axis, labeled "Ward's Distance," represents the dissimilarity or distance between clusters; the greater the vertical distance, the more different the clusters are. The horizontal dashed red line represents the chosen cut-off point, which intersects the vertical lines to define the final number of clusters. In this case, the cut-off is set to create five distinct clusters, grouping cities with similar travel profiles together and separating them from cities with different thematic characteristics. This visualization is fundamental to interpreting the thematic structure of global travel destinations.

4.3. Cluster Analysis

The cluster analysis revealed the following insights into the characteristics of each cluster. Cluster 1 is characterized by high ratings in culture, cuisine, and urban features. Cities in this group are typically vibrant urban centers with rich cultural experiences and diverse cuisines. Sample cities include Milan, New York, and Seoul. This cluster contained 47 cities. Cluster 2 also has a high focus on culture and cuisine, with an emphasis on urban areas. However, it shows a more moderate rating in nightlife compared to Cluster 1. Cities such as Guanajuato, Surabaya, and Kingston are examples of cities in this cluster. This cluster contained 248 cities, making it the largest group. Cluster 3 have high ratings in seclusion and nature, but lower scores in nightlife and cuisine. This cluster seems to represent cities that are more tranquil and nature-oriented, such as Bagan, Malabo, and San Marino. It consists of 41 cities.

Cluster 4 stand out for their emphasis on beaches, nature, and cuisine. Cities like Tampa, Bridgetown, and Rio de Janeiro are examples, highlighting destinations that are known for their natural beauty and culinary experiences. This cluster includes 45 cities. Cluster 5 focuses heavily on nature, adventure, and seclusion, with a much lower emphasis on nightlife and urban areas. Cities such as Nuuk, Livingstone, and Petra are part of this group. These cities tend to be more remote, adventure-filled destinations, and the cluster contains 48 cities. Cluster 6 emphasizes nature, seclusion, and adventure, with moderate scores in culture and cuisine. Cities like Yasawa Islands, Whistler, and Hobart are part of this group, representing destinations that are nature-centric with opportunities for adventure. This cluster contained 131 cities.

The clustering analysis revealed that cities around the world can be grouped into six distinct categories based on thematic travel preferences. Each cluster represents a unique combination of features, with some clusters emphasizing culture and urban life, while others focus on nature, adventure, or seclusion. These findings provide valuable insights into how cities cater to different types of travelers and can inform personalized travel recommendation systems. In conclusion, the results of this clustering analysis offer a deeper understanding of the diverse travel experiences that cities provide, and they highlight the importance of thematic preferences in travel planning.

4.4. Cluster Interpretation

The hierarchical clustering analysis revealed six distinct clusters, each representing cities with unique travel preferences based on their thematic ratings. These clusters provide valuable insights into the diversity of global cities and how they cater to different types of travelers. Cluster 1 is characterized by cities with high scores in culture, cuisine, and urban features. Cities like Milan, New York, and Seoul fall into this group, which can be described as cultural hubs and urban centers. These cities are known for their vibrant city life, rich history, and diverse food scenes, making them ideal destinations for travelers seeking a mix of culture, urban excitement, and culinary experiences. The emphasis on urban and cultural experiences, paired with lower seclusion, makes this cluster perfect for city-centric tourists. Cluster 2 also emphasizes culture, cuisine, and urban features, but with more balanced ratings across other categories like nature and nightlife. Cities such as Guanajuato, Surabaya, and Kingston exemplify this cluster, which can be seen as balanced cultural destinations. These cities offer a blend of urban charm, historical culture, and culinary delights while maintaining accessibility to natural attractions and moderate nightlife. They represent destinations where travelers can experience both the excitement of urban life and a touch of nature. Cluster 3 with high ratings in seclusion and nature, is characterized by cities that offer peaceful, nature-oriented experiences. Examples like Bagan, Malabo, and San Marino highlight remote, tranquil cities that are perfect for those seeking peace and solitude, as well as nature-based activities. These cities appeal to travelers looking for spiritual retreats, quiet escapes, or immersive nature experiences, far from the bustling crowds of urban areas.

Cluster 4 focuses heavily on beaches, nature, and cuisine. Cities like Tampa, Bridgetown, and Rio de Janeiro are examples of beach resorts and natural havens. These cities cater to travelers interested in combining beach relaxation with outdoor activities and culinary adventures. They represent destinations ideal for those seeking natural beauty, outdoor exploration, and local cuisine, making them attractive vacation spots for tourists interested in both leisure and adventure. Cluster 5 stands out with a focus on nature, adventure, and seclusion, while having lower ratings in nightlife and urban features. Cities like Nuuk, Livingstone, and Petra belong to this group, which can be described as adventure-oriented and secluded. These cities are typically located in more remote areas, offering unique opportunities for exploration and adventure in nature. The low emphasis on urban life and nightlife makes this cluster ideal for travelers seeking challenging adventures, nature exploration, and the freedom of secluded, untouched landscapes. Cluster 6 places high importance on nature, seclusion, and adventure, with moderate scores in culture and cuisine. Cities like Yasawa Islands, Whistler, and Hobart represent nature-centric, adventurous destinations. These cities appeal to travelers who are drawn to outdoor activities such as hiking, skiing, and exploring nature, while still offering opportunities for cultural experiences. This cluster attracts those who are passionate about both adventure and nature, seeking the perfect balance between exploration and relaxation in more isolated settings. Each cluster thus reflects a different type of travel experience, catering to varying interests from cultural exploration and urban life to nature immersion and adventure.

4.5. Comparison with Existing Methods

When compared to other clustering techniques, such as K-means clustering, AHC offers a number of advantages that are well-suited to the travel dataset at hand. Unlike K-means, which requires the number of clusters to be predefined, AHC automatically creates a hierarchy of clusters, providing more flexibility in understanding how cities relate to each other at different levels of granularity. The dendrogram visualization generated in this approach allows for a more intuitive understanding of the data, helping to identify the optimal number of clusters visually, which in this case was determined to be 6. In terms of cluster interpretability, AHC also has the advantage of retaining more detailed information about the relationships between cities, making it easier to assess the nuances in the data. On the other hand, K-means can sometimes force the data into a predetermined number of clusters, potentially leading to less meaningful groupings. While K-means may be more efficient for larger datasets, its dependency on the initial selection of centroids and the assumption of spherical clusters can make it less suitable for datasets with more complex structures, like the thematic and subjective nature of travel data.

Moreover, AHC excels when dealing with categorical and continuous mixed data, like the thematic ratings and climate information in this dataset. K-means, while effective for continuous numerical data, can struggle with mixed data types or categorical features unless explicitly adapted, which makes hierarchical clustering a better choice for our analysis. In conclusion, while both AHC and K-means have their strengths, the former provides more interpretability, flexibility, and a clearer understanding of the relationships between cities in the context of thematic travel preferences, making it a more suitable choice for this particular dataset.

5. Conclusion

The clustering analysis revealed six distinct groups of cities, each with unique thematic travel preferences, such as cultural hubs, nature-oriented destinations, and adventure-focused locations. The key findings highlighted that cities tend to cluster around specific features like culture, nature, adventure, and seclusion, offering insights into the diverse travel experiences available worldwide. This segmentation is significant as it demonstrates how cities can be categorized based on their thematic appeal, providing valuable information for understanding traveler preferences. Cities like Milan and New York fall into the cultural and urban-centered cluster, while cities like Bagan and Nuuk are more secluded and nature-oriented, catering to those seeking peace and adventure. These findings have significant implications for personalized travel recommendations, as they allow recommendation systems to align cities with user preferences based on thematic interests, such as culture, adventure, or relaxation. However, the study has limitations, such as relying on a single dataset and using AHC, which may not be the most scalable method for larger datasets. Future research could explore alternative clustering techniques like DBSCAN or K-means, or incorporate additional data sources, such as user reviews or activities, to enhance the robustness of the recommendations. Practically, these results can be implemented into travel recommendation systems by categorizing cities based on their thematic clusters

and offering users tailored travel suggestions that match their preferences for culture, nature, or adventure, thus improving the personalization of travel experiences.

6. Declarations

6.1. Author Contributions

Conceptualization: S.A.G., W.C.S.; Methodology: S.A.G., W.C.S.; Software: S.A.G.; Validation: W.C.S.; Formal Analysis: S.A.G.; Investigation: S.A.G.; Resources: W.C.S.; Data Curation: S.A.G.; Writing – Original Draft Preparation: S.A.G.; Writing – Review and Editing: S.A.G., W.C.S.; Visualization: S.A.G.; All authors have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

6.3. Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

6.4. Institutional Review Board Statement

Not applicable.

6.5. Informed Consent Statement

Not applicable.

6.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] J. Zhang, J. Zhang, and M. Gao, "A Multimodal Travel Route Recommendation System Leveraging Visual Transformers and Self-Attention Mechanisms," *Front. Neurorobot.*, vol. 18, no. November, pp. 1-14, 2024, doi: 10.3389/fnbot.2024.1439195.

[2] J. Lian and D. Liang, "Design and Application of Multiattribute Tourist Information Recommendation Model Based on User Interest," *Wirel. Commun. Mob. Comput.*, vol. 2022, no. 7, pp. 1-16, 2022, doi: 10.1155/2022/1805128.

[3] X. Xiao, C. Li, X. Wang, and A. Zeng, "Personalized Tourism Recommendation Model Based on Temporal Multilayer Sequential Neural Network," *Sci. Rep.*, vol. 15, no. 382, pp. 1-15, 2024, doi: 10.21203/rs.3.rs-5120702/v1.

[4] X. Zhang and Y. Song, "Research on the Realization of Travel Recommendations for Different Users Through Deep Learning Under Global Information Management," *J. Glob. Inf. Manag.*, vol. 30, no. 7, pp. 1-16, 2022, doi: 10.4018/jgim.296145.

[5] J. Chen, L. Huang, C. Wang, and N. Zheng, "Discovering Travel Spatiotemporal Pattern Based on Sequential Events Similarity," *Complexity*, vol. 2020, no. December, pp. 1-10, 2020, doi: 10.1155/2020/6632956.

[6] X. Nan, K. kanato, and X. Wang, "Design and Implementation of a Personalized Tourism Recommendation System Based on the Data Mining and Collaborative Filtering Algorithm," *Comput. Intell. Neurosci.*, vol. 2022, no. August, pp. 1-14, 2022, doi: 10.1155/2022/1424097.

[7] P. Yochum, L. Chang, T. Gu, and M. Zhu, "An Adaptive Genetic Algorithm for Personalized Itinerary Planning," *IEEE Access*, vol. 8, no. April, pp. 88147-88157, 2020, doi: 10.1109/access.2020.2990916.

[8] P. Alves, H. Martins, P. Saraiva, J. Carneiro, P. Nováis, and G. Marreiros, "Group Recommender Systems for Tourism: How Does Personality Predicts Preferences for Attractions, Travel Motivations, Preferences and Concerns?," *User Model. User-Adapt. Interact.*, vol. 33, no. May, pp. 1141–1210, 2022, doi: 10.21203/rs.3.rs-1762820/v1.

[9] H. Wu, H. Jin, Z. Xu, C. Liu, J. Li, and W. Huang, "Travel Mode Classification Based on GNSS Trajectories and Open Geospatial Data," *Trans. GIS*, vol. 26, no. 6, pp. 2598-2620, 2022, doi: 10.1111/tgis.12974.

[10] P. Lahagun, B. Devkota, S. Giri, and P. Budha, "Machine Learning-Based Social Media Review Analysis for Recommending Tourist Spots," *J. Eng. Sci.*, vol. 3, no. 1, pp. 45–52, 2024, doi: 10.3126/jes2.v3i1.66234.

[11] L. W. Dietz, A. Sen, R. Roy, and W. Wörndl, "Mining Trips From Location-Based Social Networks for Clustering Travelers and Destinations," *Inf. Technol. Tourism*, vol. 22, no. January, pp. 131-166, 2020, doi: 10.1007/s40558-020-00170-6.

[12] H. Alshamlan, G. Alghofaili, N. ALFulayj, S. Aldawsari, Y. Alrubaiya, and R. Alabduljabbar, "Promoting Sustainable Travel Experiences: A Weighted Parallel Hybrid Approach for Personalized Tourism Recommendations and Enhanced User Satisfaction," *Sustainability*, vol. 15, no. 19, pp. 14447, 2023, doi: 10.3390/su151914447.

[13] A. J. Sabet, M. Shekari, C. Guan, M. Rossi, F. A. Schreiber, and L. Tanca, "THOR: A Hybrid Recommender System for the Personalized Travel Experience," *Big Data Cogn. Comput.*, vol. 6, no. 4, pp. 131, 2022, doi: 10.3390/bdcc6040131.

[14] S. G. Shirley, K. Subrahmanyam, D. Susrija, and P. Akhila, "K-Means Algorithm and Clustering Technique for a Recommender System," *Int. J. Appl. Sci. Technol. Eng.*, vol. 1, no. 1, pp. 302-312, 2023, doi: 10.24912/ijaste.v1.i1.302-312.

[15] J. Sun, T. Kinoue, and Q. Ma, "Discovery of Points of Interest With Different Granularities for Tour Recommendation Using a City Adaptive Clustering Framework," *Acta Inf. Pragensia*, vol. 2021, no. 3, pp. 275-288, 2021, doi: 10.18267/j.aip.161.

[16] J. Sun, C. Zhuang, and Q. Ma, "User Transition Pattern Analysis for Travel Route Recommendation," *IEICE Trans. Inf. Syst.*, vol. E102, no. 12, pp. 2472- 2484, 2019, doi: 10.1587/transinf.2019edp7096.

[17] T. Øgaard, R. Doran, S. Larsen, and K. Wolff, "Complexity and Simplification in Understanding Travel Preferences Among Tourists," *Front. Psychol.*, vol. 10, no. October, pp. 1-9, 2019, doi: 10.3389/fpsyg.2019.02302.

[18] S. Scaramuccia, S. Nanty, and F. Masséglia, "Feedback Clustering for Online Travel Agencies Searches: A Case Study," 2020, doi: 10.48550/arxiv.2007.07073.

[19] J. Allyn, E. Brottet, E. Antok, L. Dangers, R. Persichini, N. Coolen-Allou, B. Roquebert, N. Allou, D. Vandroux, "Case Report: Severe Imported Influenza Infections Developed During Travel in Reunion Island," *Am. J. Trop. Med. Hyg.*, vol. 97, no. 6, pp. 1943-1944 , 2017, doi: 10.4269/ajtmh.17-0278.

[20] X. Cheng, "A Travel Route Recommendation Algorithm Based on Interest Theme and Distance Matching," *EURASIP J. Adv. Signal Process.*, vol. 2021, no. 57, pp. 1-10, 2021, doi: 10.1186/s13634-021-00759-x.

[21] H. Fang, C. Chen, Y. Long, G. Xu, and Y. Xiao, "DTCRSKG: A Deep Travel Conversational Recommender System Incorporating Knowledge Graph," *Mathematics*, vol. 10, no. 9, pp. 1402, 2022, doi: 10.3390/math10091402.

[22] V. K. Muneer and K. P. Mohamed Basheer, "The Evolution of Travel Recommender Systems: A Comprehensive Review," *Malaya J. Matematik*, vol. 8, no. 4, 2020, doi: 10.26637/mjm0804/0075.

[23] S. Prakki and M. Daneshyari, "Travel Recommendation System Using Graph Neural Networks," *Int. J. Comput. Artif. Intell.*, vol. 4, no. 2, pp. 06-18, 2023, doi: 10.33545/27076571.2023.v4.i2a.66.

[24] Y. Huiling and J. Jiang, "A Personalized Recommendation Technique for Travel Route Based on Fuzzy Consistent Matrix," *Mob. Inf. Syst.*, vol. 10, no. August, pp. 1-10, 2022, doi: 10.1155/2022/3974382.

[25] A. M. Wahid, T. Hariguna, and G. Karyono, "Optimization of Recommender Systems for Image-Based Website Themes Using Transfer Learning," *J. Appl. Data Sci.*, vol. 6, no. 2, pp. 952-968, 2025, doi: 10.47738/jads.v6i2.671.

[26] S. F. Pratama and A. M. Wahid, "Mining Public Sentiment and Trends in Social Media Discussions on Indonesian Presidential Candidates Using Support Vector Machines," *J. Digit. Soc.*, vol. 1, no. 2, pp. 138-151, 2025, doi: 10.63913/jds.v1i2.8.

[27] A. Morales and J. D. Cisneros-Martínez, "Seasonal Concentration Decomposition of Cruise Tourism Demand in Southern Europe," *J. Travel Res.*, vol. 58, no. 8, pp. 1389-1407, 2018, doi: 10.1177/0047287518802094.

[28] M. Qasimi, "Personalized Recommendation Intelligent Fuzzy Clustering Model for theTourism," *J. Comput. Allied Intell.*, vol. 2, no. 5, pp. 42-53, 2024, doi: 10.69996/jcai.2024024.

[29] D. Ernst and S. Dolničar, "How to Avoid Random Market Segmentation Solutions," *J. Travel Res.*, vol. 57, no. 1, pp. 69-82, 2017, doi: 10.1177/0047287516684978.

[30] S. Sorooshian, "Implementation of an Expanded Decision-Making Technique to Comment on Sweden Readiness for Digital Tourism," *Systems*, vol. 9, no. 3, pp. 50, 2021, doi: 10.3390/systems9030050.

[31] A. Alzahrani, "Impact of Dataset Scaling on Hierarchical Clustering: A Comparative Analysis of Distance-Based and Ratio-Based Methods," *Int. J. Anal. Appl.*, vol. 22, no. 36, pp. 1-14, 2024, doi: 10.28924/2291-8639-22-2024-36.

[32] E. U. Oti and M. O. Olusola, "Overview of Agglomerative Hierarchical Clustering Methods," *Brit. J. Comput. Netw. Inf. Technol.*, vol. 7, no. 2, pp. 14-23, 2024, doi: 10.52589/bjcnit-cv9poogw.

[33] K. A. Sethares, C. Y. Jurgens, and M. L. C Vieira, "Physical Heart Failure Symptom Clusters Predictive of Delay in Seeking Treatment," *Nurs. Res.*, vol. 73, no. 6, pp. 426-433, 2024, doi: 10.1097/nnr.0000000000000755.

[34] L. Sokhonn, Y.-S. Park, and M. Lee, "Hierarchical Clustering via Single and Complete Linkage Using Fully Homomorphic Encryption," *Sensors*, vol. 24, no. 15, pp. 4826, 2024, doi: 10.3390/s24154826.

[35] U. O. Eric and O. Michael, "Comparative Evaluation of Six Agglomerative Hierarchical Clustering Methods With a Robust Example," *Afr. J. Math. Stat. Stud.*, vol. 7, no. 2, pp. 1-25, 2024, doi: 10.52589/ajmss-qxph8r1n.

[36] C. Zhao, M. Johnsson, and M. He, "Data Mining With Clustering Algorithms to Reduce Packaging Costs: A Case Study," *Packag. Technol. Sci.*, vol. 30, no. 3, pp. 173–193, 2017, doi: 10.1002/pts.2286.

[37] L. Dhulipala, D. Eisenstat, J. Łącki, V. Mirrokni, and J. Shi, "Hierarchical Agglomerative Graph Clustering in Nearly-Linear Time," 2021, doi: 10.48550/arxiv.2106.05610.

[38] A. E. Olawumi and F. Dahunsi, "Load Profiling of Commercial and Residential Building Using Clustering Technique," *Niger. J. Technol.*, vol. 42, no. 2, pp. 257-263, 2023, doi: 10.4314/njt.v42i2.14.

[39] A. Muhsina and B. Joseph, "Evaluation of Different Clustering Techniques in Classifying the Vegetable Growing Panchayats of Ernakulam District, Kerala," *Int. J. Plant Soil Sci.*, vol. 34, no. 24, pp. 500-510, 2022, doi: 10.9734/ijpss/2022/v34i242666.

[40] S. Çınaroğlu, "Clustering of OECD Countries Out of Pocket Health Expenditure Time Series Data," *Res. Appl. Econ.*, vol. 8, no. 2, pp. 23-38, 2016, doi: 10.5296/rae.v8i2.9377.

[41] T. Strauss and M. von Maltitz, "Generalising Ward's Method for Use With Manhattan Distances," *Plos One*, vol. 12, no. 1, pp. e0168288, 2017, doi: 10.1371/journal.pone.0168288.

[42] G. Dağtekin, A. Kılınç, E. Çolak, A. Ünsal, and D. Arslantaş, "Classification of Oecd Countries in Terms of Medical Resources and Usage With Hierarchical Clustering Analysis," *Osmangazi J. Med.*, vol. 44, no. 4, pp. 487 – 492, 2021, doi: 10.20515/otd.978341.

[43] M. Sammour and Z. A. Othman, "An Agglomerative Hierarchical Clustering With Various Distance Measurements for Ground Level Ozone Clustering in Putrajaya, Malaysia," *Int. J. Adv. Sci. Eng. Inf. Technol.*, vol. 6, no. 6, pp. 1127–1133, 2016, doi: 10.18517/ijaseit.6.6.1482.

[44] J. Wang, Y. Xia, and Y. Wu, "Sensing Tourist Distributions and Their Sentiment Variations Using Social Media: Evidence From 5A Scenic Areas in China," ISPRS *Int. J. Geo-Inf.*, vol. 11, no. 9, pp. 492, 2022, doi: 10.3390/ijgi11090492.

[45] J. Gan, D. Zhang, F. Guo, and E. Dong, "Intensity of Tourism Economic Linkages in Chinese Land Border Cities and Network Characterization," *Sustainability*, vol. 16, no. 5, pp. 1843, 2024, doi: 10.3390/su16051843.

[46] X. Zhang, C. Song, C. Wang, Y. Yang, Z. Ren, M. Xie, Z. Tang, and H. Tang, "Socioeconomic and Environmental Impacts on Regional Tourism Across Chinese Cities: A Spatiotemporal Heterogeneous Perspective," ISPRS *Int. J. Geo-Inf.*, vol. 10, no. 6, pp. 410, 2021, doi: 10.3390/ijgi10060410.