A Comparative Analysis of DBSCAN, K-Means and Agglomerative Clustering Algorithms for Geospatial Data

Anupam Jain[1*], Khushal Rathi[2], Yuboraj Ganguly[3], Ankit Kumar[4] and Yogiraj Bhale[5]

[1,2,3,4,5] AIT CSE, Chandigarh University, Punjab 140301, India.
*anupamayushij@gmail.com, khushalrathi0@gmail.com,
yuvrajganguly09@gmail.com, ankitdkkumar@gmail.com,
yogirajb85@gmail.com

Abstract. This study presents a comparative analysis of three popular clustering algorithms, DBSCAN and KMeans, Agglomerative Clustering applied to geospatial data. We focus on their performance based on the silhouette score, examining their ability to identify meaningful clusters in noisy data. Our results show that DBSCAN outperforms KMeans and Agglomerative 9oClustering, achieving a silhouette score of 0.8646 compared to KMeans' 0.8160 and Agglomerative Clustering's 0.8160, highlighting DBSCAN's robustness in identifying clusters with irregular shapes and handling noise.

Keywords: Density-Based Spatial Clustering of Applications with Noise (DBSCAN), Non-negative Matrix Factorization (NMF), Singular Value Decomposit, Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Systematic Literature Review (SLR), k-Nearest Neighbors (k-NN), Deep Neural Network (DNN), Multilayer Perceptron (MLP),

1. Introduction

In data analysis, clustering algorithms are essential because they make it possible to find organic groupings within individual data points. Because they are easy to use and efficient, K-Means, DBSCAN, and Agglomerative Clustering are three of the most used clustering techniques. K-Means clustering separates data into K distinct clusters based on the distance between data points and cluster centroids.

Although it works effectively in many situations, its sensitivity to outliers and necessity that the number of clusters be predetermined make it less feasible for usage in real-world situations. K-Means clustering separates data into K distinct clusters based on the distance between data points and cluster centroids. Although it works well in many situations, its sensitivity to outliers and need for the number of clusters to be predetermined make it less feasible for usage in real-world situations.

Agglomerative Clustering, a hierarchical clustering method, builds clusters through a bottom-up approach, merging pairs of clusters based on a defined distance metric. While it provides a flexible

S. Bhalerao et al. (eds.), *Proceedings of the International Conference on Recent Advancement and Modernization in Sustainable Intelligent Technologies & Applications (RAMSITA-2025),* Advances in Intelligent Systems Research 192,
https://doi.org/10.2991/978-94-6463-716-8_18

framework for determining the number of clusters, it can be computationally expensive and may struggle with large datasets, especially when dealing with high-dimensional data.

In contrast, DBSCAN (spatial clustering based on density and noise) identifies clusters based on the density of data points in a region. It is very good at finding clusters of normality, and handles noise and outliers well, making it ideal for spatial data analysis. DBSCAN's density feature allows it to detect clusters of different shapes and sizes, which is often limited by K-Means and cluster analysis.

This study aims to evaluate the performance of these clustering algorithms using a geospatial dataset, focusing on their respective silhouette scores to assess cluster cohesion and separation. The silhouette score may be a broadly acknowledged metric for evaluating the quality of clustering comes about, because it measures how comparative an question is to its possess cluster compared to other clusters.

The results of our analysis demonstrate that DBSCAN achieves a silhouette score of 0.8646, outperforming both K-Means, which scored 0.8160, and Agglomerative Clustering, which also scored 0.8160. This indicates that DBSCAN provides better-defined clusters compared to its counterparts, making it a superior choice for clustering applications in geospatial data analysis. The discoveries of this investigate emphasize the significance of selecting the fitting clustering calculation based on the characteristics of the dataset, especially within the nearness of commotion and exceptions.

The results of this study offer practical insights for urban planners, policymakers, businesses, and individuals. Urban planners can use the analysis for developing optimized recommendations in infrastructure, while businesses can leverage the findings for market expansion. Smart recommendation system can be developed for induvial for making informed decisions about housing transportation and other amenities. The analysis also supports sustainable urban development strategies through efficient clustering of amenities and transportation options.

The structure of this paper is as takes after: Segment 2 depicts the strategies utilized within the consider, the datasets utilized, and the calculations utilized. Segment 3 gives an outline of the exploratory comes about, counting a comparison of the classification execution of each calculation. In Segment 4, we examine the suggestions of these discoveries and their suggestions for spatial information investigation. At long last, segment 5 concludes the paper with a rundown of key discoveries and recommendations for future investigate.

In conclusion, our study demonstrates the benefits of DBSCAN in spatial data collection and provides insights that may be useful to researchers and practitioners looking for new clustering methods.

2. Literature Review

In 2024, Fang et al. developed a multi-database framework for spatial analysis in the Pearl River Delta metropolitan area using mobile phone data from 14 million users. This study involved creating a human mobility network, examining spatial integration patterns, and examining spatial connectivity through a non-disruptive mobility network with nodes representing 500 m grid cells and mobility more than 19 million edges. Using Louvain's algorithm, this study found similar communities based on the intensity of interaction and showed how regions can create cohesive communities through activities such as work and health care. . Regardless of the exact reasons for the change, the framework identified spatial patterns and obtained a Pearson correlation coefficient of 0.89 with valid statistical data. Despite privacy concerns and sample bias, this study provides important insights for urban planning and decision-making in urban areas.[1]

In 2023, R. Alabduljabbar et al. evaluated various alternatives to improve restaurant recommendations using Foursquare.com data. They implemented and evaluated four algorithms: NMF, SVD, and SVD++, as follows to determine their effectiveness in producing appropriate recommendations based on the user. Comments and ratings This research involves pre-processing user comments and evaluating algorithms using measures such as RMSE and MAE. In normal analysis methods, SVD outperformed RMSE, while NMF performed better than MAE. SVD++ is less accurate. Content recommendations do not have exact evaluation criteria and instead rely on user feedback. The findings showed that the personalization matrix methods were good at personalizing the recommendations, but the authors suggested further optimization and evaluation strategies for the content-based models for future research.[2]

In 2023, Caiwen Li et al. improvised deep learning-based recommendation systems using SLR and classification techniques. They analyzed 799 articles and classified 140 techniques into 27 groups, featuring popular techniques such as matrix factorization, graph neural networks, and attention techniques. This study highlights the importance of collaborative neural analysis, especially in social evidence and observations. Although the detailed classification is a useful resource for future researchers, it also shows a large variation in the popularity of the methods in different regions, indicating a limited use in the world. Matrix ratio is the most commonly used method. Overall, this study provides important insights for improving recommender systems, but emphasizes the need for better generalizability across contexts.[3]

In 2022, Ahmed et al. provided thorough review of recommender system techniques was given with an emphasis on matrix generation and collaborative analysis techniques. They evaluated the effectiveness of methods such as singular value decomposition (SVD), non-negative matrix factorization (NMF), and SVD++ for handling sparse data matrices in film evidence, e.g. demonstrating their ability to accurately predict rates even when the data is imprecise. However, these methods are subject to vulnerability, which regulation can reduce. The authors also reviewed user-based and k-NN algorithms, and pointed out their challenges in scalability and computational complexity. Their findings show that matrix factorization is better than k-NN in terms of accuracy and efficiency for large datasets, but k-NN is good for small datasets due to its simplicity. This study suggests that future research should focus on hybrid approaches that combine the strengths of both methods.[4]

In 2022, Wu et al. conducted a comprehensive review of datasets and techniques in graph neural network (GNN)-based recommender systems and demonstrated the importance of different datasets for practical applications. They discussed key datasets, including the MovieLens dataset for user interaction analysis, Amazon's dataset for interactive recommendations and order, Yelp's dataset for keywords, and Gowalla's dataset for serial evidence. Additionally, they leveraged the Yoochoose and Diginetica datasets for e-commerce and event-related recommendations. This study met the common evaluation criteria for GNN-based recommendations, such as Precision@K, Recall@K and AUC. Although GNNs increase the accuracy of evidence by capturing complex relationships, they face challenges such as scaling with large graphs and the need for efficient sampling strategies. Wu et al recommended further research on dynamic graphs, improving the user's interest profile, and examining the studies you are investigating to increase robustness to noisy data and increasing the efficiency of the computer.[5]

In 2022, Rohilla et al. It focuses on the challenges of estimating quality of service in distributed systems, especially the limitations of traditional link analysis methods. They introduced a new model called group-based recommendation through deep learning , which combines DNN and MLP architectures to capture in non-inflation and high-inflation. . This approach directly handles small data, which is a common problem in real-world situations, while introducing social-services to improve forecast quality. However, the authors agreed that privacy concerns are limitations that deserve further investigation. The RLSD model was evaluated on the WS-Dream dataset, including 1974675 QoS values from 339 users and 5825 services. Based on the evaluation criteria, RLSD outperformed common analysis methods such as the UPCC, LACF, and RegionalKNN, achieved better performance in MAE and RMSE. Although the RLSD model showed improvements in predicting response time and throughput, future research may focus on parallel and distributed systems to increase scalability and efficiency.[6]

In 2022, Safavi et al. conducted a comprehensive review of POI tagging systems and compared traditional machine learning methods with deep learning models. They emphasized that while traditional approaches, such as conjoint analysis, are useful in many cases, they struggle with challenges such as parsimony and scalability. To address these issues, models such as PR-RCUC and STACP have been developed to improve evidence by incorporating time, region and user context. However, they still find it difficult to provide logical explanations for their results. Deep learning models, including RNN, LSTM, and CNN, on the other hand, outperform traditional methods by efficiently handling large data sets and capturing non-linear relationships. This study emphasizes the importance of combining geographic, physical and social factors to improve the accuracy of evidence. Using datasets such as Foursquare and Gowalla, this study used metrics such as precision and recall to evaluate model performance. Future research directions suggested by the authors include including more contextual factors, exploring in-depth reinforcement learning, and developing solutions to data inaccuracy and the problem of startups. cold.[7]

In 2022, N. Aggarwal et al. conducted a comprehensive review of Wireless Sensor Networks , highlighting their exponential growth and extensive applications across home automation, healthcare, environmental monitoring, and military sectors. The review addresses the inherent challenges in implementing WSNs due to the resource constraints of sensor nodes, emphasizing the need for energy-efficient solutions to enhance the lifespan of these networks. It notes significant research efforts aimed at minimizing energy consumption during various tasks, with the LEACH protocol identified as a foundational approach for developing advanced routing protocols focused on power efficiency. The paper compares the performance of various energy-efficient routing protocols based on key WSN characteristics and highlights crucial technological differences among the latest routing methods. This review serves as a foundational reference for future research aimed at designing improved routing protocols for WSN applications, incorporating more effective energy management techniques.[8]

In 2021, Akhtar et al. He highlighted the role of artificial intelligence models in terms of trends and data quality, and reviewed methods for predicting traffic congestion. They compared inference methods such as fuzzy logic and Bayesian networks with machine learning approaches and concluded that deep learning is very useful for handling complex datasets and extracting features and don't need to know. Although shallow ML models, including artificial neural networks and decision trees, are useful for short-term forecasts due to their robustness, they struggle to capture complex patterns. In contrast, deep learning models, such as neural and recurrent networks, are good at handling non-linear tasks, but require significant computing resources. The study highlights areas for improvement, including long-term data collection to better capture dynamic traffic conditions and include external factors such as social media and weather conditions. Challenges such as missing data weaken the accuracy of the model. Future research directions include real-time prediction, case-based learning to increase prediction accuracy, and integration of multiple vehicle parameters to improve reliability. Despite the great progress, this research needs further research into new algorithms and modification of traffic jam prediction models.[9]

In 2021, Abbasi Mod et al. proposed a context-aware tourism recommendation system designed to improve travel recommendations through a three-step process. The system includes extracting user preferences from reviews, identifying interesting features and providing recommendations based on their similarity to user preferences, and incorporating contextual features such as weather and location. User preferences are analyzed using segment-of-speech tagging and sentiment analysis, while attractive features are derived from above-mentioned information under weather conditions. This system, evaluated using TripAdvisor reviews, outperformed previous methods and showed high accuracy in accuracy, recall and F measure. However, this study acknowledges limitations, including exclusion of traffic conditions, and recommends future research to address these issues for better evidence.[10]

3. Methodology

3.1 Data Collection

The dataset utilized in this consider was sourced from Google Places APIs, which give wealthy geospatial information for different areas over distinctive districts. This dataset envelops different qualities significant to location-based clustering, counting names, pincode, conveyance status, division title, locale title, circle title, taluk, area title, state title, and geographic facilitates (longitude and scope). This differing set of highlights permits for an in-depth investigation of clustering behaviors, especially in how geographic nearness influences clustering results.

The geospatial data primarily consisted of POI details, geographic coordinates. POI data included amenities relevant to urban planning. The data enabled effective clustering using DBSCAN and supported user-specific recommendations based on proximity and accessibility.

To guarantee the strength of our investigation, we chosen a critical volume of information points—totaling roughly 1,000 one of a kind geological area. The focuses were dispersed among different clusters, speaking to different office sorts and statuses, which encourage enhances the clustering handle. The accessibility of this organized dataset empowered the usage of different clustering calculations to survey their execution precisely.

3.2 Clustering Algorithms

In order to assess the effectiveness of clustering techniques on the geospatial dataset, we implemented three distinct clustering algorithms, each selected for its unique characteristics and strengths:

- **DBSCAN** : DBSCAN was chosen for its ability to identify clusters of varying densities while effectively filtering out noise. This characteristic is particularly advantageous in geospatial clustering, where data points may not always conform to a spherical distribution. In our think about, we balanced the eps to 0.5, which decided the most extreme separate between two tests for them to be considered as within the same neighborhood. This setting permitted DBSCAN to distinguish clusters based on the thickness of information focuses and isolated exceptions viably.

- **KMeans Clustering**: KMeans may be a broadly utilized centroid-based clustering calculation that segments the dataset into a indicated number of clusters. This calculation works by initializing k haphazardly and doling out each information point to the closest center, redetermining the centers based on the chosen focuses. Although KMeans works well and is easy to implement, it focuses on the original location of the centroids and assumes that the clusters are spherical. In this study, we selected an optimal k-value through the knee method and ensured that large enough clusters were formed to represent the underlying data distribution.
- **Agglomerative Clusters:** This sorting algorithm creates a list of clusters by recombining adjacent clusters. There is no need to determine the number of clusters in advance, which allows for data analysis. Variables with different correlation criteria (simple, complete, simple) were used to evaluate the impact on the synthesis process. By creating a dendrogram, we evaluate the clustering process and identify the most accurate clusters based on geographic proximity. Each of these algorithms was given a set of data to compare their performance under similar conditions.

3.3 Evaluation Metrics

To evaluate the cumulative performance of each algorithm, the silhouette indicator was used as the main evaluation criteria. The silhouette symbol shows how similar the data point of the cluster is compared to other clusters, and provides a cluster's density and separation scale. Scores range from -1 to 1, with higher silhouette scores indicating better defined clusters, with scores closer to 1 indicating good clustering of symptoms and scores closer to 0 indicating overlapping clusters. This criterion is important for understanding the performance of various clustering methods, especially for distinguishing noisy and meaningful clusters.

In addition to the silhouette score, the number of noise points identified by DBSCAN was recorded to evaluate its performance in handling outliers. The capability of DBSCAN to identify and exclude noise is particularly valuable in geospatial data, where the presence of noise can significantly affect clustering outcomes. The methodology employed in this research provides a structured approach to clustering evaluation, allowing for a comprehensive comparison of the performance of DBSCAN, K-Means, and Agglomerative Clustering on geospatial data.

4. Result Analysis

In this study, we implemented three algorithms: DBSCAN, K-Means and Agglomerative Clustering, based on a dataset containing spatial data. The performance of each algorithm was evaluated using the silhouette score, a metric that measures the similarity of objects in a cluster to other clusters. Silhouette scores range from -1 to 1, with higher scores indicating more distinct clusters. The results of the silhouette scores for each algorithm cluster are as follows:

•DBSCAN: **Fig. 1** shows that the silhouette score achieved during the analysis is 0.8646.

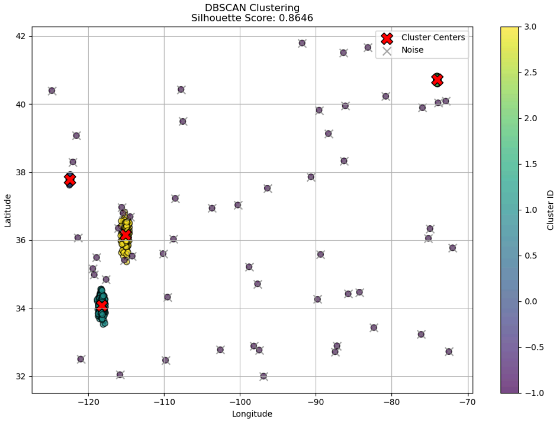
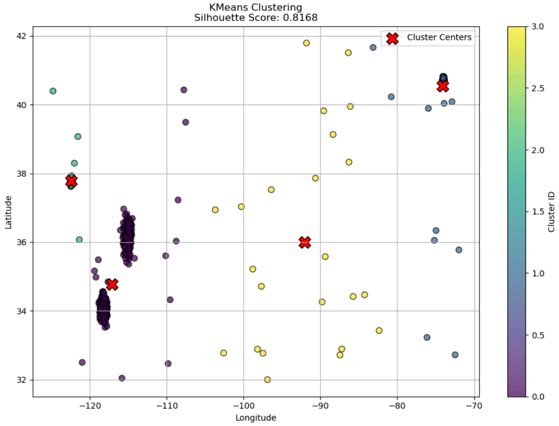


Fig. 1. DBSCN Clustering

- K-MEAN: **Fig. 2** shows achieved silhouette index is 0.8160.



.

Fig. 2. K-Means Clustering

- Agglomerative: **Fig. 3** depicted that the achieved silhouette score is 0.8160.

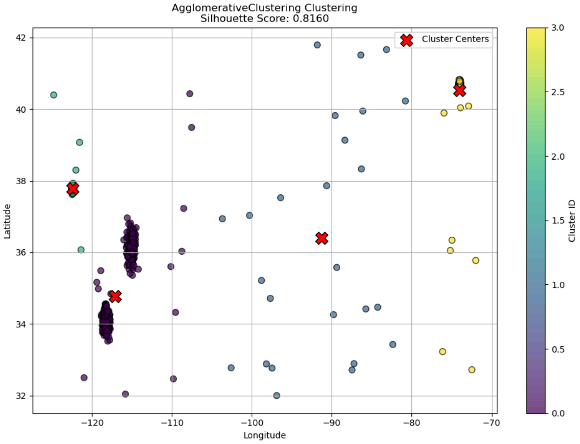


Fig. 3. Agglomerative Clustering

From the results, it can be seen that DBSCAN is more powerful than K-Means and Agglomerative Clustering in terms of ghost results. A higher silhouette index in DBSCAN indicates that the clusters created are better and better separated than the clusters created by the other two algorithms. This indicates that DBSCAN is very suitable for this data set, due to its ability to identify clusters of different shapes and densities.

One of the significant advantages of the DBSCAN algorithm is its capability to identify noise points within the dataset. Noise points are data points that do not belong to any cluster, often considered outliers. This characteristic is especially crucial in real-world datasets where noise can significantly impact the results of clustering.

In our dataset, DBSCAN was able to successfully recognize between center focuses, border focuses, and commotion focuses as shown in **Fig. 4**.

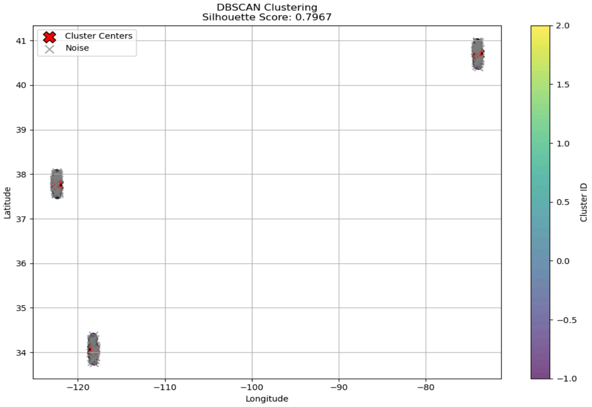


Fig. 4. Noise Point Identification with DBSCAN(Number of noise points=93)

The recognizable proof of commotion focuses not as it were makes strides the quality of the clustering comes about but too gives experiences into the characteristics of the dataset. Understanding the nature and dispersion of these clamor focuses can be important for ensuing information investigation or preprocessing steps, assist educating the choice of clustering calculations for future ponders.

5. Conclusion

This study evaluated the performance of three clustering algorithms—DBSCAN, K-Means, and clustering—on a dataset obtained from the Google Maps API. The results showed that DBSCAN achieved a high silhouette index of 0.8646, indicating its good ability to identify meaningful clusters in complex datasets. This capability is important for applications in internal services, where accurate aggregation is critical for decision making.

Additionally, DBSCAN's ability to detect noisy signals sets it apart from other algorithms and is more capable of handling latency. The findings show that using DBSCAN can improve urban data analysis and provide valuable insights for stakeholders in areas such as urban planning and logistics. Future research could investigate the use of DBSCAN in larger datasets and consider hybrid approaches to improve clustering results.

References

1. Fang, B., Li, M., Huang, Z., Yue, Y., Tu, W., & Guo, R. (2024). Revealing multi-scale spatial synergy of mega-city region from a human mobility perspective. *Geo-spatial Information Science*, 1-16.
2. Alabduljabbar, R. (2023). Matrix Factorization Collaborative-Based Recommender System for Riyadh Restaurants: Leveraging Machine Learning to Enhance Consumer Choice. Applied Sciences, 13(17), 9574.
3. Li, C., Ishak, I., Ibrahim, H., Zolkepli, M., Sidi, F. and Li, C., 2023. Deep Learning-Based Recommendation System: Systematic Review and Classification. IEEE Access.
4. Ahmed, M., Ansari, M. D., Singh, N., Gunjan, V. K., BV, S. K., & Khan, M. (2022). Rating‐Based Recommender System Based on Textual Reviews Using IoT Smart Devices. Mobile Information Systems, 2022(1), 2854741.
5. Wu, S., Sun, F., Zhang, W., Xie, X., & Cui, B. (2022). Graph neural networks in recommender systems: a survey. ACM Computing Surveys, 55(5), 1-37.
6. Rohilla, V., Kaur, M., & Chakraborty, S. (2022). An Empirical Framework for Recommendation-based Location Services Using Deep Learning. Engineering, Technology & Applied Science Research, 12(5), 9186-9191.
7. Safavi, S., Jalali, M., & Houshmand, M. (2022). Toward point-of-interest recommendation systems: A critical review on deep-learning Approaches. Electronics, 11(13), 1998.
8. Aggarwal, N., Sharma, N., & Bhale, Y. (2022, April). Performance Analysis of Power Efficient Routing Protocols for wireless sensor networks: a survey. In 2nd IEEE International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE) (pp. 270-274)..
9. Akhtar, M., & Moridpour, S. (2021). A review of traffic congestion prediction using artificial intelligence. *Journal of Advanced Transportation*, *2021*(1), 8878011.
10. Abbasi-Moud, Z., Vahdat-Nejad, H., & Sadri, J. (2021). Tourism recommendation system based on semantic clustering and sentiment analysis. Expert Systems with Applications, 167, 114324.