© APR 2026 | IRE Journals | Volume 9 Issue 10 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I10-1716179

Development Of an Intelligent Location Data Clustering System Using Regular Expressions and Dbscan

ONYEDIKACHUKWU O. IKECHUKWU-ONYENWE[1], DORIS ASOGWA[2], IKECHUKWU EKENE ONYENWE[3]

[1,2,3]*Department of Computer Science Nnamdi Azikiwe University, Awka*

Abstract- ***Intelligent location data clustering systems are used to group spatial information such as GPS coordinates, addresses, and points of interest into meaningful clusters based on geographic proximity and similarity. In many developing countries, including Nigeria, the absence of a standardized address structure creates significant challenges for postal delivery, logistics management, emergency response, and urban planning. Address data often contain spelling variations, abbreviations, incomplete structures, and inconsistent formatting, making accurate geocoding and spatial analysis difficult. This study presents the development of an intelligent location data clustering system that integrates Regular Expressions (Regex) for address normalization and the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm for spatial clustering. The system preprocesses unstructured address data, standardizes them according to the Nigerian Postal Service (NIPOST) addressing framework, and groups similar addresses into spatial clusters. The approach eliminates reliance on external geocoding services and provides a locally controlled solution suitable for limited-resource environments. Experimental results demonstrate that the proposed system improves clustering accuracy and provides meaningful geographic groupings of address data. The system can be applied in postal delivery optimization, urban planning, logistics operations, and public service management.***

Index Terms- ***Location Data Clustering, Natural Language Processing (NLP), Regular Expressions, DBSCAN, Machine Learning, Address Standardization, NIPOST, Geospatial Analysis.***

I. INTRODUCTION

Location-based information systems play a critical role in modern digital infrastructure. Accurate location data enable efficient postal delivery, emergency response, logistics planning, and urban development. However, in many developing countries such as Nigeria, address systems remain inconsistent and poorly standardized. This results in fragmented location information that is difficult to process automatically.

Many Nigerian addresses contain typographical errors, inconsistent abbreviations, missing postal codes, and informal landmark references. These inconsistencies create challenges for conventional geocoding systems, which rely on well-structured address formats. Consequently, postal services, logistics companies, and government agencies face difficulties in identifying and clustering geographic locations accurately.

Artificial Intelligence (AI) and text-processing techniques provide opportunities to address these challenges. Regular Expressions (Regex) offer a rule-based mechanism for detecting and standardizing address patterns within textual data. By extracting key address components such as street names, local government areas, and postal codes, Regex helps transform unstructured text into machine-readable data.

Once address data are standardized, machine learning clustering techniques can be used to group similar locations. One of the most effective algorithms for spatial clustering is the Density-Based Spatial Clustering of Applications with Noise (DBSCAN). DBSCAN identifies clusters based on data density and can effectively detect irregular cluster shapes while identifying outliers or noise within datasets.

The primary aim of this research is to develop an intelligent location data clustering system capable of processing unstructured address data and grouping them into spatial clusters.

II. LITERATURE REVIEW

A. Sharma (2023) highlighted the crucial importance of location data across a variety of industries, including e-commerce, banking, insurance, and supply chain management. Maintaining accurate and reliable address information is essential for data verification and guaranteeing correct and timely delivery of goods, both now and going forward. This investigation used a new approach to address standardization that utilizes Named Entity Recognition (NER). NER, a technology that categorizes text into predefined labels like person, address, organization, region, and geographic coordinates, is shown to be valuable for identifying different characteristics of an address. The developed system has broad potential uses, spanning from preventing fraudulent activities to automating translation processes and aiding in data discovery. Natural Language Processing (NLP) is acknowledged as a vital aspect of NER, with its capabilities extending to text recognition, computer-assisted translation, and the identification of text's meaning.

The proposed system accurately *Hassini et al. (2023),* proposed a hybrid model combining REGEX and machine learning to improve spatial entity recognition in text. This is dependent on language-specific REGEX; and lacked the adaptability to informal address formats whereas the proposed system specializes in a method of Hybridizing Regular Expressions (REGEX) as a pre-filtering tool with machine learning models to enhance the precision and recall of entity recognition in semi-structured text information, which is fundamental to the effective parsing of addresses.

Mehta et al. (2023) presented a parallel geocoding system that employed four geocoding services to handle imprecise road traffic incident (RTI) address information in Lagos, Nigeria. The methodology effectively generated significant spatial trends related to traffic injuries, identifying high-risk areas and differences in ambulance response durations. It showed that even data with poor address quality can yield valuable insights when processed using several geocoding engines. The study encountered ongoing issues with unclear addresses, positional inaccuracies greater than 1 kilometer, and a lack of address standardization prior to geocoding. The model included techniques like text cleaning using regular expressions or machine-learning-based grouping to improve precision. The application could be used with postal, population, or service provision datasets.
Kebe et al (2019) Introduced a multi-agent geocoding system designed for regions lacking standardized address systems. Developed a representative local address structure and employed collaborating agents to process and match address data. The system enhanced geocoding accuracy and delivery efficiency in developing countries with limited mapping resources and the new system encourages the use of automated set of rules and has ML integration and semantic context detection embedded in it.

The researchers *Gupta, Gupta, Garg, and Garg (2021)* delved into a semantic address matching technique, a sophisticated natural language processing (NLP) application powered by deep learning. The objective is to accurately identify a particular address from a pool of possible matches. After scrutinizing prevalent methodologies and acknowledging their shortcomings, they introduce a semantic-based strategy designed to resolve the inadequacies of established systems. These often struggle with challenges like replicated or abbreviated address data. Their proposed solution utilizes Optical Character Recognition (OCR) to glean address information from invoices, thereby constructing a comprehensive dataset. This data then undergoes processing using the BM25 ranking function to prioritize the most pertinent entries. To further refine the outcome and pinpoint the best possible match, the top candidates were subsequently evaluated using the BERT model. The findings of this research investigation reveal that this innovative approach substantially improves both the ac` curacy and completeness of address matching compared to current leading-edge methodologies.

Lin et al (2019) under the title Learning Architecture for Semantic Address Matching introduced a deep learning model achieving high accuracy in semantic address matching tasks, emphasizing contextual similarity between address entries. The model was limited in interpretability but the new system provided contextual linkage with geographic or administrative hierarchies.

III. METHODOLOGY

© APR 2026 | IRE Journals | Volume 9 Issue 10 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I10-1716179

The Methodology adopted for this work is Hybrid Methodology. This system follows a Design Science Research (DSR) approach. The core methodology involves analyzing an existing geocoding constraint identified in a Limited Resource Setting (LRS), proposing an indigenous, data-centric solution, and developing an autonomous system to implement this solution. The prediction system software was configured using Python was selected due to its powerful libraries, ease of use, and versatility in handling complex tasks and Folium (Python library) used for creating interactive web designs and Plotly- a graphing library that can be used to create interactive and quality geographic Information System (GIS) maps.

This methodology employs a three-phase approach: Data Pre-processing, Pattern Recognition and Standardization, and Autonomous Clustering.

i. Data Pre-processing (The Regex Engine): This phase addressed the unstructured nature of the raw location address data, which involves the use of Regular Expressions (Regex).
ii. Pattern Recognition and Standardization (The NIPOST Layer): This stage imported native, hierarchical spatial data into the information, adopting a NIPOST Postal Code Mapping and Hierarchy.
iii. Autonomous Location Clustering (The AI Layer) This final phase generates the precise spatial units for analysis. Techniques of Density-Based Spatial Clustering of Applications with Noise (DBSCAN) or a similar unsupervised clustering algorithm, applied to the standardized location features.

Experimental Tools

Python

Python is a high-level programming language widely used in artificial intelligence, data analysis, and machine learning applications. Its simplicity and extensive library support make it suitable for implementing geospatial data analysis and clustering algorithms.

Regular Expressions (Regex)

Regular Expressions are pattern-matching techniques used for text processing and data extraction. In this system, Regex was used to parse unstructured address strings and extract meaningful components such as street names, postal codes, and administrative regions.

DBSCAN Clustering Algorithm

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is an unsupervised machine learning algorithm used to identify clusters within spatial datasets. DBSCAN groups together points that are closely packed while marking points in low-density regions as noise.

NIPOST Address Database

The Nigerian Postal Service (NIPOST) address structure provides a hierarchical addressing system that includes postal codes, delivery zones, and administrative regions. The system uses this structure to validate and assign standardized postal codes to address data.

IV. EXISTING SYSTEM

The existing system address administration framework in Nigeria, largely overseen by the Nigerian Postal Service (NIPOST), depends significantly on manual procedures and fixed postcode allocations. While endeavors like the parallel geocoding approach presented by Mehta et al. (2023) offer advancements for spatial pattern analysis, these solutions are often limited to particular applications and lack independent capacity in processing unstructured address information. The existing system introduced a concurrent geocoding procedure designed to enhance the analysis of imprecise datasets frequently found in low- and middle-income countries (LMICs). This procedure is then implemented and assessed using a road traffic injury (RTI) dataset from Lagos State, Nigeria, reducing location inaccuracies during geocoding by integrating results from four different commercial geocoding platforms. The agreement among the outputs from these platforms is examined, and spatial representations are created to offer an understanding of the spread of RTI events within the study area. This work underscores the significance of location-based data analysis in LMICs, enabled by contemporary tools, for health resource distribution and, ultimately, patient well-being.

© APR 2026 | IRE Journals | Volume 9 Issue 10 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I10-1716179

V. PROPOSED SYSTEM

The primary objective of this system was to pre-process and standardize low-fidelity address data using indigenous knowledge (NIPOST structure) and algorithmic methods (Regex) to produce high-fidelity, structured location clusters before any costly commercial geocoding is performed. This would significantly increase the accuracy and consistency of spatial analysis while reducing operational costs.

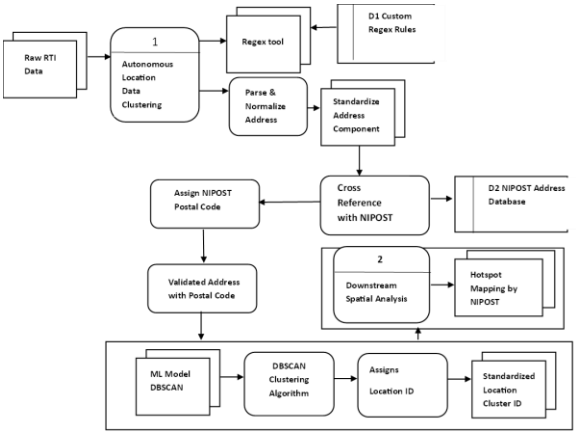

Fig 1. A Data Flow Diagram of the New System

The main menu is the homepage that contains all components that made up the system. It consists of the location module, the clusters module and the pending module. Detailed description of the main menu can be seen as;

a. Locations: This is a module that contains the RTI data location of address that has been ingested into the repository of the system.

b. Clusters: This module contain the clustered address locations with their unique cluster identifier. It depicts the total number of unique spatial groups identified by the AI.

c. Pending: These are records that have been added but are awaiting final clustering or validation.

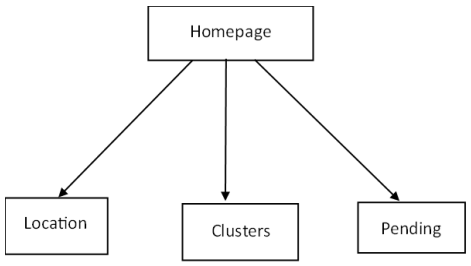

Fig 1b. The workflow menu diagram of the system

VI. RESULT AND ANALYSIS

The input data was collected and entered in the location clustering system. The data needed is a location address data within Nigeria which must contain the house number, street, district, state in line with the NIPOST standardized framework within the country.

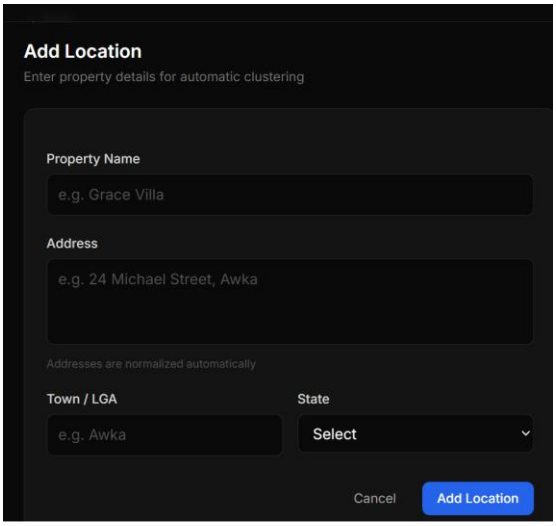

Fig 2a. Input representation of the system

The system produces real-time feedback in the form of table with which the property clusters can be displayed and the file could be downloaded in excel format for further information on the clusters and the confidence score.

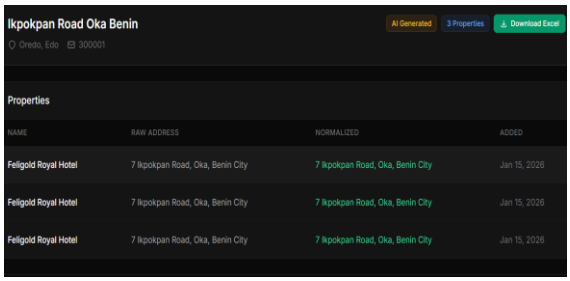

Fig 2b. Output representation of the system

© APR 2026 | IRE Journals | Volume 9 Issue 10 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I10-1716179

6.1 Dataset representation

The system was tested using:

1. Realistic Nigerian address samples
2. Addresses with spelling errors, abbreviations, and informal formats
3. Duplicate and incomplete address entries
4. Sample NIPOST postal code and district reference data

Test Case	Expected Result	Actual Result	Status
Address Cleaning	Noise-free text	Noise-free text	Pass
Postal Code Assignment	Correct NIPOST code	Correct NIPOST code	Pass
Clustering	Accurate cluster grouping	Accurate cluster grouping	Pass
Hotspot Mapping	Clear hotspot visualization	Clear hotspot visualization	Pass

Fig 3. Test data of the New System

VII . PERFORMANCE EVALUATION

The results demonstrate strong performance across all evaluation metrics, indicating that the proposed system effectively clusters address data in alignment with official district classifications while maintaining strong structural cluster integrity. The high precision and recall values indicate that the system accurately assigns address records to their appropriate NIPOST districts while ensuring minimal loss of relevant data. The F1-score further confirms the balanced performance of the system. Additionally, the clustering quality was further evaluated using the Silhouette Score, which measures intra-cluster compactness and inter-cluster separation. The system achieved a silhouette score of 0.794, indicating that address clusters are both well-defined and clearly separated. This result confirms that the DBSCAN-based approach effectively captures spatial density patterns while minimizing cluster overlap across NIPOST districts.

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Where:

True Positive (TP): Address correctly clustered into its real NIPOST district.

False Positive (FP): Address clustered into a district it does not belong to

False Negative (FN): Address belonging to a district but missed or marked as noise

True Negative (TN): Correctly excluded address

Silhouette score (S):

$$S = \frac{b - a}{\max(a, b)}$$

Where:

a = average intra-cluster distance

b = average nearest-cluster distance

The obtained silhouette score of 0.794 indicates strong cluster cohesion and clear separation between clusters. Since silhouette values range between -1 and 1, a value close to 0.8 reflects well-defined spatial groupings with minimal overlap. This confirms that the selected DBSCAN parameters (ε and MinPts) were appropriately tuned and that the resulting clusters represent meaningful geographic address groupings.

Table 1. Combined Evaluation Results

Evaluation Category	Evaluation Metric	Obtained Value	Interpretation
External Evaluation	Accuracy	0.990	Accuracy score of 0.981 measures the proportion of correctly classified addresses relative to the total number of evaluated address records (excluding noise). The high accuracy value indicates that the clustering output closely matches the official district assignments.

© APR 2026 | IRE Journals | Volume 9 Issue 10 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I10-1716179

External Evaluation	Precision	0.982	Indicates that 98.2% of addresses assigned to clusters truly belong to the correct NIPOST district, demonstrating minimal false-positive clustering.
External Evaluation	Recall	0.991	Shows that 99.1% of valid address records were successfully identified and clustered, indicating excellent coverage.
External Evaluation	F1-Score	0.986	Reflects a strong balance between precision and recall, confirming the overall robustness of the clustering approach.
Internal Evaluation	Silhouette Score	0.794	Demonstrates high intra-cluster compactness and clear inter-cluster separation, confirming good clustering structure and spatial coherence.

A confusion matrix was generated to compare the predicted cluster-based district assignments with the actual NIPOST district labels. The confusion matrix provides a detailed breakdown of classification performance by indicating correctly and incorrectly assigned address records.

The diagonal elements of the confusion matrix represent correctly classified addresses (True Positives), while the off-diagonal elements represent misclassifications (False Positives and False Negatives). The observed matrix shows strong diagonal dominance, indicating that the majority of address records were correctly assigned to their respective districts.

The minimal off-diagonal values suggest a low misclassification rate, thereby confirming that the DBSCAN clustering results closely align with the verified district ground truth. This validates the effectiveness of the density-based clustering approach for spatial address grouping.

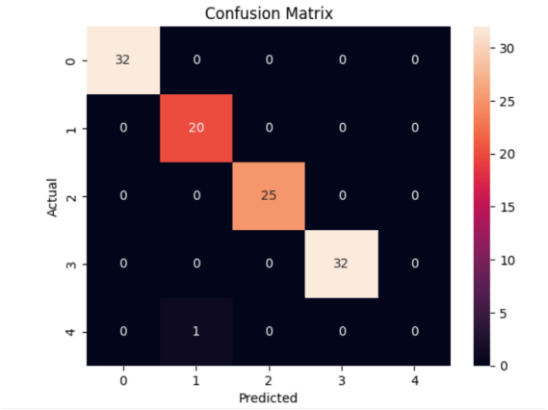

Fig 4. A Confusion Matrix

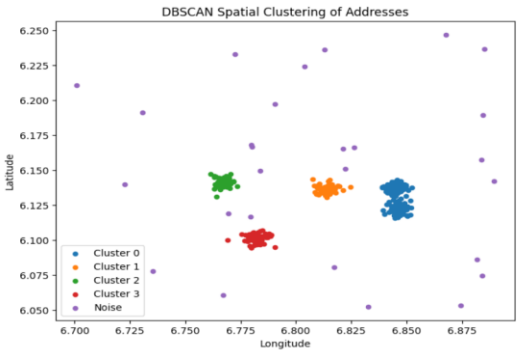

Fig 4b. Scattered Plot Diagram

VIII. CONCLUSION

This study presented the development of an intelligent location data clustering system that integrates Regular Expressions and the DBSCAN clustering algorithm to process unstructured Nigerian address data. The proposed system provides an automated framework for address standardization and spatial clustering and can support postal delivery optimization, urban planning, logistics operations, and emergency services.

REFERENCES

[1] Barrero, D. F., Camacho, D., & R-Moreno, M. D. (2009). *Automatic web data extraction based on genetic algorithms and regular expressions.*

[2] Carter, B. A., Hubert, L. A., & Walton, A. C. (2007). *Applications of regular expressions.*

[3] Frenz, C. M. (2008). *Introduction to searching with regular expressions.*

[4] Jalal, A. A. (2020). *Text mining: Design of interactive search engines based on regular expressions.*

[5] Riaño, D., Piñon, R., Molero-Castillo, G., Bárcenas, E., & Velázquez-Mena, A. (2020). *Regular expressions for web advertising detection.*

[6] Shahbaz, M., McMinn, P., & Stevenson, M. (2015). *Automatic generation of valid and invalid test data using web searches and regular expressions.*

[7] Evtimova, M. (2023, August). Validation algorithm for aligning postal addresses available on the Internet. In *Proceedings of the International Conference on Advanced Mathematical and Computational Sciences (ICAMCS)* (pp. 75–80). IEEE. https://doi.org/10.1109/ICAMCS59110.2023.00019

[8] Fatima, B. (2022, June 18). Nitda reaffirms drive for a digitally inclusive, innovation-led Nigerian economy. https://nitda.gov.ng/nitda-reaffirms-drive-for-a-digitally-inclusive-innovation-led-nigerian-economy/8986/

[9] Ghimire, B. R., Maharjan, B., Shrestha, S., & Nakarmi, S. (2021). Metric house address generation with semi-automatic geospatial web-based technology. *International Journal of Research Publications, 74*(1). https://doi.org/10.47119/ijrp100741420211851

[10] Geoapify GmbH. (2025, September). *Address standardization guide: What it is & methods.* https://www.geoapify.com/address-standardization-guide/

[11] Gupta, V., Gupta, M., Garg, J., & Garg, N. (2021). Improvement in semantic address matching using natural language processing. In *2021 2nd International Conference for Emerging Technology (INCET)* (pp. 1–5). https://doi.org/10.1109/INCET51464.2021.9456342

[12] Hassini, N., Mahmoudi, K., & Faiz, S. (2023, December). A hybrid approach for spatial information extraction from natural language text. In *2023 20th ACS/IEEE International Conference on Computer Systems and Applications (AICCSA)* (pp. 1–8). https://doi.org/10.1109/AICCSA59173.2023.10479267

[13] Hu, Y., Mao, H., & McKenzie, G. (2019). A natural language processing and geospatial clustering framework for harvesting local place names from geotagged housing advertisements. *International Journal of Geographical Information Science, 33*(4), 714–738. https://doi.org/10.1080/13658816.2018.1458986

[14] Kebe, A. M., Faye, R. M., & Lishou, C. (2019). Multi agent-based addresses geocoding for more efficient home delivery service in developing countries. In G. Mendy, S. Ouya, I. Dioum, & O. Thiaré (Eds.), *E-infrastructure and e-services for developing countries* (pp. 294–304). Springer.

[15] Lin, Y., Kang, M., Wu, Y., Du, Q., & Liu, T. (2019). A deep learning architecture for semantic address matching. *International Journal of Geographical Information Science, 34*(3), 559–576. https://doi.org/10.1080/13658816.2019.1681431