East African Scholars Journal of Engineering and Computer Sciences
Abbreviated Key Title: East African Scholars J Eng Comput Sci
ISSN: 2617-4480 (Print) & ISSN: 2663-0346 (Online)
Published By East African Scholars Publisher, Kenya
Volume-9 | Issue-2 | Mar-Apr-2026 | DOI: https://doi.org/10.36349/easjecs.2026.v09i02.004

OPEN ACCESS

Review Article

Healthcare Facility Clustering in Uganda: Geospatial, Ownership and Service-Based Segmentation for Evidence Based Planning

Abubakhari Sserwadda[1]*, Edwin Amanya[1], Teddy Phionah Nalwadda[1], Lincoln Keinebagaza[1], Matias Kabagambe[1], Excellence Favor[1]

[1]School of Engineering and Technology, Soroti University, Soroti, Uganda

Article History
Received: 19.02.2026
Accepted: 13.04.2026
Published: 29.04.2026

Journal homepage:
https://www.easpublisher.com

Quick Response Code

Abstract: In order to find coherent clusters of facilities based on geographic coordinate, user/inspector ratings, and the range of services provided, this study applies unsupervised machine learning to 6,520 health facility records from throughout Uganda. Three stable clusters that correspond with broad regional divisions in Uganda were produced by K Means following systematic cleaning, robust imputation of missing ratings, removal of obvious geographic outliers, feature scaling, and objective selection of the number of clusters using the Silhouette method: a Central/Eastern cluster with high facility density and diversity of services, a Northern cluster with a higher proportion of government-operated facilities and fewer specialized services, and a Southern/Western cluster with more balanced ownership and a relative strength in maternity and laboratory services. We also used centroid analysis, a multilabel binarization procedure to extract the prevalence of critical services, and the distribution of care systems (GOVT, PNFP, and PFP) to further characterize the clusters. The clusters' remarkable correlation to established regional borders was validated spatially using an administrative district shapefile. The findings are examined in light of Uganda's health planning requirements, and suggestions for focused action, more data gathering, and additional modelling are made.

Keywords: Clustering, Geospatial, Health Facilities and Planning.

1 INTRODUCTION

Uganda's regions continue to have unequal access to healthcare [1, 2]. National surveys and scholarly evaluations document the concentration of private facilities in urban areas, the shortage of resources in portions of Northern Uganda, and the disparities in non-governmental presence among districts [3].

Rigorous, data-driven assessments that combine facility location, service availability, and quality indicators would improve the nation's health planning architecture [4]. Clustering is used in this paper to produce such a synthesis. The study creates an actionable regional segmentation of Uganda's healthcare facilities by combining geospatial coordinates (latitude, longitude), a proxy quality indicator (facility rating), and a methodical extraction of services provided [5]. This segmentation can help district health officers (DHOs), the Ministry of Health (MoH), NGOs, and donors plan resources and interventions [6].

2 Related Work

Three related topics may be found in the literature: Uganda's national health system reports and surveys [1]; geospatial and machine learning applications to facility mapping in low- and middle-income countries (LMICs) [5]; and methodological resources for service availability analysis and clustering [7].

2.1 Uganda-Specific Planning and Facility Inventories

Uganda's Ministry of Health and the Uganda Bureau of Statistics (UBOS) publish regular health sector performance reports, facility inventories, and SARA/SDI style assessments [1, 2]. These documents provide the empirical basis that motivates clustering: uneven facility density, maternal health challenges in certain districts, and differences in service readiness driven by ownership or donor investment [3].

***Corresponding Author:** Abubakhari Sserwadda
School of Engineering and Technology, Soroti University, Soroti, Uganda

2.2 Geospatial and Machine Learning Applications

Researchers have used unsupervised learning, accessibility modeling, and spatial analysis on facility datasets in East Africa and other regions [5-8]. These studies demonstrate how robust maps of service deserts and specialized service clusters may be created by merging geolocation with service availability data [7].

2.3 Clustering, Scaling, and Validation Methods

When interpretability is necessary and the number of clusters is moderate, K Means and other centroidbased techniques are frequently employed for geographical segmentation [5-7]. Standard methods for verifying cluster quality include the Silhouette coefficient, mean intra-cluster distance, and visual inspection through GIS overlays [5-9]. When the dataset contains continuous, categorical, and textual columns, it is advised to use principled imputation for missing quality metrics and service vectorization using MultiLabelBinarizer [8].

3 Problem Definition and Preliminaries

3.1 Goal

Create interpretable cluster centers and service and ownership profiles for each cluster that policy actors may use by segmenting a national health facility registry into meaningful clusters that represent geographic, ownership, and service provision features [5, 1].

3.2 Data Schema

The working dataset is the uploaded CSV file Full collected hospital data 6K+ - All healthcare facilities (1).csv [10]. Kisejjere, R., Kakande, A., & Sserwadda, A. (2025). Uganda Healthcare Facilities' Features and Services [Data set]. Mendeley Data. https://doi.org/10.17632/58K7ZPWT84 Key fields used in this analysis are:

- Facility name (string)
- Services (text: comma-separated list)
- Latitude (float)
- Longitude (float)
- Rating (float, contains NaNs)
- Care system (categorical: GOVT, PRIVATE, NGO, FBO, etc.)
- Additional fields (operating _ hours, website, phone number, Subcounty, mode of payment) available but not central to the present clustering [1-7].

3.3 Constraints and Assumptions

- Although ratings are acknowledged as a nonstandardized measure (various raters, uncertain measuring technique), they are used as an ordinal/continuous proxy for quality [5-7]. To lessen sensitivity to extremes, missing data are imputed using the median [8].
- Geographical coordinates (about lat: −2 to +5, lon: 29 to 36) that are obviously outside of Uganda's valid bounding box were considered recording errors and eliminated.
- Services are heterogeneous textual items; before binarization, cleaning and normalization (trimming, lowercasing, and collapsing synonyms like "maternity" vs. "maternal care") were used [8].

4 Proposed Method

This section provides a full description of the data processing pipeline, the clustering methodology, and the analysis steps used to characterize clusters. The level of detail is intentionally high so the procedure is reproducible and understandable to nonspecialist stakeholders.

4.1 Data Cleaning and Preprocessing

Load dataset from the provided CSV. Inspect column types and missingness. Examine value ranges for latitude and longitude using min/max statistics and visual scatter plots to detect anomalies [5].

Rating imputation: compute the median rating from available entries (median = 4.0 in our dataset). Replace NaN rating values with the median. Rationale: median reduces influence of outliers and is appropriate for skewed or ordinal quality proxies [8].

Geographical outliers removal: filter out records with latitude or longitude outside the broadly accepted Uganda extent (conservative bounding box: latitude $\leq +10$ flagged; explicit outlier at lat ≥ 10 removed in the initial pass). Log the removed entries for audit [9].

Care system normalization: standardize categorical entries (e.g., 'Govt', 'GOVT', 'government' → GOVT) and create a small dictionary to map known synonyms [9].

Services cleaning and normalization: apply the following normalization to the services string column:

- Convert to lowercase
- Remove punctuation
- Replace common synonyms (e.g., 'maternal care' ↔ 'maternity', 'lab' ↔ 'laboratory') • split by commas or semicolons into service tokens
- Trim whitespace and deduplicate tokens per facility [8]

Binarization: use a MultiLabelBinarizer or equivalent to convert the services lists into a 0/1 matrix where each column is a normalized service [8].

Feature selection for clustering: for the initial geographic + rating clustering we used exactly three numerical features: latitude, longitude, and rating. This keeps the clusters interpretable in geographic terms while allowing rating to slightly influence separation where quality differences correlate with geography [5].

Feature scaling: standardize the numeric columns using StandardScaler (zero mean, unit variance) so that the rating does not dominate because of scale differences [5].

4.2 Determining the Number of Clusters (K)

We calculated clustering quality metrics for k between 2 and 10 using the following criteria to prevent arbitrary decisions: Davies Bouldin and the Calinski Harabasz Index as secondary checks, and the Silhouette Score as the major criterion. The three cluster solution was confirmed by Calinski Harabasz and visual examination of GIS overlays; the Silhouette analysis yielded a maximum at $k = 3$ with an estimated score of 0.593. The Silhouette Coefficient $s(i)$ for a single data point i is calculated as:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad (1)$$

Where:

- $a(i)$ is the average distance from i to all other data points in the same cluster (cohesion).
- $b(i)$ is the minimum average distance from i to data points in any other cluster (separation).

The overall Silhouette Score for the clustering solution is the mean of $s(i)$ over all data points.

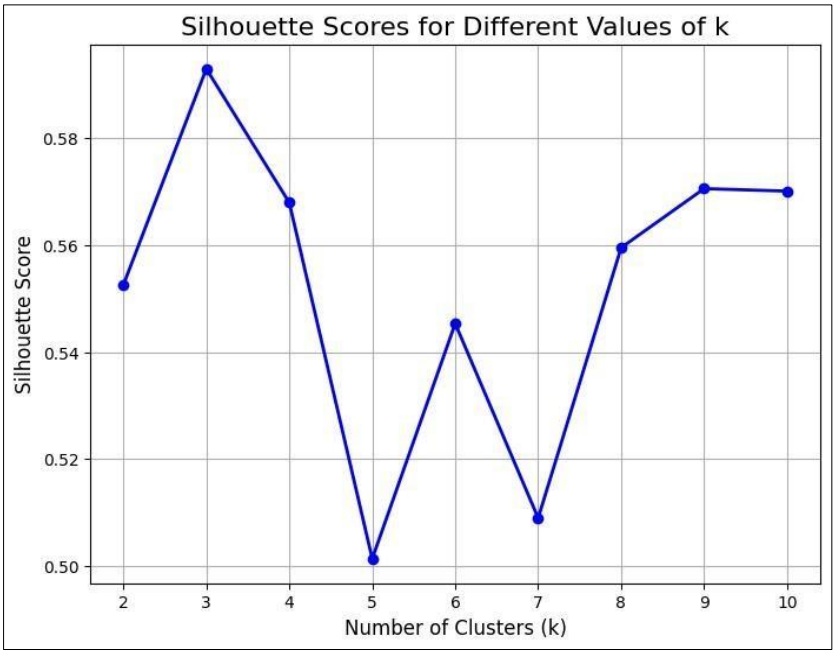

Figure 1: Silhouette Scores for Different Values of k

4.3 K-Means Clustering

With *n clusters* = 3 and *random state* = 42, KMeans was fit on the scaled features. We recorded:

- Per-facility cluster labels
- Number of facilities per cluster
- Cluster centroids in scaled space, then inverse transformed to original lat/lon/rating for interpretation

K-Means was chosen for clarity of centroids and ease of mapping cluster membership directly to geographic coordinates.

4.4 Characterization of Clusters

We computed the following per cluster:

- Counts and percent share of total facilities
- Mean and distribution of ratings
- Centroid latitude/longitude and mapping onto the Uganda shapefile
- Care system composition: percentage breakdown of GOVT vs PNFP vs PFP
- Service provision profile: for each normalized service, percentage of facilities in the cluster offering it

For service features we limited the detailed comparison to those services that appear in at least 10% of facilities in any cluster to reduce noise from rare services.

4.5 Geographical Validation

Where available, was used to overlay cluster points and cluster centroids over district boundaries. This step tested whether cluster patterns aligned with known administrative and regional boundaries.

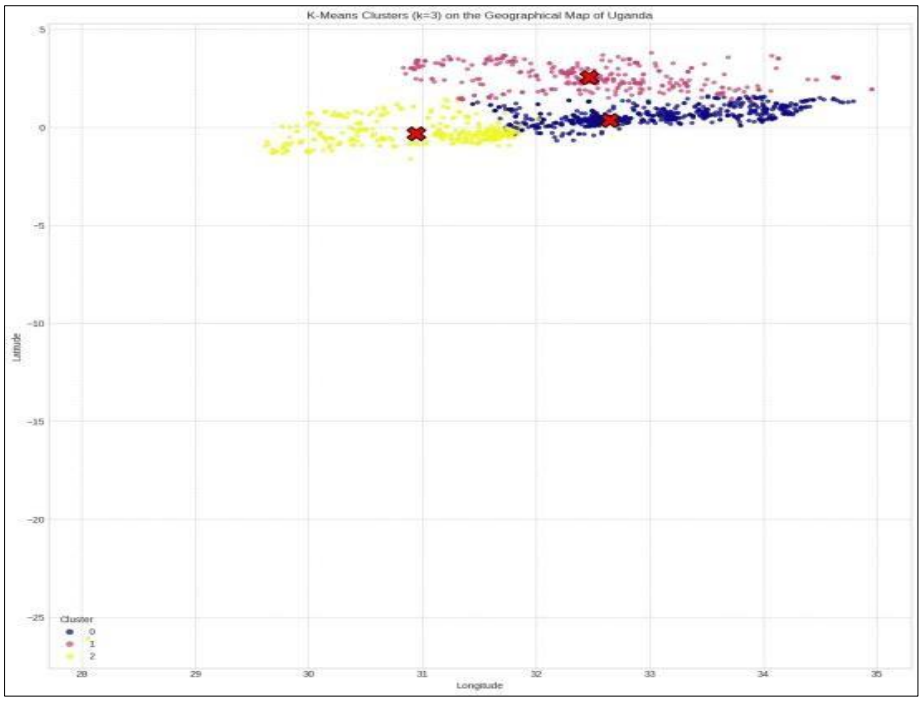

Figure 2: Geographical validation

4.6 Visualization

We created a number of figures that were meant to be included in the deliverable report and the notebook:

- Scatter plot of facilities colored by cluster (longitude vs. latitude), with centroids marked
- Stacked bar chart of care system composition per cluster.

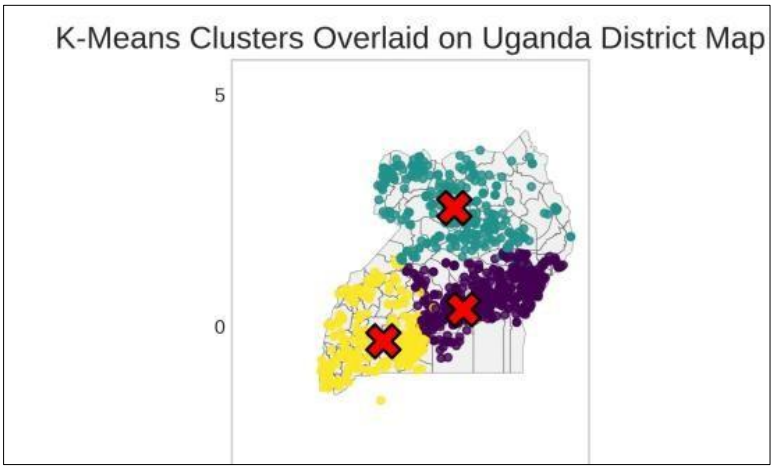

Figure 3: Placeholder for Cluster Scatter Plot • Grouped bar chart showing percentage offering of key services per cluster • Map overlay (shapefile) with clusters to visually validate geographic coherence

5 Experiments and Quantitative Results

This section lists the primary computational experiments, the key metrics observed, and quotations of the principal numeric results that inform the subsequent discussion.

5.1 Dataset after Cleaning

Original Records: 6,520
Removed Outliers: 1 (latitude ≥ 10)
Final records used for clustering: 6,519

5.2 Rating Imputation

Median rating used for imputation: 4.0
Rationale: robust central tendency measure appropriate for ordinal quality proxies

5.3 Optimal K selection

Silhouette scores (k=2..10): peak at $k = 3$ (score ≈ 0.593). Secondary metrics (Calinski-Harabasz and Davies-Bouldin) support separation quality for $k = 3$.

5.4 K-Means Clustering Results

Table 1: Cluster Centroid Metrics

Cluster	Count	Mean Lat	Mean Long	Mean Rating
0	5508	0.368	32.650	4.016
1	437	2.561	32.468	3.923
2	574	-0.308	30.941	3.808

(*n clusters* = 3)

5.5 Care System Distribution

Cluster 0: higher share of PRIVATE and FBO facilities, consistent with urban private market and mission hospitals.

Cluster 1: greater share of GOVT facilities, consistent with public provision strategies in the North

Cluster 2: mixed ownership with significant NGO and FBO presence.

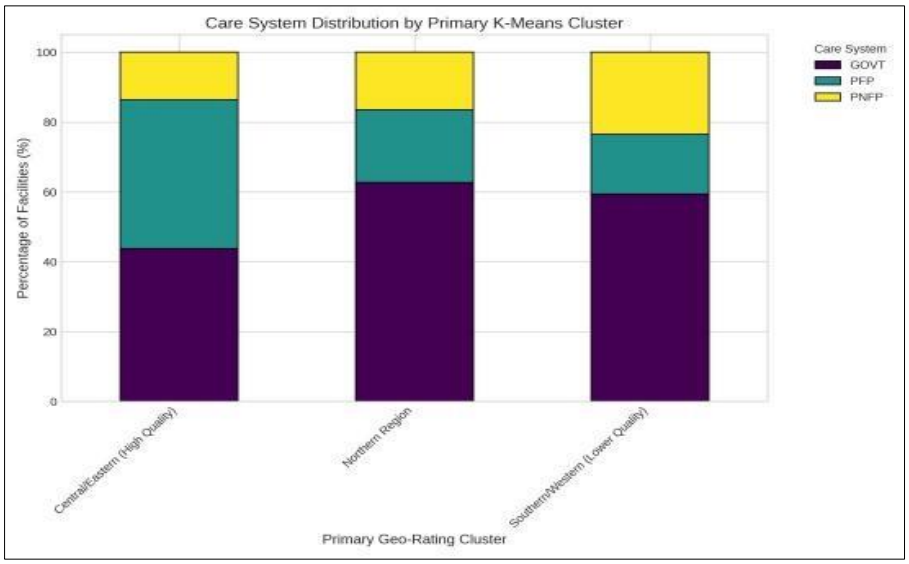

Figure 4: Care System Distribution per Cluster

5.6 Service Provision (Key Services, ≥ 10% Threshold)

Most commonly available services across clusters: outpatient, general care, maternity, child health, laboratory Cluster contrasts: Cluster 0 has broadest service portfolio and highest percentages across most services; Cluster 1 shows fewer diagnostic services and fewer specialized services; Cluster 2 displays stronger maternity and laboratory service presence relative to Cluster 1.

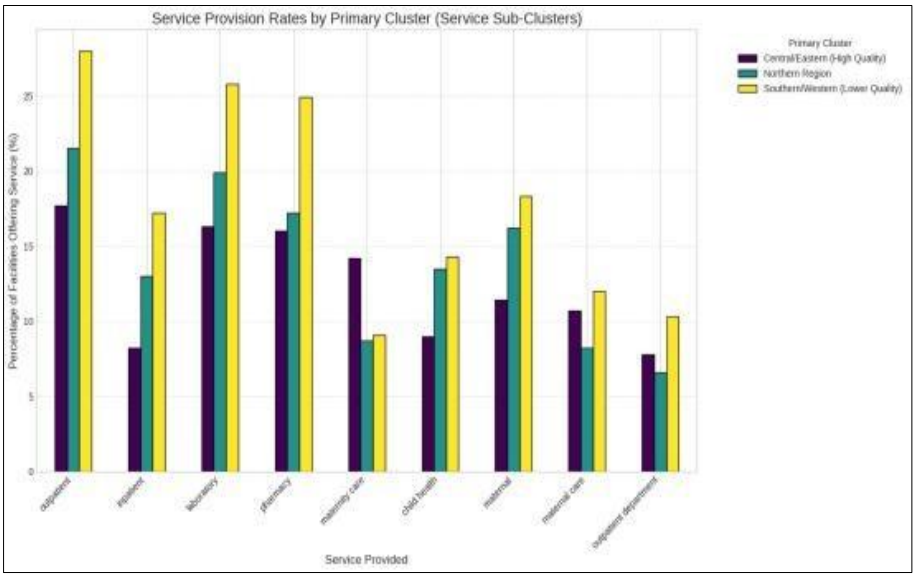

Figure 5: Service provision Profile per Cluster

5.7 Operational Hours Analysis

The analysis of 'Average Operational Duration by Primary K-Means Cluster' reveals that healthcare facilities in the Northern Region (Cluster 1) maintain the highest average duration (≈ 21.5 hours), suggesting a necessity for near 24/7 operation to ensure continuous access in less-serviced, rural areas. This contrasts with the more developed Central/Eastern (High Quality) Region (Cluster 0), which has an intermediate duration (≈ 18.4 hours), indicative of more structured hours potentially focused on specialized care, supported by a higher concentration of facilities. Finally, the Southern/Western (Lower Quality) Region (Cluster 2) exhibits the shortest duration (≈ 17.5 hours), possibly reflecting resource constraints or a lower demand for extended operating hours, ultimately showing that operational duration serves as an indicator of the pressure on facilities to guarantee accessibility in their specific geographical and developmental context.

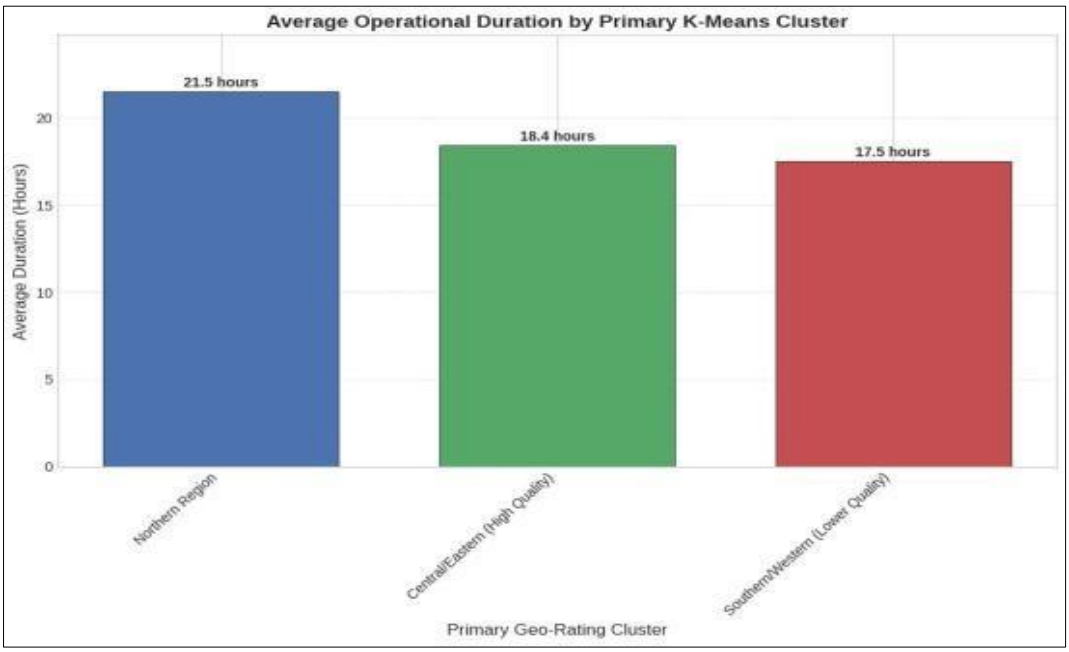

Figure 6: Operational Time Analysis

6. DISCUSSION (INTERPRETATION AND POLICY IMPLICATIONS)

This section synthesizes the findings and discusses them in practical terms for health planning in Uganda.

6.1 What the Clusters Tell Us about Regional Service Ecosystems

Three large service ecosystems that fit Uganda's socioeconomic geography are shown by the clustering. The majority of facilities and the richest service portfolio are found in the Central/Eastern cluster, which encompasses the Kampala metropolitan area and a number of urban center routes. This is in line with the notion that more specialized facilities and a more competitive private health market are supported by urban and periurban areas.

The Northern cluster continues to require focused investments in supply chain resilience, maternal and child health strengthening, and diagnostic capacity due to its greater reliance on government institutions and lower prevalence of diagnostics and specialized services. These insights could be used by international partners and government planners to prioritize referral system strengthening in the northern areas, implement training programs, or direct mobile diagnostic units.

With somewhat higher maternity and laboratory service rates, the Southern/Western cluster seems more balanced; this could be due to past NGO and FBO efforts in maternal and child health in those areas. Therefore, resource allocation strategies might prioritize growing outpatient and referral services while maintaining and improving current maternal health capacity.

6.2 Practical Recommendations for Decision Makers

- **Targeted infrastructure investments:** consider investing in laboratory capacity and emergency obstetric care in high-need districts identified from the Northern cluster.
- **Public– private partnership (PPP) opportunities:** examine districts in the Central/Eastern cluster for PPPs that can expand subsidized services where private facilities dominate but vulnerable populations remain.
- **Service readiness and quality monitoring:** harmonize rating procedures across districts to obtain a standardized quality metric for future iterations; consider linking facility ratings to Service Delivery Indicator (SDI) or SARA protocols.
- **Data improvement:** encourage routine geolocation audits and a standard taxonomy for services to reduce text noise and improve downstream analyses.

6.3 Limitations and Caveats

- **Rating data quality:** the ratings are used as a proxy for facility quality, but their source, rater standards, and recency vary; conclusions about service quality should therefore be cautious.
- **Service text heterogeneity:** despite cleaning and synonym mapping, the services field may hide inconsistent reporting; some services may be under- or over-reported.
- **Temporal dynamics omitted:** this analysis is cross-sectional. Facilities open, close, and change service offerings over time; longitudinal analysis is recommended for programmatic decision making.

7 Recommendations for Further Work

- Incorporate population denominator data (population per Subcounty/district) to compute facility-to-population ratios and identify true service deserts.
- Add road network and travel time modeling using OpenStreetMap and routing tools to evaluate effective accessibility rather than straight line distance.
- Ingest facility utilization metrics such as outpatient visits and inpatient admissions to align supply with demand.
- Standardize facility quality metrics using structured SARA/SDI audits to obtain more reliable quality measures.
- Explore hierarchical clustering of services to identify functionally similar facility groups (e.g., primary care hubs vs referral centers).
- Perform time series clustering if multiple snapshots of the dataset exist to detect growth, decline, or service shifts over time.

8. Expanded Discussion and Policy Implications

The study's clustering methodology offers a more thorough, data-driven understanding of the spatial and functional distribution of Uganda's healthcare facilities. The findings show structural trends inherent in Uganda's socioeconomic geography, population distribution, and health system evolution, going beyond just identifying three clusters. The largest and most service-diverse cluster, Central/Eastern, serves as an example of how infrastructure density, urbanization, and the expansion of the private sector work together to provide a healthcare environment that is more competitive, service-rich, and highly regarded. The higher ratings shown in this cluster are partly owing to the fact that Kampala and the neighboring metropolitan districts have historically drawn private investment because of their dense population, improved road systems, greater purchasing power, and stronger regulatory presence.

The Northern cluster, on the other hand, represents the residual consequences of past underinvestment, interruptions caused by conflicts, a greater geographic dispersion of settlements, and a greater dependence on public amenities. The region's lower facility density and predominance of public ownership have an impact on the services offered, especially the limited availability of specialized care and diagnostics. This is consistent with the results of MoH's SARA and AHSPR reports, which show ongoing differences in staffing, equipment functionality, and preparedness indicators between northern and central regions.

The healthcare environment is more balanced in the Southern/Western cluster. Important medical schools and regional referral hospitals are located in the area, which affects the availability of more reliable laboratory and maternity care services. The cluster's mixed ownership composition and intermediate average rating point to a changing health system in which PNFP and government players both have important responsibilities. The cluster's service profile is further supported by the existence of robust PNFP and FBO networks in western Uganda, particularly those connected to the Uganda Catholic Medical Bureau and Uganda Protestant Medical Bureau.

These trends indicate the significance of region specific policies from a policy perspective. While Northern Uganda might profit from focused diagnostic expansion, incentives for private sector penetration, and investments in health worker retention, Central Uganda might need more stringent regulation of the private sector and improved quality control procedures. In order to connect peripheral institutions to better-equipped centers, Western Uganda may need to expand its referral systems and provide logistical help.

The importance of incorporating geospatial analytics into national health planning is further shown by the clustering results. Automated analytics can be used by organizations like MoH, UBOS, and district authorities to estimate service shortfalls, identify underserved areas, model ideal facility placement, and assess how future investments might change the healthcare landscape. Additionally, NGOs and development partners can use the clusterbased insights to more precisely tailor interventions. For example, they could prioritize maternal services in parts of Cluster 2 or laboratory strengthening in Cluster 1.

9 Limitations and Opportunities for Future Work

Although the study offers a thorough analytical pathway, a number of limitations need to be noted. First, the dataset's facility evaluations might not be standardized or validated, which means they might represent user opinions rather than impartial quality indicators. Second, service-based clustering may be impacted by the granularity of service lists, which are self-reported by facilities. Third, the dataset lacks a temporal component, making it unable to record adjustments in facility distribution, ownership, or service improvements over time. Furthermore, the dataset lacks

data on patient load, staffing, stock outs, referral patterns, and health outcomes—all of which are necessary for a comprehensive assessment of the health system.

To improve the clustering, future research could incorporate trip time models, socioeconomic factors, population density rasters, or HMIS performance data. Road networks and accessibility models (such friction surfaces) could provide more realistic depictions of healthcare access in the actual world. The Ministry of Health could assess the effects of national health interventions over time by using temporal extensions to enhance trend analysis. Predictive models for resource allocation could be further improved by incorporating mobile-based patient satisfaction data or community health worker coverage. Lastly, using hierarchical or deep learning-based clustering techniques may reveal more complex geographical distinctions than the three groups found.

10. CONCLUSION

This study used a multilayered machine learning and geospatial analytical approach to examine Uganda's healthcare system. Three main clusters representing Central/Eastern, Northern, and Southern/Western Uganda were found by the study using systematic data cleaning, scaling, silhouette based optimization, K Means clustering, and service distribution analysis. In terms of facility density, ownership structures, ratings, and service availability, each cluster showed unique trends. Strong proof that the clusters are significant regional representations rather than arbitrary mathematical groupings was supplied by the geographical validation utilizing district shapefiles.

Policymakers, district health officers, nongovernmental organizations, and development partners that are looking for evidence-based approaches to improve Uganda's health system may find great value in the insights acquired. Cluster 0 stresses the necessity of regulatory supervision in urban private markets; Cluster 1 emphasizes the need for focused public sector investment and diagnostic growth; and Cluster 2 identifies chances to strengthen maternity services and referral networks. This study's overarching framework demonstrates how machine learning models may help with equitable monitoring, strategic planning, and health care optimization.

Future research should develop temporal clustering models, combine predictive analytics to forecast future service requirements, and include demographic, infrastructure, and health outcome data in order to progress this study. Computational analysis must be incorporated into planning procedures as Uganda continues to implement its Health Sector Development Plans (HSDP I–III) in order to provide all individuals with equitable healthcare.

11 Recommendations for Stakeholders

Building on the findings of this analysis, several targeted recommendations can support Uganda's ongoing efforts to strengthen healthcare systems and reduce regional disparities.

11.1 Ministry of Health (MoH)

- **Strengthen Diagnostic Capacity in Northern Uganda** – Expand laboratory infrastructure, supply chains, and biomedical maintenance systems in Cluster 1, which shows limited diagnostic service availability.
- **Develop a National Quality Rating Framework** – Standardize facility rating methods to improve accuracy, comparability, and usefulness for public information systems.
- **Integrate ML-Based Dashboards into DHIS2 –** Deploy clustering and geospatial analytics tools into district-level dashboards to help planners identify underserved areas in real time.

11.2 District Health Offices (DHOs)

Prioritize Outreach and Mobile Clinics in low density and underserved subcounties, particularly within Northern Uganda.

- **Promote Public–Private Partnerships (PPP)** in clusters with limited private-sector participation to improve service variety and coverage.

11.3 Private and PNFP Sectors

- **Encourage Equitable Expansion** by offering incentives for private facilities to operate in Northern regions.
- **Support Training and Capacity Building** especially in specialized services such as radiology, maternal health, and emergency care.

11.4 Development Partners and NGOs

- **Focus interventions** on high-need clusters such as expanding maternal services in specific Southern/Western districts.
- **Invest in digital data collection systems** to improve accuracy of service reporting and readiness assessments.

12 Implications for Future Research and Data Development

The machine learning methods used in this work show great promise for assisting with national health planning. However, larger datasets might greatly improve analysis in the future. Integrating is one of the main opportunities.

- **Population distribution data** to compute facility-to-population ratios at parish and subcounty levels.
- **Road networks and travel-time models** to evaluate true accessibility beyond geographic distance.

- **Health outcomes metrics** such as disease prevalence, maternal mortality, or immunization coverage.
- **Staffing data** including clinician-to-patient ratios.
- **Facility functionality data** such as electricity reliability, drug-stock consistency, and equipment uptime.
- **Temporal analytics** to track how clusters evolve over time, especially during policy shifts or epidemics.

These enhancements would allow Uganda to transition from descriptive clustering into predictive modelling, enabling planners to anticipate future needs and optimize interventions.

13 Final Remarks

This study's thorough clustering analysis shows how useful machine learning and spatial analytics are for comprehending Uganda's healthcare system on a large scale. The methodology offers a multifaceted view that traditional reports frequently overlook by integrating geolocation, facility ratings, ownership structures, and service availability. Furthermore, the approach's robustness is validated by the clusters' alignment with actual administrative and socioeconomic regions. Integrating cuttingedge analytical techniques would be essential as Uganda continues to carry out the Health Sector Development Plans (HSDP II and III). This study adds to the increasing amount of evidence that supports data-driven decision-making in low- and middle-income health systems and lays the groundwork for such integration.

REFERENCES

- J. e. a. Aboagye, "Spatial analysis of healthcare facility distribution in Ghana," 2020.
- M. &. S. P. Rahman, "Machine learning for public health planning.," 2020.
- M. e. a. Haque, " Health service readiness and clustering analysis.," 2019.
- M. o. H. Uganda, "Annual Health Sector Performance Report.," 2022.
- M. S. o. P. H. ., "Service Availability and Readiness Report," 2020.
- R. K. A. &. S. A. ..Kisejjere, "Uganda Healthcare Facilities' Features and Services [Data set].," Mendeley Data., (2025). [Online]. Available: https://doi.org/10.17632/58K7ZPWT84.
- U. M. o. Health, "Health Facility Inventory Report.," 2021.
- UBOS, "Statistical Abstract.," 2021.
- UBOS," Uganda Demographic and Health Survey.," 2022.
- Y. e. a. Molla, "Spatial distribution of maternal health services in East Africa," 2020.

Cite This Article: Abubakhari Sserwadda, Edwin Amanya, Teddy Phionah Nalwadda, Lincoln Keinebagaza, Matias Kabagambe, Excellence Favor (2026). Healthcare Facility Clustering in Uganda: Geospatial, Ownership and Service-Based Segmentation for Evidence Based Planning. *East African Scholars J Eng Comput Sci, 9*(2), 39-47.