Clustering Techniques for Modeling Village Infrastructure Development

Nataliia Kussul[1,2], Bohdan Potuzhnyi[1,3] and Vlada Svirsh[1,3]

[1] Institute of Physics and Technology, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Beresteiskyi Ave, 37, Kyiv, 03056, Ukraine
[2] Space Research Institute NASU-SSAU, Ukraine
[3] Engineering and Informational Technology Department, Bern University of Applied Science, Quellgasse, 21, Biel/Bienne, 2502, Switzerland

Abstract

This paper presents a comprehensive framework for enhancing rural village infrastructure in Ukraine, utilizing advanced clustering techniques alongside OpenStreetMap (OSM) geospatial data analysis. At its core, the study aims to categorize villages based on infrastructure quality, identify developmental gaps, and inform strategic planning for improvement. Employing spatial data analysis and machine learning, particularly through a novel application of the KMeans clustering method, the research outlines a systematic approach to analyzing infrastructure disparities. It innovatively combines proximity measures and graph-based details to create detailed infrastructure profiles for each village, enabling precise identification of areas in need. The methodology includes a unique reorganization step post-cluster formation to rank clusters by infrastructure quality, a move that enhances the study's ability to target strategic development initiatives effectively. Tested across various Ukrainian villages, the approach not only highlights infrastructure imbalances but also suggests a model for broader application, potentially integrating with socio-economic data for comprehensive development strategies. This work underscores the importance of informed policymaking and cost-effectiveness, aiming to contribute significantly to the strategic enhancement of rural infrastructure in diverse economic conditions.

Keywords

Spatial Data Analysis, Machine Learning, Clustering, Geoinformatics, Village Classification

1. Introduction

In recent years, the development of rural infrastructure has emerged as a critical challenge for policymakers and researchers alike, not only for Ukraine, a country grappling with the profound impacts of conflict and economic hardship due to its ongoing war with Russia but also for nations worldwide. The disparity in infrastructure quality between urban and rural areas exacerbates social and economic inequalities, hindering the overall development of the nation. This study aims to address this challenge by introducing an innovative framework for modeling and enhancing village infrastructure, leveraging advanced clustering techniques and OpenStreetMap (OSM) geospatial data. Our primary objectives are to:

1. Develop a methodology for categorizing villages based on infrastructure quality, utilizing spatial data analysis and machine learning clustering techniques.
2. Identify developmental gaps within rural communities by analyzing geospatial data from OSM, and setting benchmarks for infrastructure improvement.
3. Evaluate the effectiveness of the proposed clustering technique in informing strategic development planning for rural areas.

COLINS-2024: 8th International Conference on Computational Linguistics and Intelligent Systems, April 12–13, 2024, Lviv, Ukraine

✉ nataliia.kussul@lll.kpi.ua (N. Kussul); bohdan.potuzhnyi@gmail.com (B. Potuzhnyi); vlada.svirsh25@gmail.com (V. Svirsh)
iD 0000-0002-9704-9702 (N. Kussul); 0009-0001-2282-7012 (B. Potuzhnyi); 0009-0004-4006-2705 (V. Svirsh)

Ukraine's rural regions, characterized by their diverse and dynamic infrastructure, present unique challenges that necessitate sophisticated analytical tools. The traditional approaches to infrastructure assessment often fall short of capturing the nuanced differences across rural communities, leading to generalized and sometimes ineffective intervention strategies. By integrating spatial data analysis with machine learning, this research offers a more nuanced and comprehensive tool for policymakers and developers to identify critical infrastructure needs and prioritize interventions.

Furthermore, this study is part of a broader project, "Information Technologies of Geospatial Analysis for the Development of Rural Areas and Communities", commissioned by the Ministry of Education and Science of Ukraine. This project underscores the national significance of enhancing rural infrastructure as a means to foster equitable development across the country.

The outcomes of this study are expected to validate a methodology that can be adapted for broader geographic applications, integrating additional socioeconomic data layers, and paving the way for future automation and real-time updates. By doing so, this research is set to make significant contributions to the strategic enhancement of village infrastructure, with an emphasis on cost-effectiveness and informed policymaking, particularly under the current economic conditions faced by Ukraine.

2. Related works

In recent years, the intersection of geospatial analysis, machine learning (ML), and rural infrastructure development has emerged as a critical research domain, driven by the urgent need to address disparities between urban and rural areas. This literature review delves into pioneering studies that have laid the groundwork for our innovative framework aimed at enhancing village infrastructure through clustering techniques and OpenStreetMap (OSM) geospatial data. Our primary objectives include developing a robust methodology for village categorization based on infrastructure quality, identifying developmental gaps, and informing strategic planning for rural areas, particularly in the context of Ukraine's unique socio-economic challenges.

2.1. Multifunctional rural development (MRD) and spatial analysis

The concept of MRD offers a holistic approach to rural development, acknowledging the multifaceted roles that rural areas play beyond agriculture, including cultural preservation, recreation, and biodiversity conservation. Long et al. (2022)[1] provide an extensive review of MRD in China, showcasing its significant impact on rural revitalization by promoting economic diversification and sustainability. Their work emphasizes the importance of spatial data analysis in understanding the spatial distribution of multifunctional activities, which directly aligns with our objective to categorize villages by infrastructure quality using spatial analysis. By examining the spatial complexities of multifunctionality, our study extends the application of MRD principles to the development of rural infrastructure in Ukraine, proposing a data-driven approach to identify and prioritize areas for intervention.

2.2. Youth employment, empowerment, and socioeconomic integration

Addressing the socioeconomic dimensions of rural development, Geza et al. (2022)[2] explore the dynamics of youth employment and empowerment in South Africa's rural economy. Their scoping review highlights the critical need for inclusive strategies that integrate youth into the rural economy, underscoring the potential of combining geospatial data with socioeconomic factors to create more comprehensive development strategies. This perspective is integral to our methodology, which seeks to incorporate socioeconomic data layers into our clustering analysis, thereby providing insights into how infrastructure development can support broader societal goals, including youth empowerment and employment.

2.3. Community-based tourism and infrastructure development

Community-based tourism (CBT) represents a sustainable approach to leveraging local assets for economic development while preserving cultural and natural heritage. Arintoko et al. (2020)[3] discuss strategies for CBT village development in Indonesia, emphasizing the role of strategic planning and community participation. This study's emphasis on local data and community characteristics mirrors our approach to utilizing OSM geospatial data for village infrastructure development. By incorporating community engagement and local characteristics into our analysis, we aim to identify infrastructure gaps and opportunities for sustainable tourism development, aligning with broader community development goals.

2.4. Techno-economic analyses and smart village models

The transition towards "Smart Villages" involves integrating technology into rural development strategies to enhance sustainability, governance, and quality of life. Aziiza & Susanto (2020)[4] and Stojanova et al. (2021)[5] provide valuable insights into the smart village model, highlighting the importance of technological solutions and local governance in achieving sustainable rural development. Our study draws upon these insights, applying machine learning techniques to analyze geospatial data for infrastructure development. By envisioning a future where technological integration supports rural infrastructure, our methodology aligns with the smart village concept, advocating for data-driven decision-making to inform policy and planning.

2.5. Geospatial analysis in rural development

Emerging studies on geospatial analysis, such as the work by Yailymova et al. (2023)[6], have demonstrated the power of spatial data in assessing the quality of life and infrastructure in rural areas. These studies underscore the potential of geospatial technologies to identify regional disparities and inform targeted interventions. Our research builds upon these foundations, leveraging OSM data and machine learning to offer a nuanced view of rural infrastructure in Ukraine, highlighting the utility of spatial analysis in guiding development efforts under diverse economic conditions.

2.6. Clustering techniques and their application

The application of clustering techniques in rural development offers a methodological advancement that allows for the nuanced categorization of villages based on various attributes, including infrastructure quality. Studies like Golalipour et al. (2021)[7] review the evolution of clustering methods across different domains, providing a theoretical basis for our enhanced clustering approach. By integrating these techniques with geospatial analysis, our study aims to refine the categorization process, making it a cornerstone of our methodology for assessing and improving village infrastructure.

In synthesizing the insights gleaned from the examined body of literature, it becomes evident that the intersection of spatial data analysis, machine learning, and rural development holds profound potential for transformative impact on village infrastructure enhancement strategies. The incorporation of multifunctional rural development principles, socio-economic integration, community-based tourism approaches, and the adoption of smart village models underscores a shift towards more inclusive, sustainable, and technology-driven rural development practices. Our study leverages these insights to propose an innovative framework that not only categorizes villages based on infrastructure quality using advanced clustering techniques but also integrates open-source geospatial data for comprehensive analysis. By doing so, we aim to address the nuanced challenges of rural infrastructure development in Ukraine, offering a model that is both scalable and adaptable to other contexts.

3. Methods and materials

The core of our research revolves around a rigorous methodology that leverages a fusion of spatial data analysis, machine learning clustering techniques, and rich geoinformatics data from OpenStreetMap (OSM). This section outlines the methods and materials used in our study, providing a blueprint of our approach to modeling village infrastructure development.

3.1. Materials

3.1.1. Geospatial data overview

The dataset, central to our study on modeling village infrastructure development, encompasses 28381 entries, each entry corresponding to a unique geographical location within rural Ukraine. These entries are meticulously detailed across 41 informative columns, offering a panoramic view of the rural infrastructure landscape through the lens of geospatial analytics. However, in this research, we will only use 17 descriptions of type distance to the closest object and 13 graphed descriptions.

3.1.2. Infrastructure proximity measures

A significant portion of the dataset results from proximity measures, quantifying the closeness of each village to various critical infrastructure elements, thereby serving as a fundamental component for spatial analysis. All these objects are grouped by types and described in Table 1. These objects were created and discussed as part of the research conducted by Yalimova et al. [6].

Table 1
Infrastructure proximity data

Type	Objects	Description
Roads	RD_m1_NEAR, RD_m2_NEAR, RD_m3_NEAR	These objects describe the distance to major, regional, and rural roads.
Cities	CITY2_NEAR, Kyiv_NEAR_	These objects describe the distance to the nearest city and capital of Ukraine
Parks	LokPark_NE, NatPark_NE, regPark_NE	These objects describe the distance to the nearest local park, national park, and regional park
Elevators	Elevators_	Distance to the closest elevator
Kindergarten	Kinder_NEAR	Distance to the closest kindergarten
Bank	Bank_NEAR_	Distance to the closest bank

Church	Cerkva_NEA	Distance to the closest church
Education	Education_	Distance to the education
Hotels	Hotels_NEA	Distance to the closest hotel
Library	Library_NE	Distance to the closest library
Hospital	Likarni_NE	Distance to the closest hospital
Shop	Magaz_NEAR	Distance to the closest shop

3.1.3. Graph-based representations

Graph-based columns (graph_city, graph_local_park, etc.) contain JSON-formatted data in the format of an array, offering an in-depth look at the nearest facilities of various types. This detailed representation allows for an advanced analysis of the infrastructure network, including the examination of connectivity and the spatial distribution of essential services and amenities. A more detailed description of each type of column and an example of the object is given in Table 2. In general, all these objects will have the 'id_type' which describes from which column of the original data the 'id' was received from, as result it gives us the ability to retrieve new data without the need to rebuild the whole graphed structure. 'id', 'distance', 'pos_x', 'pos_y' are always presented in this dataset.

Table 2
Graph-based data

Type	Example object	Description

Graph_city	{
	"id_type": "admin4Pcod",
	"id": "UA5323810100",
	"distance": 15874.531539053276,
	"pos_x": 814898.792138855,
	"pos_y": 5565562.434442552
	}
		Just a standard description of the city that is located near a village with all basic descriptions presented.

Graph_local_park	{
	"id_type": "KodPZF",
	"id": "0253UA0708041",
	"distance": 3695.6829331930207,
	"KatObPZF": "Reserve",
	"AreaPZF": 120,
	"pos_x": 807527.095508027,
	"pos_y": 5577813.50041332
	}
		Description of the local parks, except standard, has Area description, and KatObPZF.

Graph_national_park,
graph_regional_park
	{
	"id_type": "KodPZF",
	"id": "0153UA0200001",
	"distance": 10908.01990446648,
	"AreaPZF": 12028.42,
	"pos_x": 813282.7981963563,
	"pos_y": 5570259.240116742
	}
		Description of the national and regional parks, has area inside of it

Graph_bank,
graph_kindergarten,
graph_library
	{
	"id_type": "osm_id",
	"id": "668736377",
	"distance": 24589.561415524415,
	"pos_x": 830942.7863028484,
	"pos_y": 5593380.706043951
	}
		Description of the bank, kindergartens, and libraries

Graph_church,
graph_edu,
graph_hotel,
graph_medicine,
graph_shop
	{
	"id_type": "osm_id",
	"id": "298759370",
	"distance": 6992.808427912616,
	"fclass": "class",
	"pos_x": 816907.4829874948,
	"pos_y": 5580679.407344543
	}
		Description of the churches, education, hotel, medicine, and shop units, on the place of the ‘fclass’ will be the object that describes to which class it is corresponding for example in edu it will be school, for church it will be Christian, and so on.

Graph_elevator	{
	"id_type": "id",
	"id": 746,
	"distance": 2657.6559562758794,
	"pos_x": 807667.363751653,
	"pos_y": 5579215.811847648
	}
		Description of the elevator object

3.2. Methods

3.2.1. Custom vector creation

In this part, we create unique vectors for each type of Point of Interest (POI) by combining data from two sources: graph-based spatial details mentioned in Section 3.1.3 and distance measurements to key infrastructure features discussed in Section 3.1.2. This combination provides a detailed infrastructure snapshot for every village, making it possible to group villages more accurately based on their infrastructure characteristics. Our aim is to capture a detailed picture of village infrastructure to support better analysis and grouping.

3.2.2. Clustering

Here, we enhance the standard KMeans clustering approach by introducing a step that reorganizes clusters after they are formed. This step ranks clusters by the quality of their infrastructure, from most to least in need of improvement, using data points (centroids) from the clustering process. This method not only makes it easier to understand the infrastructure status of each cluster but also offers a systematic way to analyze various types.

3.2.3. Usage of heatmap with normalized data

We use heatmaps to visualize the results of our clustering, focusing on the normalized data points for each cluster. This technique simplifies the presentation of complex data, highlighting how similar or different clusters are regarding infrastructure. By normalizing the data, we ensure a fair comparison across clusters, making the heatmap an effective tool for quickly identifying key trends and differences.

3.2.4. Distribution analysis

This section examines how villages are spread across the clusters to understand the overall state of infrastructure development. Using histograms, we can see the shape and spread of infrastructure quality among the villages, identifying both common trends and outliers. This analysis helps quantify the variety in infrastructure quality, pointing out areas that may need more attention and guiding where improvements could be made. Through this careful examination, we aim to provide a clear picture of infrastructure development needs.

4. Experiment

In this section, we outline the methodological framework of our experiment, focusing on the creation of vectors for Points of Interest (POIs), the clustering process, and the computation of village infrastructure quality scores. Our approach synthesizes proximity measures with graph-based spatial details to construct a detailed infrastructure profile for each village. The process unfolds in the following steps:

1. Preprocessing and Vector Creation: Initially, we preprocess the geospatial data, addressing any inconsistencies and missing values to ensure uniformity. Subsequently, we create distinct vectors for various POIs by amalgamating distance measures to key infrastructure elements with graph-based spatial insights. This step forms the bedrock for our comparative analysis of village infrastructure.
2. Clustering of Vectors: Utilizing the crafted vectors, we engage in clustering, employing an enhanced version of the KMeans algorithm complemented by a reorganization step. This process categorizes villages into clusters based on infrastructure quality, enabling us to identify specific areas needing attention.
3. Quality Score Computation: Post-clustering, we calculate an infrastructure quality score for each village. This score integrates the average infrastructure category with cluster rankings, providing a quantifiable measure of each village's infrastructure status.

This concise overview encapsulates our experimental approach, laying a structured pathway to dissecting and understanding the complex landscape of rural infrastructure through a data-driven lens.

4.1. Creation of vectors for POI types

The foundation of our analysis lies in the construction of unique vectors for various POIs by merging distance measures to key infrastructure features with graph-based spatial details. This amalgamation provides a nuanced snapshot of each village's infrastructure, facilitating a granular comparison and grouping based on infrastructure characteristics. The vectors incorporate distances to the nearest facilities of various types, as outlined in Table 3. Notably, for graph-based columns such as 'graph_city' and 'graph_local_park', JSON-formatted data provides a detailed look at the nearest facilities, enhancing our spatial analysis capability. In instances where data for a particular POI within a village is missing, a placeholder value of 2147483647 is used to signify the absence, ensuring consistency in vector dimensions across the dataset. The terms obj_{1..5} represent the distances to the nearest objects of a given POI category to the village, further enhancing our vector.

Table 3
POI type vector format

Type	Vector
Roads	(RD_m1_NEAR, RD_m2_NEAR, RD_m3_NEAR)
Cities	(Kyiv_NEAR_, CITY2_NEAR, obj_1, obj_2, obj_3, obj_4, obj_5)

Local parks	(LokPark_NE, obj_1, obj_2, obj_3, obj_4, obj_5)
National parks	(NatPark_NE, obj_1, obj_2, obj_3, obj_4, obj_5)
Regional parks	(regPark_NE, obj_1, obj_2, obj_3, obj_4, obj_5)
Combined Parks	(cluster value of local parks, cluster value of regional parks, cluster value of national parks)
Banks	(bank_NEAR_, obj_1, obj_2, obj_3, obj_4, obj_5)
Church	(cerkva_NEA, obj_1, obj_2, obj_3, obj_4, obj_5)
Edu	(education_, obj_1, obj_2, obj_3, obj_4, obj_5)
Elevators	(elevators_, obj_1, obj_2, obj_3, obj_4, obj_5)
Hotels	(hotels_NEA, obj_1, obj_2, obj_3, obj_4, obj_5)
Kindergartens	(kinder_NEA, obj_1, obj_2, obj_3, obj_4, obj_5)
Library	(library_NE, obj_1, obj_2, obj_3, obj_4, obj_5)
Medicine	(likarni_NE, obj_1, obj_2, obj_3, obj_4, obj_5)
Shop	(magaz_NEAR, obj_1, obj_2, obj_3, obj_4, obj_5)
Overall	(roads, cities, parks, banks, churches, edu, elevs, hotels, kindergartens, libraries, medicine, shops)

4.2. Clustering methodology

Our clustering approach extends beyond the standard application of KMeans. After initial cluster formation, we undertake a novel reorganization step where clusters are ranked by the aggregated quality of their infrastructure, determined using the centroids obtained from the clustering process. This re-ranking allows us to systematically assess infrastructure quality across clusters, from those most in need of improvement to the least. This step not only aids in the interpretability of the clusters but also in identifying specific infrastructure imbalances.

To ensure the validity of our clustering approach, we utilized the silhouette score as a metric to evaluate the cohesion and separation of clusters, confirming the optimal number of clusters (k=10) for our analysis. Additionally, we meticulously handled missing or incomplete data points in the dataset to maintain the integrity of our clustering process.

4.3. Heatmap visualization and normalization

For an intuitive representation of our clustering results, we employed heatmaps to visualize the normalized centroids for each cluster. This method allows for a straightforward comparison of infrastructure presence across clusters, highlighting disparities and commonalities with ease. By normalizing the data before visualization, we ensure a balanced comparison, facilitating the identification of pivotal infrastructure trends and discrepancies.

4.4. Distribution analysis and quality score calculation

The distribution of villages across the identified clusters was examined to gauge the overall landscape of infrastructure development. Through histograms, we visualized the spread of infrastructure quality, identifying prevalent trends and outliers. This analysis was pivotal in quantifying the diversity in infrastructure quality, and pinpointing areas of potential focus for development initiatives.

A crucial outcome of our experiment is the formulation of a quality score for each village, calculated as follows:

$$quality = 0.8 * \left(\frac{1}{12}\sum_{1}^{12} category_i\right) + 0.2 * cluster \quad (1)$$

This formula integrates the average infrastructure category score with the cluster ranking, offering a nuanced metric of infrastructure quality. This composite score encapsulates both the diversity of available infrastructure within a village and its comparative standing among other villages.

4.5. Implementation insights

Throughout the experiment, we encountered and overcame several challenges, particularly in data preprocessing and the handling of missing data. The adoption of Python libraries, such as scikit-learn for clustering and seaborn for heatmap generation, was instrumental in our analysis. Custom scripts were developed to transform JSON-formatted spatial data into analyzable vectors, showcasing the adaptability and depth of our methodological framework.

5. Results

As a result of our experiment, we obtained 17 visualizations, including 2 3D plots for roads and parks, 14 heatmap plots, and 1 histogram plot. We begin our analysis with Figure 1, which effectively illustrates the creation of clusters and their correspondence to the anticipated outcomes. A notable observation from the figures is the identification of the types of Points of Interest (POIs) that are more prevalent in Ukraine. This results in fewer groups with a lower availability of certain POIs, such as those depicted in Figure 3 (local parks), Figure 8 (churches), Figure 9 (education), Figure 11 (hotels), Figure 12 (kindergartens), Figure 14 (medicine), and Figure 15 (shops).

Diving deeper into Figure 1, we see the validity of our approach as well as the first meaningful insights that could be derived. For instance, most of the villages have close access to regional and rural roads (types 2 and 3, respectively). We can clearly define the set of villages within cluster 3, which need better connections to regional roads. As a result, these villages are likely to develop more evenly, and building regional roads would require less financial investment than constructing national roads (road type 1).

Examining Figure 2, we can discern that there are three clusters of villages with proximity to five cities, two clusters with four nearby cities, one cluster with three, and two clusters with two cities in the vicinity. This distribution suggests that many villages benefit from a variety of cities close to them, which could significantly ease future development efforts. The presence of nearby cities can facilitate infrastructure improvements and enhance the overall well-being of residents, as the establishment of cities often entails substantial investment and the advancement of related amenities.

Continuing our interpretation of the results for the upcoming Figure 3, which delineates the clusters created for local parks, we can observe that most village clusters have at least three local parks nearby. This finding underscores the fact that village residents have quick access to places that enhance both the emotional and overall well-being of community members. While parks are generally accessible, indicating room for improvement, it is important to note that such enhancement is not just a matter of need but also pertains to the emotional well-being and attractiveness of the region.

Examining the descriptions of national and regional park clusters, we see that most villages do not have quick access to national and regional parks. While proximity to these places may not be vital for the villagers' day-to-day lives, being near these areas undoubtedly promotes the overall development of the surrounding region. It enhances tourism, which in turn stimulates the construction of hotels and shops, supporting the local economy and capitalizing on this interest.

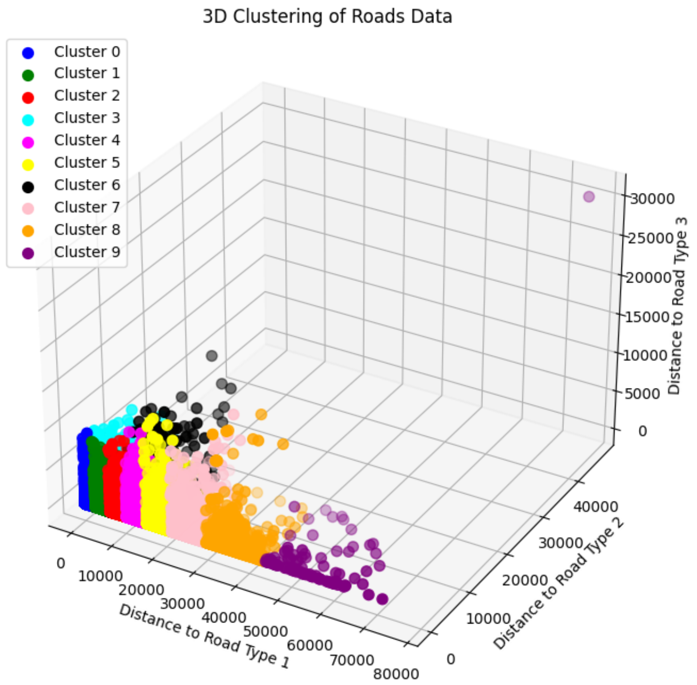

Figure 1: Clusters of road types, with each color representing a different cluster. The order of the clusters can be seen on the top-left side, where 0 indicates the cluster with the closest distance to the objects, and 9 is the farthest from all other roads.

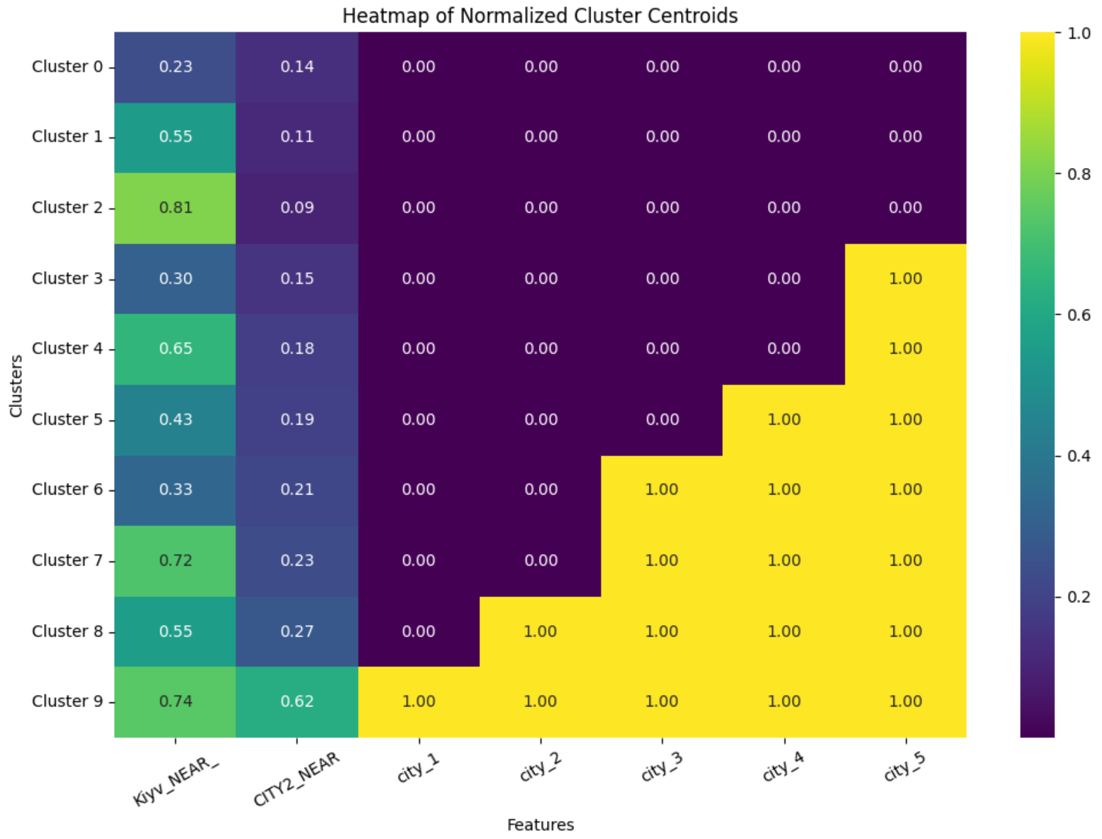

Figure 2: Heatmap of city clusters, where each row represents the distribution from 0 to 1 for each type of value within the vector, and each column indicates the corresponding parameter.

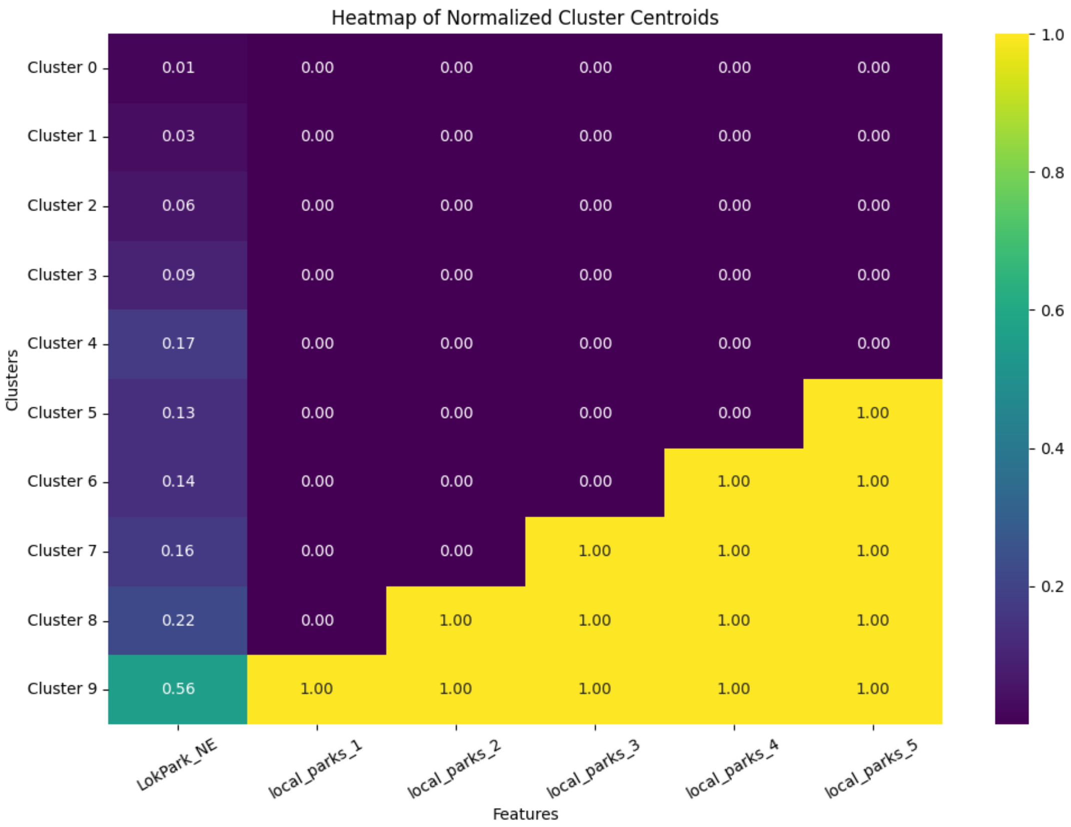

Figure 3: Heatmap of local park clusters, where each row represents the distribution from 0 to 1 for each type of value within the vector, and each column indicates the corresponding parameter.

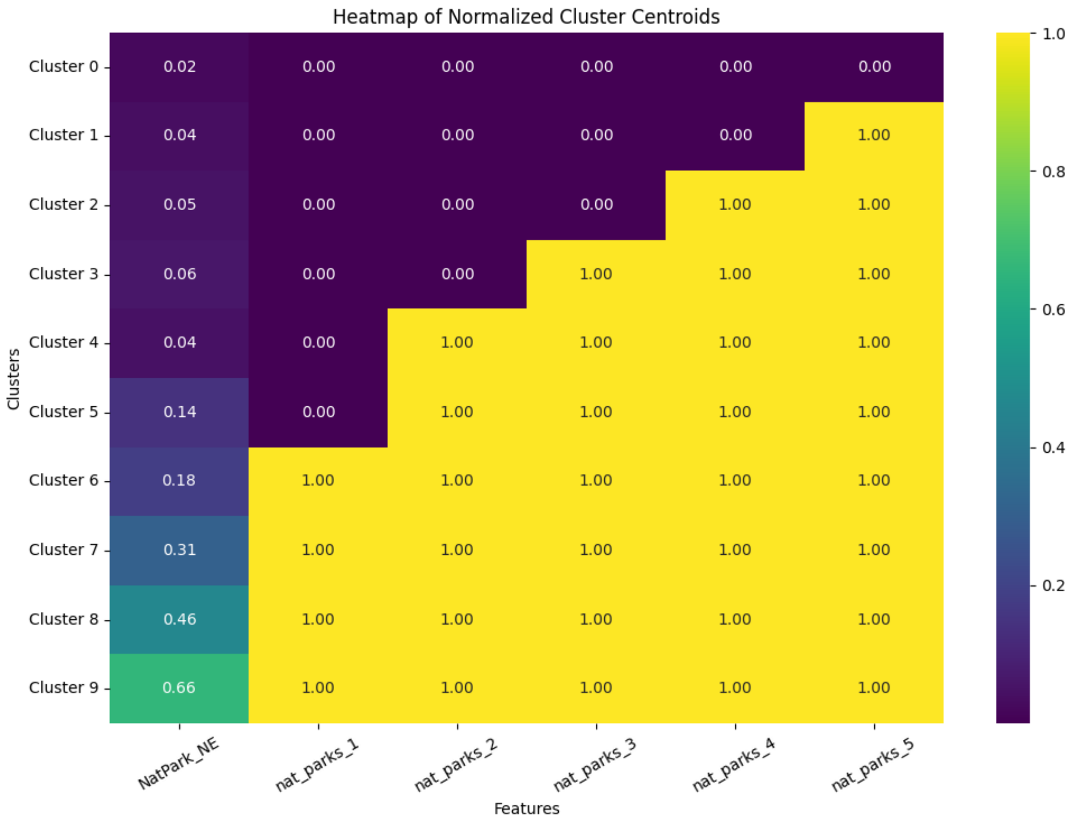

Figure 4: Heatmap of national park clusters, where each row represents the distribution from 0 to 1 for each type of value within the vector, and each column indicates the corresponding parameter.

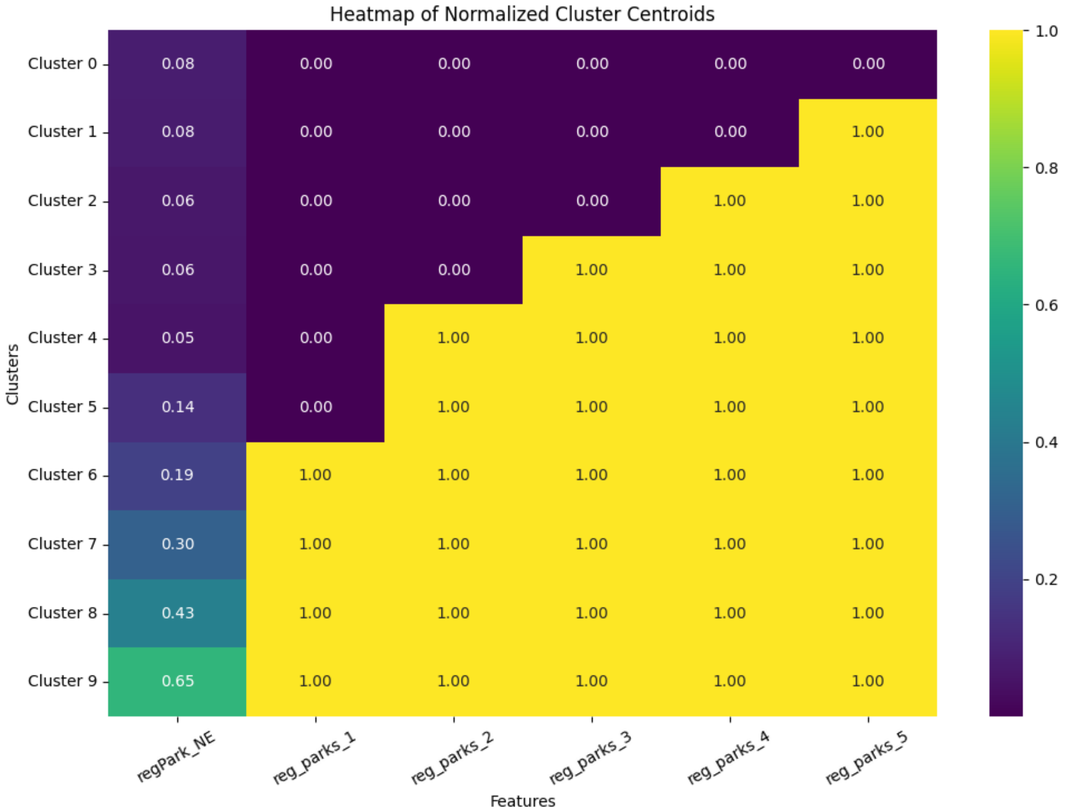

Figure 5: Heatmap of regional park clusters, where each row represents the distribution from 0 to 1 for each type of value within the vector, and each column indicates the corresponding parameter.

As seen in Figure 6, there is a prevalence of cluster descriptors indicating a lower number of local parks, with greater diversity observed in the national and regional parks. While access to basic amenities like local parks offers fundamental coverage, the establishment of national and regional parks throughout Ukraine is crucial for the sustained support of local communities.

The subsequent figures depict various aspects of the social life in villages and can be categorized into two groups: those that represent developed and easily accessible facilities (including churches as seen in Figure 8, education centers in Figure 9, elevators in Figure 10, hotels in Figure 11, kindergartens in Figure 12, medical facilities in Figure 14, and shops in Figure 15), and those that are more complex to access, such as banks and libraries (shown in Figures 7 and 13, respectively).

Addressing the first group, it is encouraging to observe that villages have prompt access to education, medical services, and kindergartens, as these are crucial for residents' well-being and community performance. While hotels and shops may seem less critical in terms of immediate resource accessibility, their presence is essential for fostering small businesses and providing access to goods. Furthermore, hotels enhance a village's capacity to welcome tourists, which is instrumental for sustainable development.

Regarding the second group, libraries are transforming with the widespread availability of the internet and mobile connectivity, acquiring new information has become predominantly digital, emphasizing the need to focus on the quality of these digital services. Nevertheless, libraries also serve as hubs for socialization, an aspect that could be supplemented by improving other communal spaces, such as local parks.

The distribution of banks remains a significant issue for ensuring convenient access to financial services. Although Ukraine boasts a well-developed e-banking system, physical banking access is still vital, especially for the elderly and tourists who may prefer direct interaction or require assistance. Thus, improving the physical banking infrastructure could render regions more attractive and comfortable, addressing any financial concerns that might arise.

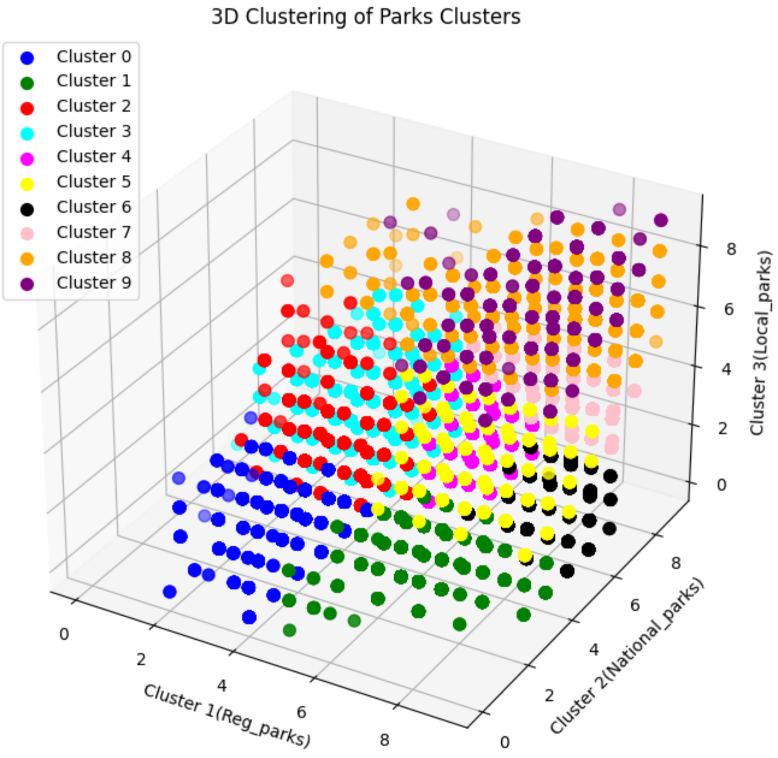

Figure 6: Figure 6: 3D plot of park clusters, with each color representing a different cluster. The order of the clusters can be seen on the top-left side, where 0 indicates the cluster closest to the objects, and 9 is the farthest from all other parks.

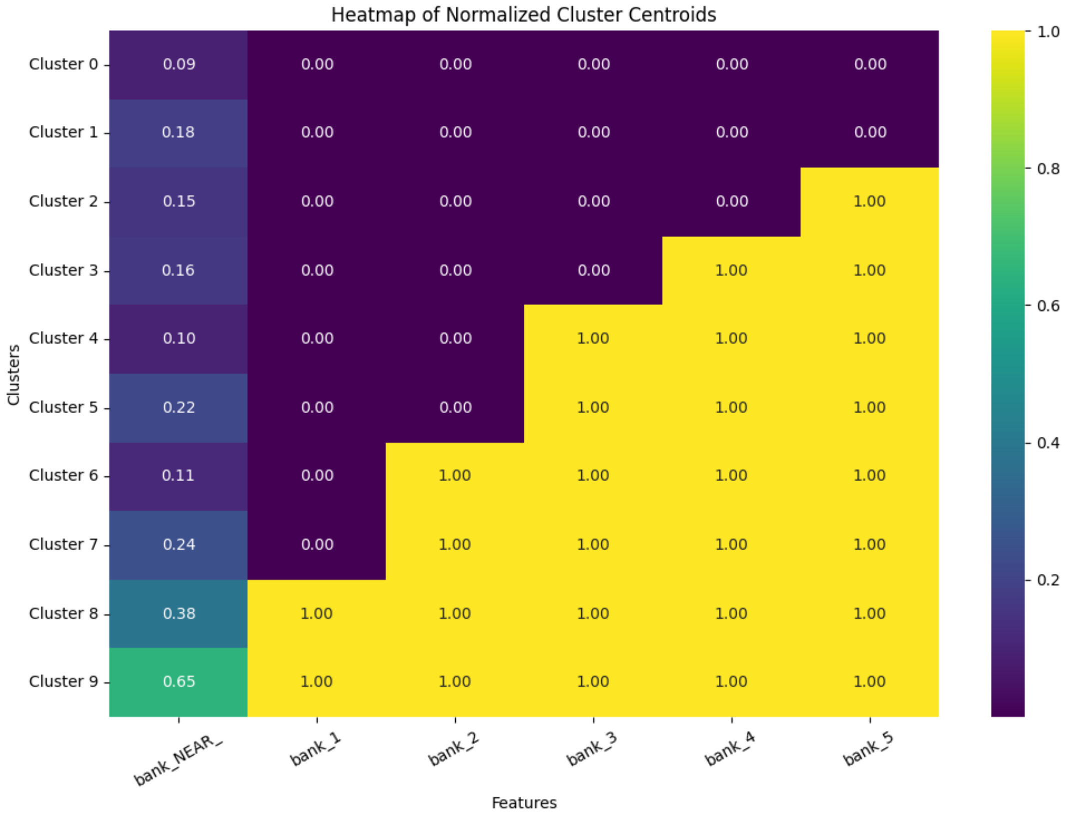

Figure 7: Heatmap of bank clusters, where each row represents the distribution from 0 to 1 for each type of value within the vector, and each column indicates the corresponding parameter.

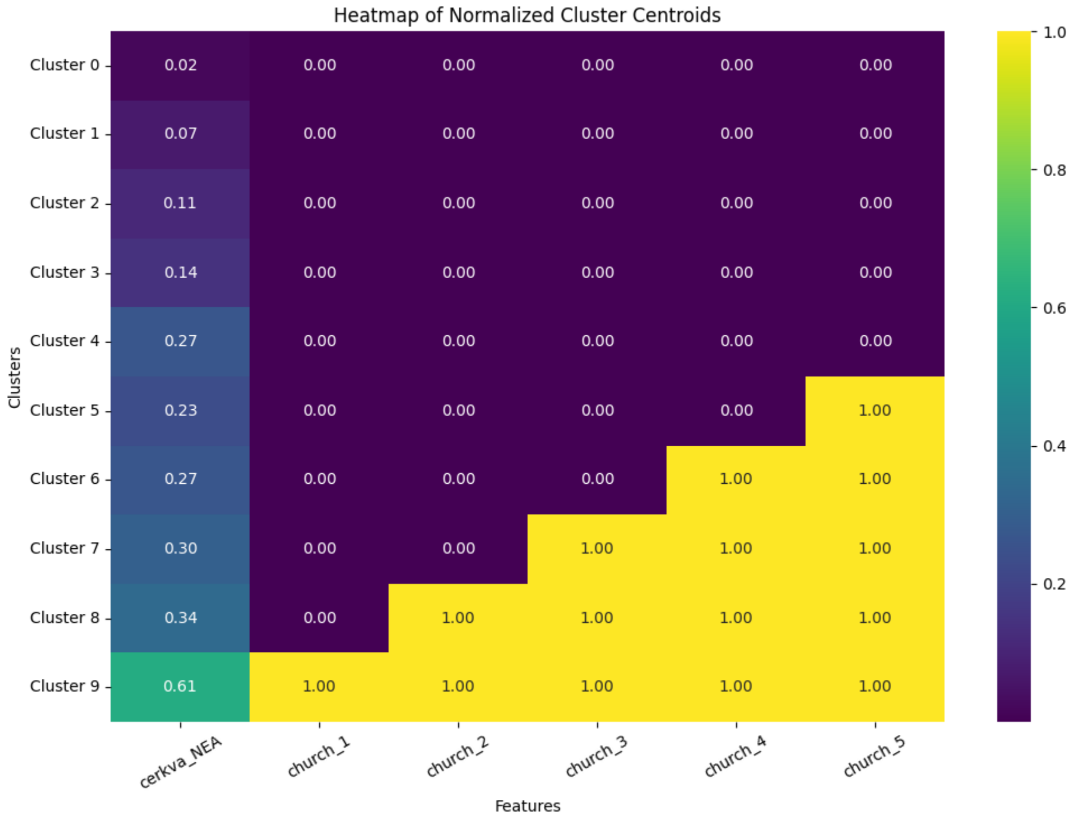

Figure 8: Heatmap of church clusters, where each row represents the distribution from 0 to 1 for each type of value within the vector, and each column indicates the corresponding parameter.

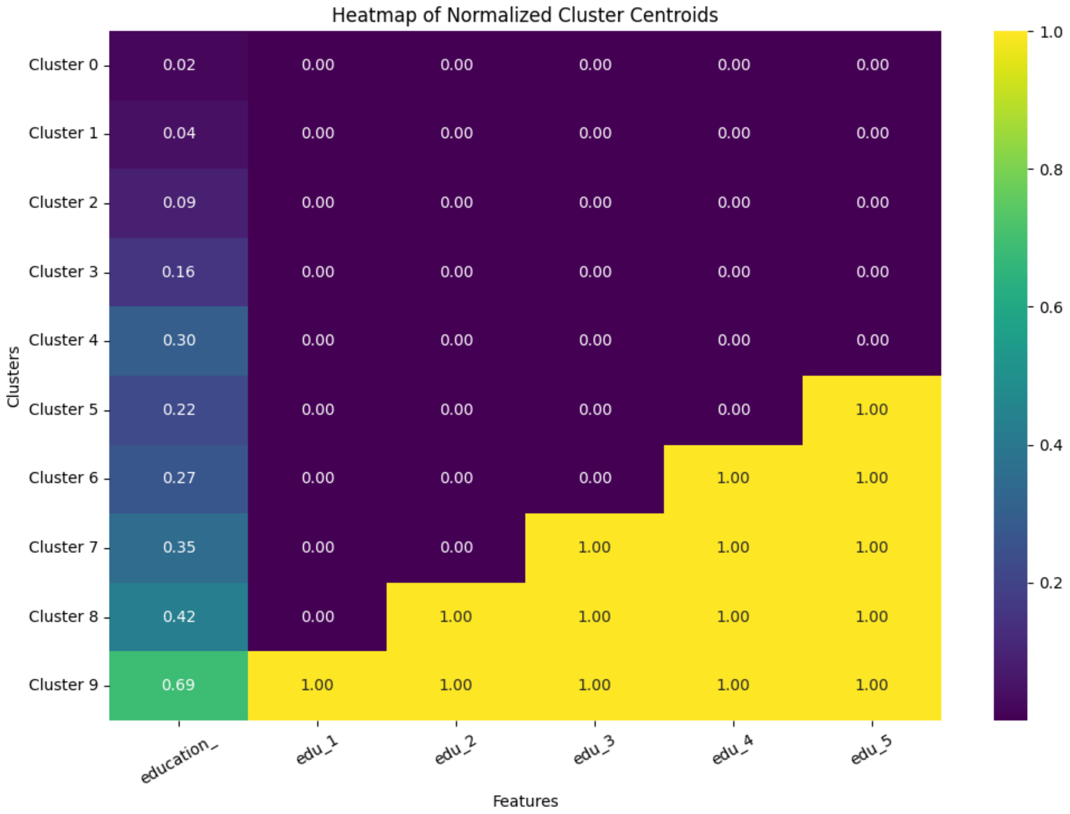

Figure 9: Heatmap of education clusters, where each row represents the distribution from 0 to 1 for each type of value within the vector, and each column indicates the corresponding parameter.

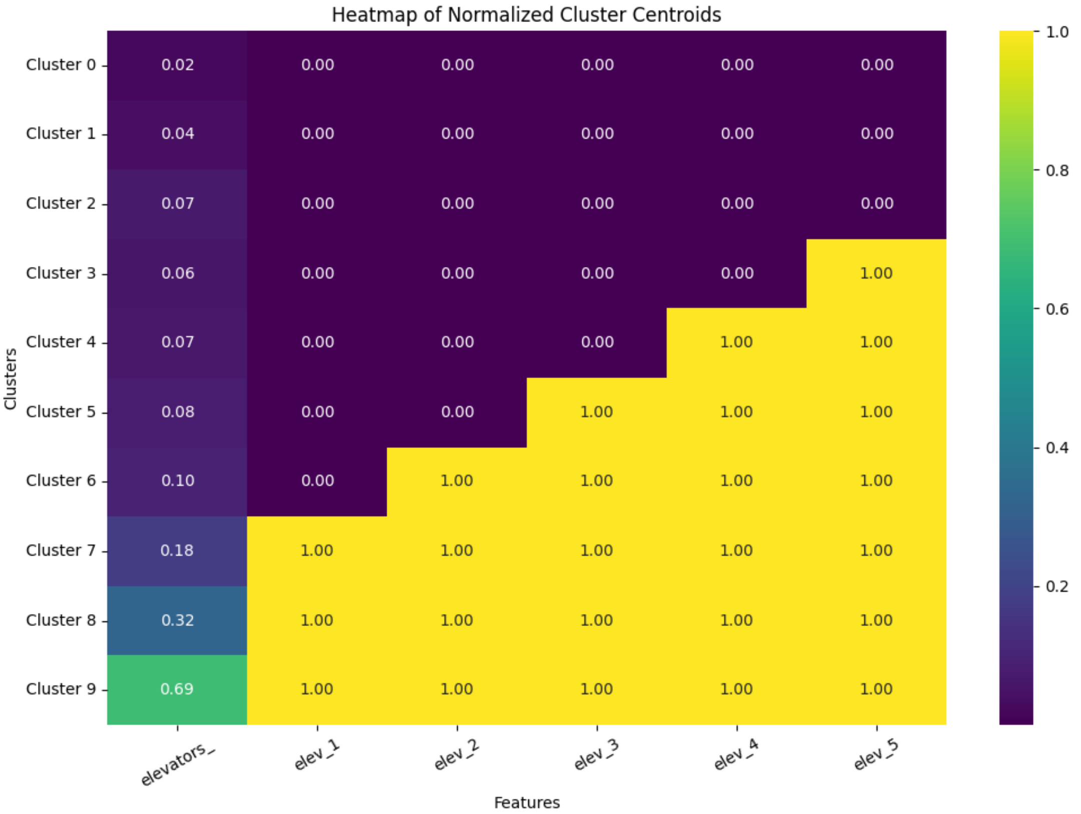

Figure 10: Heatmap of elevator clusters, where each row represents the distribution from 0 to 1 for each type of value within the vector, and each column indicates the corresponding parameter.

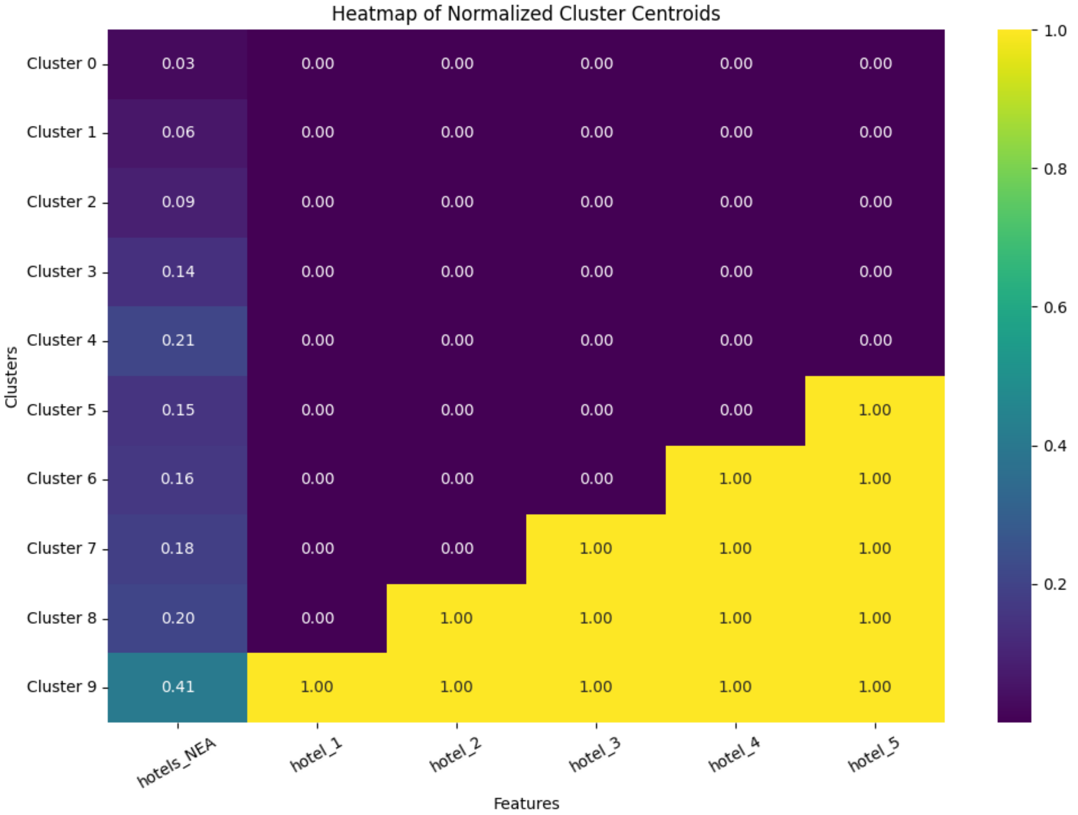

Figure 11: Heatmap of hotel clusters, where each row represents the distribution from 0 to 1 for each type of value within the vector, and each column indicates the corresponding parameter.

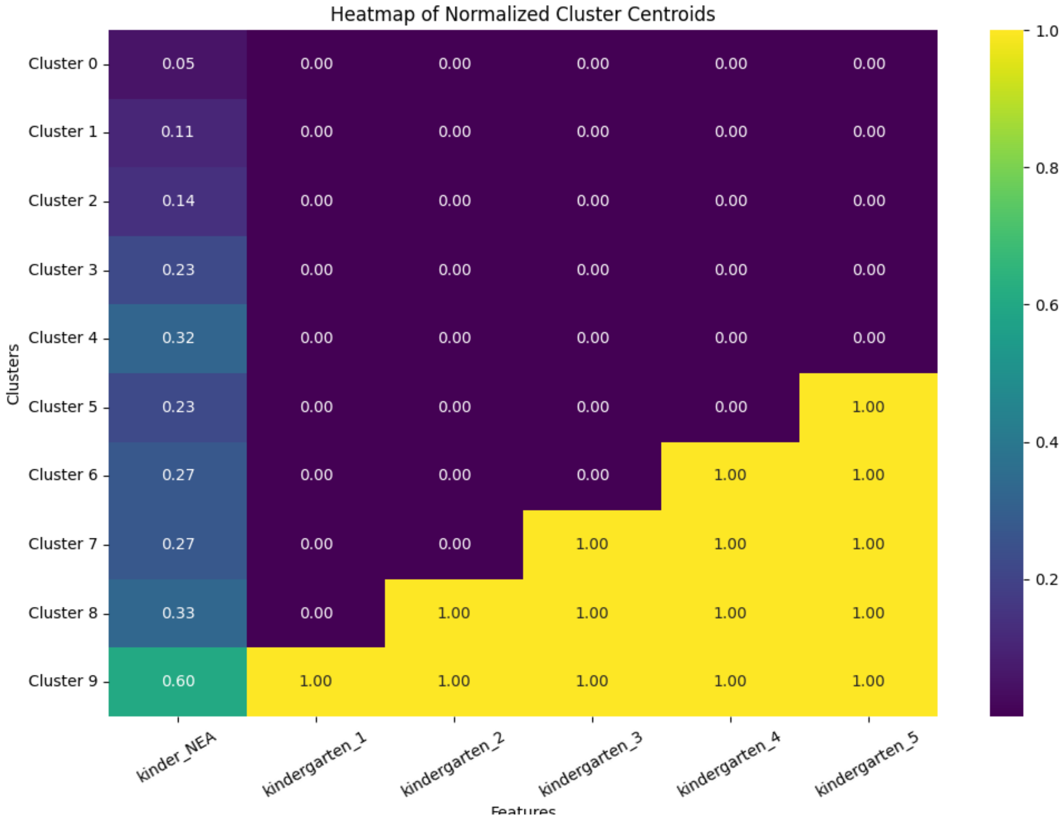

Figure 12: Heatmap of kindergarten clusters, where each row represents the distribution from 0 to 1 for each type of value within the vector, and each column indicates the corresponding parameter.

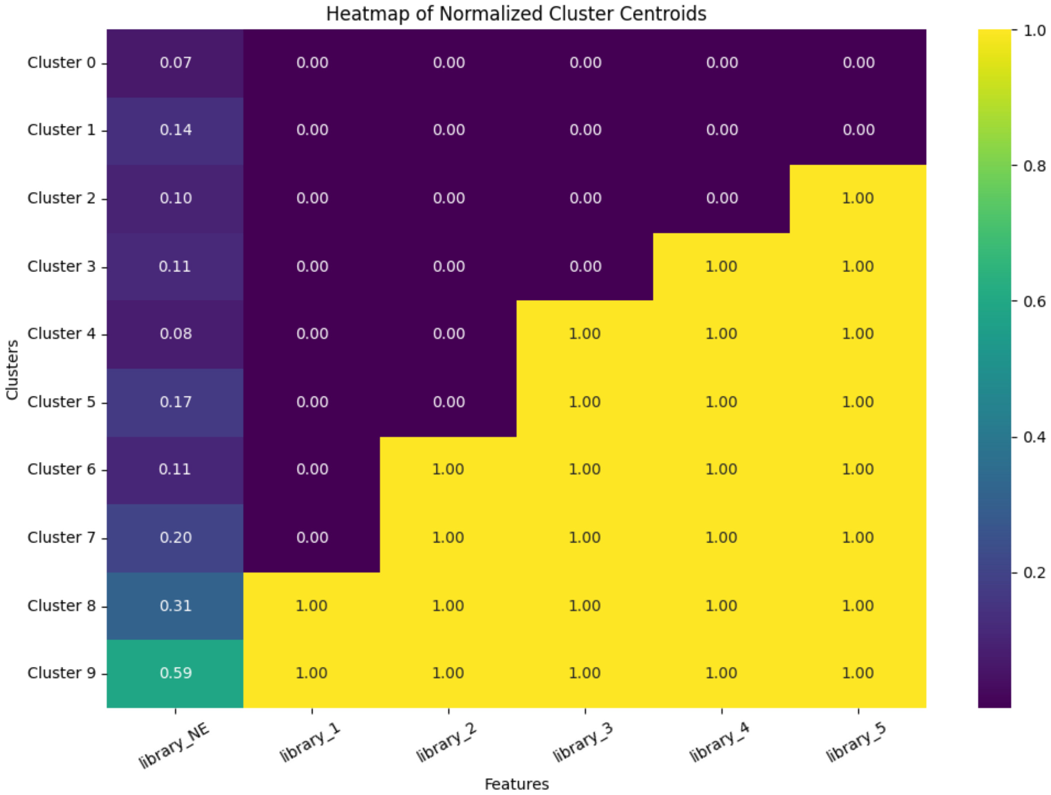

Figure 13: Heatmap of libraries clusters, where each row represents the distribution from 0 to 1 for each type of value within the vector, and each column indicates the corresponding parameter.

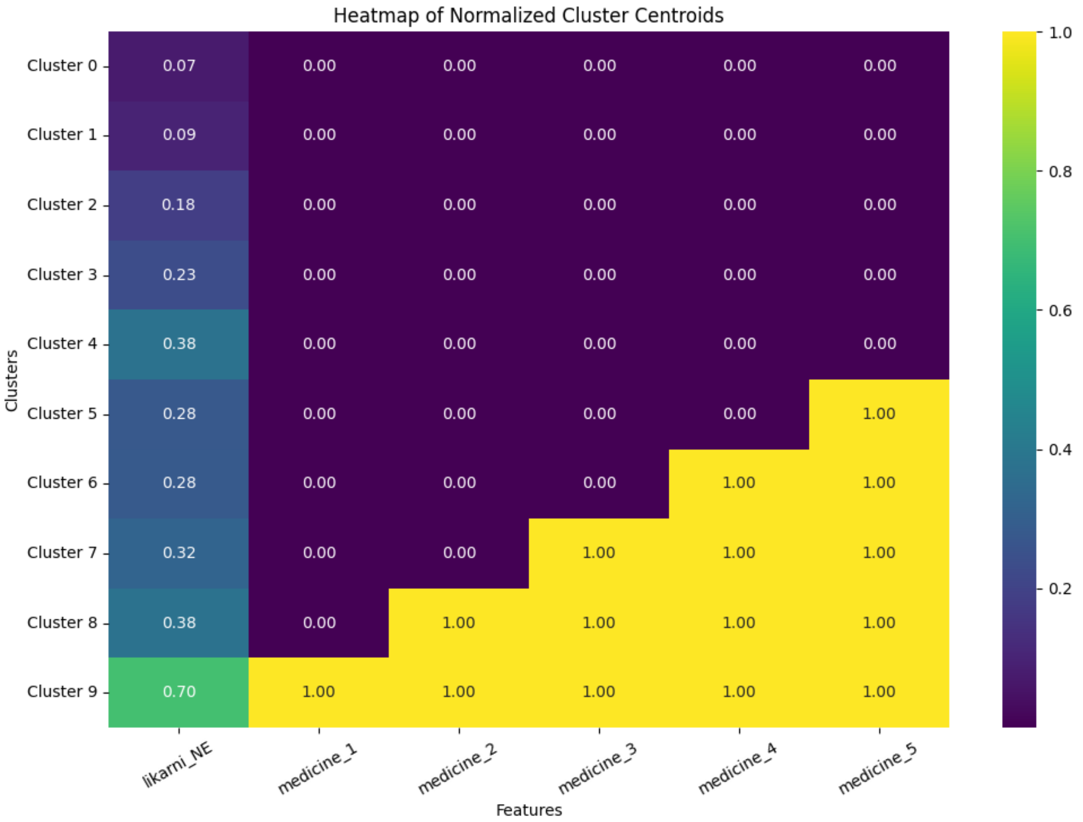

Figure 14: Heatmap of medicine clusters, where each row represents the distribution from 0 to 1 for each type of value within the vector, and each column indicates the corresponding parameter.

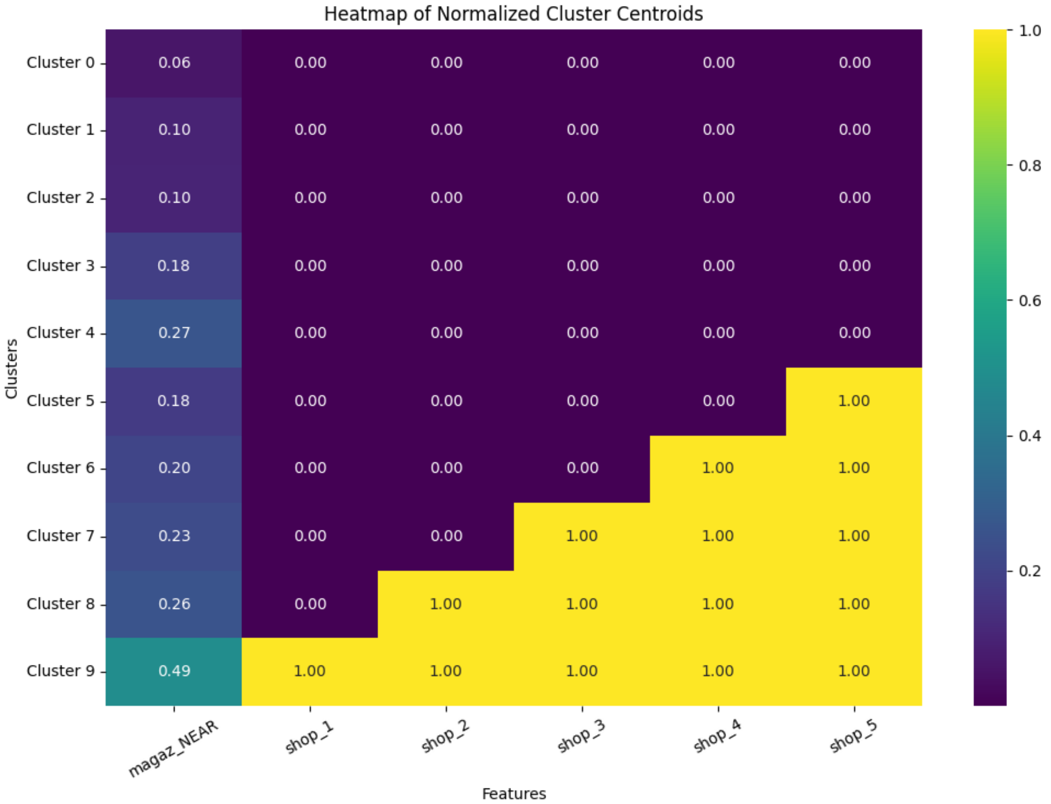

Figure 15: Heatmap of shop clusters, where each row represents the distribution from 0 to 1 for each type of value within the vector, and each column indicates the corresponding parameter.

When examining the overall cluster results, we can discern the distribution of these clusters. Although there are some discrepancies, such as the library descriptions in cluster 2 and road descriptions in cluster 1, Figure 17 reveals a distribution somewhat akin to a standard distribution with a mean of 0.3. This confirms that our grading method and approach have provided a correct and descriptive way to analyze our data.

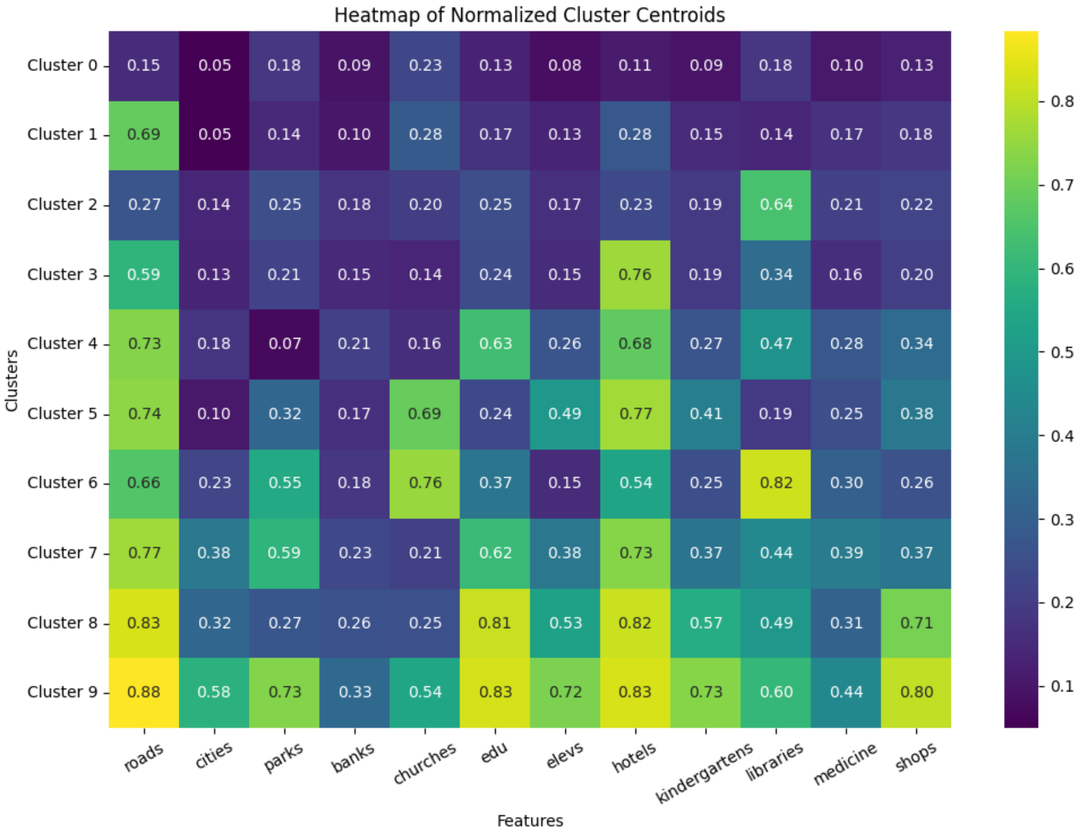

Figure 16: Heatmap of overall clusters, where each row represents the distribution from 0 to 1 for each type of value within the vector, and each column indicates the corresponding parameter.

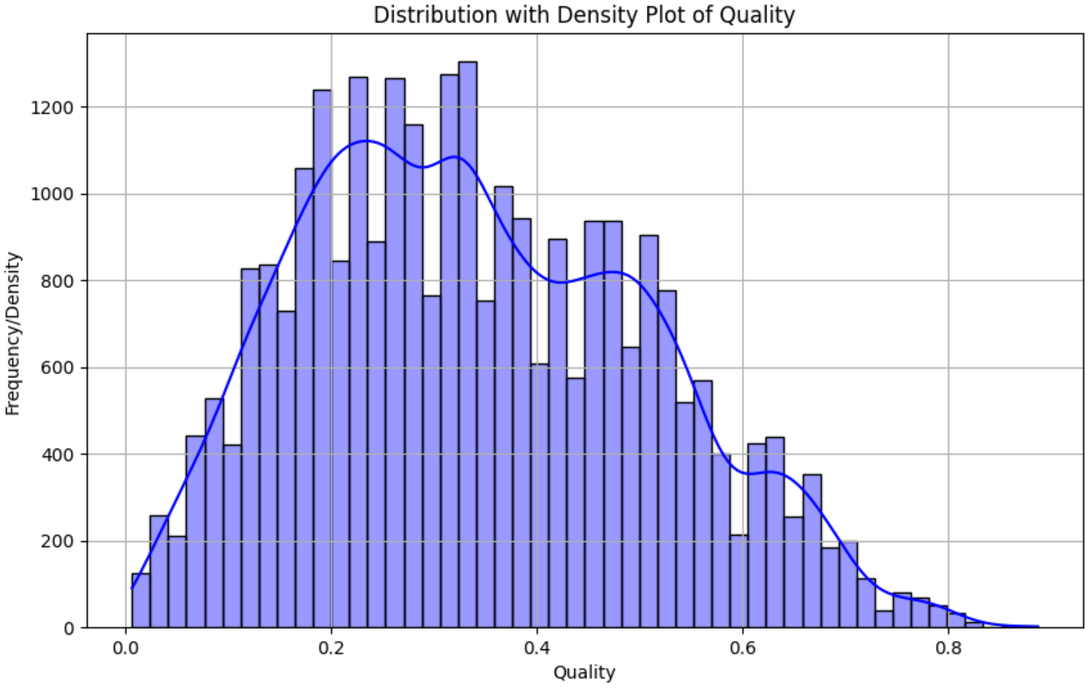

Figure 17: Histogram plot of the overall quality of all villages.

6. Discussion

This research has presented a multifaceted approach to analyzing and improving the infrastructure of villages through advanced clustering techniques and the integration of OpenStreetMap geospatial data. The nuanced methodology has successfully categorized villages based on the quality of infrastructure, highlighting disparities, and indicating potential developmental strategies.

The complexity of the method lies in its layered approach to data analysis. By constructing a comprehensive infrastructure profile for each village, our study moves beyond traditional clustering methods. The novel framework of custom vector creation, incorporating both proximity measures and detailed graph-based representations, has allowed for a more precise grouping of villages. This precision is crucial in a country like Ukraine, where rural infrastructure development is uneven and can greatly benefit from such targeted analysis.

Our findings reveal that the clustering process, augmented by our custom reorganization step, does indeed correspond with the actual infrastructure scenarios on the ground. The effectiveness of this approach is evident in the distinct clustering of villages, as depicted in the heat maps and 3D plots. These visual representations have been instrumental in understanding the clustering outcomes and have highlighted the variability in infrastructure access across Ukraine.

The distribution analysis, as evidenced by the histogram of the overall quality, confirms the validity of our grading method. The standard-like distribution with a mean quality score of 0.3 underscores the method's capability to provide a descriptive and accurate assessment of infrastructure quality. This analysis has been particularly revealing in identifying clusters with fewer amenities, such as those highlighted for local parks and educational facilities.

The future enhancements of this study could be substantial. Envisioning a system that not only assesses but also prescribes steps for infrastructure improvement could have profound implications for policy and planning. The potential for such a system to operate autonomously and provide real-time updates would revolutionize infrastructure development strategies. Furthermore, the addition of new descriptors to the graph structure and the ability to cluster graphs of similar sizes with descriptive node information would extend the model's applicability and precision.

Another aspect of future work involves addressing the method's scalability and adaptability to other geographic contexts. The model's performance in different environments and its ability to handle diverse data types are areas ripe for exploration. Moreover, integrating socioeconomic data could provide a more holistic view of the villages' needs, further refining development strategies.

Incorporating the insights from Figure 18 into our discussion, we observe a distinct spatial pattern in the distribution of infrastructure quality across Ukraine. The map vividly illustrates that villages in the western part of Ukraine appear more developed in terms of infrastructure access, while the southern and eastern regions exhibit signs of infrastructural deprivation. Central areas display a moderate level of development, suggesting a more balanced access to infrastructure. This spatial differentiation is not only crucial for immediate strategic planning but also for the design of long-term, sustainable development policies. Future research will aim to correlate these geographical patterns with socio-economic parameters such as population density, economic health, regional investment, and other critical metrics. By integrating additional vital descriptors and testing against these parameters, we anticipate a richer, more multifaceted understanding of rural development challenges. This research has laid the groundwork by providing a comprehensive method for assessing village infrastructure quality, proven effective in the Ukrainian context. The approach demonstrates the potential of geospatial analysis and machine learning to transform data into actionable insights, paving the way for policies that are not just reactive but proactively shaped by predictive analytics and a deep understanding of regional development dynamics.

In summary, the methodology presented here substantiates the transformative capabilities of machine learning and spatial data analysis within the domain of infrastructure development, particularly for rural upliftment. While the approach is intricate, it yields transparent, actionable insights with the potential to guide the efficient distribution of resources, an advantage that cannot be overstated in periods of economic hardship. Looking ahead, the evolution and application of this methodology are poised to facilitate more equitable and well-informed development strategies, underpinning the next steps towards rural prosperity and contributing to the broader narrative of sustainable development in the global context.

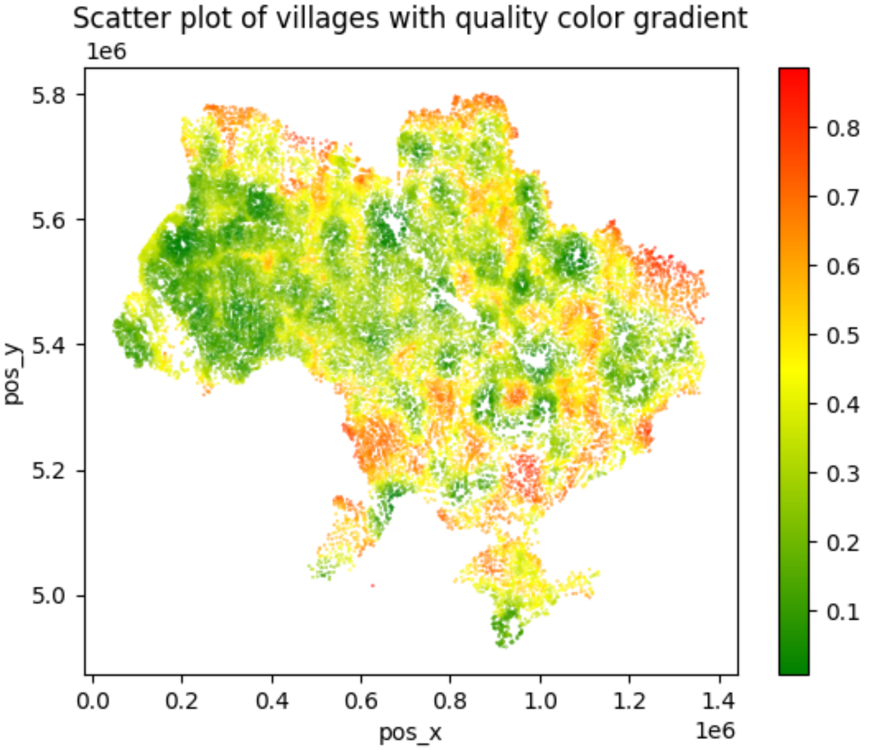

Figure 18: Scatter plot of villages with a quality color gradient. Each point on the grid represents a village; the color indicates the quality as defined before, where green represents better quality and red represents poorer quality.

7. Conclusion

This study has successfully demonstrated a comprehensive framework for assessing and enhancing the infrastructure of villages through the strategic use of clustering techniques and the exploitation of OpenStreetMap geospatial data. The developed methodology categorizes villages based on infrastructure quality, thus facilitating targeted development and setting a benchmark against average quality standards. In a context as diverse and challenging as Ukraine's rural landscape, the approach has proven to be exceptionally effective in pinpointing areas with infrastructure deficits and projecting potential enhancements.

The robustness of our method is evident from the multi-dimensional data analysis, integrating proximity measures with graph-based spatial information to yield a granular view of village infrastructure. The application of this method has provided a detailed infrastructure profile for each village, an essential step for any nuanced analysis and grouping based on infrastructure characteristics.

Our clustering process, reinforced by a post-clustering reorganization step, has not only simplified the interpretation of infrastructure statuses across clusters but also highlighted the systematic variation in infrastructure quality. The graphical heatmaps and 3D visualizations have offered a clear and immediate understanding of our clustering outcomes, illustrating the diversity and discrepancies in infrastructure access throughout Ukraine.

Furthermore, the distribution analysis and the subsequent quality score calculation have served as a validation of our methodology. The resultant grading system, underscored by a mean quality score, underscores the descriptiveness and accuracy of our approach, while also revealing key insights into the areas most in need of development.

The implications of this research are far-reaching. It provides a viable, cost-effective pathway to improve the overall infrastructure situation in Ukraine, especially in rural regions that have long been underserved. In the face of current economic challenges, the method's potential to drive efficient resource allocation cannot be overstated. The continued refinement and application of this methodology bear the promise of fostering more equitable and informed development practices, significantly enhancing the quality of life for rural populations.

As we look ahead, the adaptability of this methodology to other geographic contexts and the integration of socioeconomic data stand as exciting avenues for future exploration. This project has not only contributed a novel approach to infrastructure analysis but has also laid the groundwork for future innovations that can sustain and advance the development of rural infrastructure.

8. Acknowledgments

The study has been supported by the Ministry of Education and Science of Ukraine through the project “Information Technologies of Geospatial Analysis for the Development of Rural Areas and Communities" state registration number: 0123U102838. The authors are grateful to the project for providing the raw data for this analysis.

9. References

[1] The Long, H., Ma, L., Zhang, Y., & Qu, L. (2022). Multifunctional rural development in China: Pattern, process and mechanism. Habitat International, 102530. https://doi.org/10.1016/j.habitatint.2022.102530.

[2] Geza, W., Ngidi, M. S. C., Slotow, R., & Mabhaudhi, T. (2022). The Dynamics of Youth Employment and Empowerment in Agriculture and Rural Development in South Africa: A Scoping Review. Sustainability, 14(9), 5041. https://doi.org/10.3390/su14095041.

[3] Arintoko, A., Ahmad, A.A., Gunawan, D.S., & Supadi, S. (2020). Community-based tourism village development strategies: A case of Borobudur tourism village area, Indonesia. GeoJournal of Tourism and Geosites, 29(2), 398–413. https://doi.org/10.30892/gtg.29202-477.

[4] Aziiza, A. A., & Susanto, T. D. (2020). The Smart Village Model for Rural Area (Case Study: Banyuwangi Regency). IOP Conference Series: Materials Science and Engineering, 722, 012011. https://doi.org/10.1088/1757-899X/722/1/012011.

[5] Stojanova, S., Lentini, G., Niederer, P., Egger, T., Cvar, N., Kos, A., & Stojmenova Duh, E. (2021). Smart Villages Policies: Past, Present and Future. Sustainability, 13(4), 1663. https://doi.org/10.3390/su13041663.

[6] H. Yailymova, B. Yailymov, N. Kussul, A. Shelestov, "Geospatial Analysis of Life Quality in Ukrainian Rural Areas," in Proceedings of the 13th International Conference on Dependable Systems, Services and Technologies (DESSERT), Athens, Greece, 2023, pp. 1-5. https://doi.org/10.1109/DESSERT61349.2023.10416517.

[7] Golalipour, K., Akbari, E., Hamidi, S.S., Lee, M., & Enayatifar, R. (2021). From clustering to clustering ensemble selection: A review. Engineering Applications of Artificial Intelligence, 104, 104388. https://doi.org/10.1016/j.engappai.2021.104388.

[8] Vendoti, S., Muralidhar, M., & Kiranmayi, R. (2021). Techno-economic analysis of off-grid solar/wind/biogas/biomass/fuel cell/battery system for electrification in a cluster of villages by HOMER software. Environmental Development and Sustainability, 23, 351–372. https://doi.org/10.1007/s10668-019-00583-2.

[9] Nugraha, L.F., Sulistyowati, L., Setiawan, I., & Noor, T.I. (2022). Alternative Community-Based Village Development Strategies in Indonesia: Using Multicriteria Decision Analysis. Agriculture, 12, 1903. https://doi.org/10.3390/agriculture12111903.

[10] Supandi, A., Saefuddin, A., & Sulvianti, I. D. (2020). Two step Cluster Application to Classify Villages in Kabupaten Madiun Based on Village Potential Data. Xplore, 10(1). https://doi.org/10.29244/xplore.v10i1.272.

[11] Cvar, N., Trilar, J., Kos, A., Volk, M., & Stojmenova Duh, E. (2020). The Use of IoT Technology in Smart Cities and Smart Villages: Similarities, Differences, and Future Prospects. Sensors, 20(14), 3897. https://doi.org/10.3390/s20143897.

[12] Liu, Y., et al. (2023). Simple Contrastive Graph Clustering. IEEE Transactions on Neural Networks and Learning Systems. https://doi.org/10.1109/TNNLS.2023.3271871.

[13] Kapoor, N., Ahmad, N., Nayak, S.K., Singh, S.P., Ilavarasan, P.V., & Ramamoorthy, P. (2021). Identifying infrastructural gap areas for smart and sustainable tribal village development: A data science approach from India. Journal of Jimei University International Edition, https://doi.org/10.1016/j.jjimei.2021.100041.

[14] Y. Liu, X. Ke, W. Wu, M. Zhang, X. Fu, J. Li, J. Jiang, Y. He, C. Zhou, W. Li, Y. Li, Y. Song, X. Zhou, Geospatial characterization of rural settlements and potential targets for revitalization by geoinformation technology, Sci. Rep. 12 (2022) 8399. https://doi.org/10.1038/s41598-022-12294-2.

[15] B. Herfort, S. Lautenbach, J. Porto de Albuquerque, J. Anderson, A. Zipf, A spatio-temporal analysis investigating completeness and inequalities of global urban building data in OpenStreetMap, Nat. Commun. 14 (2023) 3985. https://doi.org/10.1038/s41467-023-33956-z.

[16] Ye, X., & Wei, Y. D. (2013). Geospatial Analysis of Regional Development in China: The Case of Zhejiang Province and the Wenzhou Model. Pages 445-464. https://doi.org/10.2747/1538-7216.46.6.445.

[17] Chanak, P., & Banerjee, I. (2021). Internet-of-Things-Enabled Smart Villages: An Overview. IEEE Consumer Electronics Magazine, 10(3), 12-18. https://doi.org/10.1109/MCE.2020.3013244.

[18] Cai, C., Zaghloul, M., & Li, B. (2022). Data Clustering in Urban Computational Modeling by Integrated Geometry and Imagery Features for Probabilistic Navigation. Applied Sciences, 12(24), 12704. https://doi.org/10.3390/app122412704.

[19] Adamowicz, M., & Zwolińska-Ligaj, M. (2020). The “Smart Village” as a Way to Achieve Sustainable Development in Rural Areas of Poland. Sustainability, 12(16), 6503. https://doi.org/10.3390/su12166503.

[20] Wang, Q., Luo, S., Zhang, J., & Furuya, K. (2022). Increased Attention to Smart Development in Rural Areas: A Scientometric Analysis of Smart Village Research. Land, 11(8), 1362. https://doi.org/10.3390/land11081362.

[21] Zavratnik, V., Kos, A., & Stojmenova Duh, E. (2018). Smart Villages: Comprehensive Review of Initiatives and Practices. Sustainability, 10(7), 2559. https://doi.org/10.3390/su10072559.

[22] Zavratnik, V., Podjed, D., Trilar, J., Hlebec, N., Kos, A., & Stojmenova Duh, E. (2020). Sustainable and Community-Centred Development of Smart Cities and Villages. Sustainability, 12(10), 3961. https://doi.org/10.3390/su12103961.

[23] Visvizi, A., & Lytras, M.D. (2020). Sustainable Smart Cities and Smart Villages Research: Rethinking Security, Safety, Well-being, and Happiness. Sustainability, 12(1), 215. https://doi.org/10.3390/su12010215.

[24] Abdulrazzak, H.N., Hock, G.C., Mohamed Radzi, N.A., Tan, N.M.L., & Kwong, C.F. (2022). Modeling and Analysis of New Hybrid Clustering Technique for Vehicular Ad Hoc Network. Mathematics, 10(24), 4720. https://doi.org/10.3390/math10244720.

[25] Suprayoga, G.B., Bakker, M., Witte, P., et al. A systematic review of indicators to assess the sustainability of road infrastructure projects. Eur. Transp. Res. Rev. 12, 19 (2020). https://doi.org/10.1186/s12544-020-0400-6