Proceedings of the 2024 Winter Simulation Conference
H. Lam, E. Azar, D. Batur, S. Gao, W. Xie, S. R. Hunter, and M. D. Rossetti, eds.

UNCOVERING SOCIOECONOMIC FEATURES IN PAVEMENT CONDITIONS THROUGH DATA MINING: A TWO-STEP CLUSTERING MODEL

Tamim Adnan[1], and Abdolmajid Erfani[1]

[1]Dept. of Civil and Environmental Engineering, Michigan Technological University, Houghton, MI, USA

ABSTRACT

Across the United States, individuals regardless of their socioeconomic backgrounds deserve equitable access to high-quality roads and highways. This research delves into the use of data mining methods to examine access quality, focusing on pavement condition through the International Roughness Index and socioeconomic factors, by exploring the Highway Performance Monitoring System (HPMS) dataset. Data mining serves as an exploratory process, unveiling and visualizing valuable yet not immediately evident insights within extensive datasets. Through data mining with two-step clusters, k-means, and hierarchical agglomerative clustering, we examined over 8 million records from HPMS and U.S. census data over four years. Our findings highlight the impact of socioeconomic elements—such as urbanization, income, and demographic composition—on pavement quality, beyond traffic, weather, and technical specifications. These insights emphasize the need for incorporating social equity into pavement maintenance and budgeting strategies, underscoring the significant role socioeconomic factors play in pavement performance.

1 INTRODUCTION

Road infrastructure is pivotal in various dimensions of socio-economic advancement, enhancing communication networks, creating employment opportunities, mitigating poverty, and accelerating progress in education, healthcare, and many other sectors (Timilsina and Hochman 2020). Therefore, maintaining the physical and functional performance of road infrastructure within an acceptable range is crucial, necessitating robust operation and maintenance (O&M) efforts by the Federal Highway Administration (FHWA) and state highway agencies (SHAs) as part of their transportation asset management strategies (FHWA 2007).

The rapid deterioration of the transportation assets, alongside limited financial budgets, the impacts of climate change and natural disasters, as well as legislative requirements, has complicated decision-making processes for SHAs on allocating fundings on transportation assets across their region (Evan 2023; Kothari et al. 2022). Disproportionate allocation of resources for maintaining transportation assets can lead to unequal access to high-quality roads and networks for individuals from diverse backgrounds. Recent initiatives have acknowledged the significance of social equity in the development of transportation asset management plans (Khalife et al. 2023), with both the FHWA and SHAs aiming to integrate equity as a critical facet of asset management, alongside performance-based and environmental considerations. However, there remains a scarcity of research evaluating the state of equity concerning pavement assets nationwide. Crucially, pinpointing the existing disparities in access quality among various socioeconomic groups is vital for improved future planning and addressing these gaps.

Examining the influence of socioeconomic factors on pavement asset management presents a multifaceted challenge, as pavement condition is determined by a variety of elements. These include the technical design of the pavement, traffic levels, environmental conditions, variations in soil and subgrade,

as well as maintenance and rehabilitation efforts. Thus, employing advanced data mining techniques can be advantageous for uncovering hidden patterns in data while managing the multiple driving factors influencing pavement condition. In this study, we utilized a two-step clustering approach, specifically crafted to unearth meaningful insights from datasets that are complex, large, and diverse (Moscoso et al. 2024; Radovic et al. 2017). This method is useful to analyze the data from mixed fields and exclude the outliers from the raw dataset (IBM 2024). The complex and extensive nature of the Highway Performance Monitoring System (HPMS) data, used for monitoring pavement conditions across the nation, justifies the use of data mining techniques like clustering to uncover concealed socioeconomic patterns.

This study aims to uncover existing disparities in road access quality among various socioeconomic groups throughout the U.S. It explores patterns in reported pavement conditions, alongside traffic, climate, regional, and other environmental factors affecting pavement condition, using data from the HPMS database. This is coupled with socioeconomic data including total population, average household income, employment status, and racial demographics from different regions, as gathered from the census bureau. The objective is to utilize clustering techniques to group samples with similar patterns relating to technical aspects, environmental conditions, traffic, road types, and socioeconomic characteristics together. This approach aims to assess how socioeconomic attributes vary within clustered data, with other influencing factors being controlled. The subsequent sections of this paper are structured as follows: initially, a literature review focusing on recent research related to social equity in transportation asset management is presented. This is followed by the methodology section, which details the processes of data collection and processing, including the implementation of the two-step clustering method. The analysis and discussion of the results are then provided. The paper concludes with a summary and the key findings of the study.

2 LITERATURE REVIEW

2.1 Social Equity in Transportation Asset Management

The US Department of Transportation, in its Equity Action Plan, articulates equity as "the steady and systematic fair, just, and impartial treatment of every individual, encompassing those from underserved communities historically deprived of such fairness". Accordingly, social equity covers fair distribution of resources, opportunities, benefits, and burdens across society (Seyedrezaei et al. 2023). Asset management is one of the major subsectors of transportation where equity is still marginally considered (Khalif et al. 2023). According to a few recent studies, the relationship between socioeconomic status and asset conditions indicates disparities.

For example, Gandy, Armanios, and Samaras (2023) conducted an analysis linking the National Bridge Inventory (NBI) condition ratings with US Census Bureau tract data. Their findings indicated that bridges situated in lower-income and disadvantaged community tracts are more prone to being in poor condition. Erfani et al. (2024) performed a statistical analysis to assess the relationship between social vulnerability characteristics and pavement condition statuses throughout the US. Ultimately, the research findings by Erfani et al. (2024), utilizing Spearman correlation and the Random Forest (RF) machine learning classification model, uncovered that American groups facing socio-demographic disadvantages—characterized by higher populations and percentages of African Americans, as well as challenges related to transportation, housing, and language barriers—encounter disparities in accessing high-quality pavements. In another study (Townsend 2022), the YOLO5 model was used to develop an Artificial Intelligence based framework to look at pavement distresses and sociological variables such as racial, economic, educational, and occupational characteristics. A significant gap in pavement conditions was found between the advantaged and disadvantaged groups because of the household income disparity. Cavallaro et al (2020) applied social spatial equity to track the profitable groups from the advancement in High-Speed Rail in the northwestern part of Italy. The finding showed that high income people are getting advantages of the advancements. Here are several examples of research investigations that examine the current

socioeconomic status of populations alongside asset conditions and infrastructure development, employing both statistical and machine learning methodologies (Erfani and Frias-Martinez 2023).

In addition to research studies, the US Government Accountability Office (GAO) report published in 2022 raised important issues for SHAs and FHWA to consider when allocating resources for transportation asset maintenance. This research focuses solely on analyzing 2019 HPMS data using spatial linear regression models and discovers that most US highway pavements—spanning approximately 220,000 miles—are in either good or fair condition. However, it also identifies that highway pavements are less likely to be in good condition in urban regions, areas with higher family poverty rates, and localities with a greater percentage of underserved racial and ethnic groups. There have been a few studies that have examined how equity can be incorporated into asset management policies and practices in the transportation sector. Kothari et al. (2022) proposed a highway maintenance and rehabilitation (M&R) model including three sustainable components, namely environment, economic, and social equity, to reduce greenhouse gas emissions. Poor pavement conditions are susceptible to more greenhouse gas emissions by vehicles due to their higher fuel consumption. Therefore, Kothari et al. (2022) optimized the proposed M&R model with the NSGA II optimization algorithm for a sustainable asset management plan, including environmental, economic, and social equity.

Moreover, there has been an ongoing body of research within transportation and construction sector to investigate social equity in various topics. Studies mostly focus on achieving two main goals: enhancing overall access to services and facilities and ensuring a more equitable distribution of this access across different regions (Ortega, López and Monzón 2012; Welch and Mishra 2013; Cavallaro et al. 2020; Seyedrezaei et al. 2023). Several studies have established that the assessment of social equity is influenced by factors such as income, population size, race or ethnicity, education levels, urbanization, and employment. There are some metrics available to measure social equity such as range, variance, measure of variations, log variance, Theli's entropy measure, and Gini index (Ramjerdi 2006). Although some studies have explored pavement asset management, statistical and supervised machine learning models often struggle to analyze complex databases where pavement condition is influenced by a wide array of factors. These include technical design, traffic flow, environmental conditions, and socioeconomic factors. As a result, clustering methods, especially those employing multi-step advanced models, are becoming increasingly popular for more precisely uncovering the intricate patterns present in large datasets.

2.2 Two Step Cluster Analysis

Two-step cluster analysis is emerging as a favored data mining technique for pattern detection in sizable datasets that incorporate various parameters. Several studies have utilized this approach for diverse applications. Common unsupervised machine learning methods like K-means cluster analysis and hierarchical clustering, including agglomerative cluster analysis, are widely used to categorize data with similar traits. Radovic et al. (2017) implemented two-step cluster analysis on NBI bridge data, where initial pre-clusters formed by K-means were further aggregated through hierarchical agglomerative clustering, showcasing the effectiveness of this strategy in analyzing complex data. Two-step cluster analysis serves as an effective tool for knowledge discovery, capable of processing both categorical and interval data concurrently while reducing dimensions in vast datasets. This algorithm is proficient, creating clusters from either categorical or continuous data and adapting to a variable number of clusters (Moscoso et al. 2024; Radovic et al. 2017).

In the conventional single clustering method, such as when using k-means to define a predetermined number of clusters within a large dataset, each data point is allocated to the nearest cluster based on centroid distance. Subsequently, the k-means algorithm recalculates the centroids of these clusters by averaging the distances between the cluster centroids and all associated data points (Abdulhafedh, 2021). In usual single clustering approach When the k-means clusters are assigned to create a specific number of clusters for a large dataset, it gives all data points to their closest clusters by centroid distance. Then, the k-means model recomputes the centroids of the clusters by calculating the average distance between the centroid of the

clusters and all the data points (Abdulhafedh, 2021). The mathematical equation of k-means clustering is given below:

$$J = \sum_{i=1}^{m} \sum_{i=1}^{k} w_{ik} ||x - \mu_i||)^2$$

where J is the objective function that the k-means algorithm minimizes. m is the total number of data points, k is the total number of clusters, w_{ik} = 1 if the datapoints are x_i, belongs to cluster k, otherwise, w_{ik} = 0, x_i is the i-th data point, μ_k is the centroid of the cluster k, $||x_i - \mu_k||^2$ is the squared Euclidean distance between the data point x_i and μ_k (Dabbura 2018).

On the other hand, hierarchical clusters measure the hierarchical order among clusters and group them accordingly (Abdulhafedh, 2021). Given the fact that, hierarchical clusters are computationally expensive for large datasets while k-means are inexpensive and only efficient for the data which has interval data points (Radovic et al. 2017). The mathematical equation for agglomerative clustering involves calculating the mean distance between elements of each cluster. This process is fundamental in the hierarchical clustering method, where clusters are generated hierarchically by merging two clusters at a time (Murtagh and Contreras 2012).

$$d = \sqrt{\sum_{i=1}^{k} (x_i - y_i)^2}$$

2.3 Research Objective

This research aims to explore how socio-economic factors influence asset management and consequently pavement conditions across different regions in the United States. Utilizing a data mining approach called two-step cluster analysis, the study endeavors to reveal socio-economic trends affecting pavement quality, while also considering technical design, traffic, and environmental and climate factors. The primary objective is to analyze the prevailing pavement condition patterns for people of various races, income levels, and geographical areas, emphasizing the need for social equity in transportation asset management. By conducting a social equity analysis on pavement conditions and identifying social patterns, this study aims to make a significant contribution, advocating for equitable and high-standard pavement maintenance for all communities.

3 METHODOLOGY

The methodology section outlines the processes of data collection, preparation, and preprocessing. It details the development of a two-step clustering approach, enhancing the k-means and hierarchical agglomerative clustering models through fine tuning procedures. A visual summary of this methodology is provided in Figure 1.

3.1 Data Collection and Preparation

The HPMS database provides detailed insights into the use, maintenance, and condition of U.S. highways, offering a broad overview of national highway health. Data collected based on availability from 2017 to 2020, it amassed 8,487,369 records nationwide, covering pavement performance (indicated by the International Roughness Index, IRI) and road details like type, surface, base thickness, traffic, and location. The IRI is regarded as one of the premier metrics for measuring pavement performance. This is due to its uniformity across different regions and pavement types, its strong correlation with ride quality, broad acceptance within the industry, and its ability to predict future pavement conditions (Corley-Lay 2014). This study collected social data from the census bureau where the American Community Survey (ACS) provides annual estimates for income, education, employment, health insurance coverage, and housing

costs and conditions for residents of the United States. As per the literature review, social equity is associated with employment status, race, population, and household income. Therefore, this study collected one year's estimated ACS data over the years of collected HPMS data.

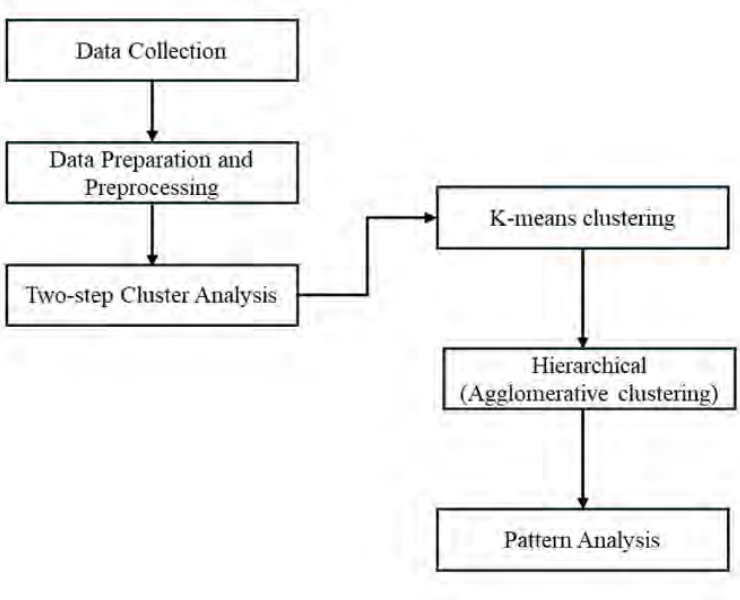

Figure 1: The methodology.

3.2 Data Preparation

A critical step in our research methodology involved the preparation of datasets obtained from the HPMS and social parameters for the years 2017, 2018, 2019, and 2020. Data preparation process comprised several essential steps, as outlined below, to ensure the data's integrity based on geographical information and usability for further analysis. Table 1 displays the chosen parameters, encompassing traffic, climate, and social characteristics.

To facilitate temporal analysis, we appended a 'Year' column to each dataset, which allowed us to effectively track and analyze temporal trends. We began handling missing values by replacing instances of the character 'N' with NaN (Not a Number) to standardize the representation of missing data. We then identified columns with missing values and calculated the mean for each, excluding NaN entries. This enabled us to impute missing values with their respective means, ensuring the dataset's statistical integrity. For feature selection and processing, we evaluated the data types in each column to confirm their suitability for numerical computations or modeling tasks, performing necessary type conversions. Additionally, we analyzed variable relationships through correlation and removed features exhibiting collinearity.

3.3 Two Step Cluster Analysis

The two-step clustering approach begins with an initial pass through the dataset, arranging the raw data into pre-clusters or subclusters. Following this, the second phase employs a hierarchical clustering technique to detect patterns within these pre-clusters, ultimately forming the final set of clusters (IBM 2024). Determining the optimal number of clusters for k-means in the first step presented a challenge. However, by utilizing the elbow method, it was suggested that the ideal number of clusters is 6. This was further supported by the Davies-Bouldin Score, which confirmed that range of 6 -13 clusters would be the most suitable choice. Despite differing recommendations for the optimal number of clusters from various metrics,

the study sought to identify the maximum number of groups within the dataset while ensuring satisfactory metric scores. Ultimately, the decision was to establish 10 clusters for KMeans, where both the silhouette and Davies-Bouldin scores were within acceptable limits (Table 2).

Table 1: Selected parameters and their descriptions.

Parameters	**Description**
Year	Data were collected in each calendar years (2017 – 2021)
Urban_code_mode	When the code 9999 its rural, and 9998 it's small urban and all other codes are urban (FHWA 2016)
AADT mean	Annual Average Daily Traffic
IRI	International Roughness Index
Relative Humidity	Annual Average Daily Average Relative Humidity in each county (FIPS)
Precipitation	Water equivalent of annual surface precipitation (measured in millimeters) in each county (FIPS)
Mean Household income	Mean Household income (dollars) in each county (FIPS)
Total Population	Total population in each county (FIPS)
Median Age	Median age of people in each county
Employment Rate on Population 16 years and over	Employment/Population Ratio on the Population 16 years and over.
Unemployment Rate on Population 16 Years and Over	Unemployment rate on Population 16 Years and Over in each county (FIPS)
White Alone	Only White people in each county (FIPS)
Black or African American alone	Only Black or African American alone people in each county (FIPS)
Asian alone	Only Asian alone people in each county (FIPS)
Some other race alone	Only some other race alone people in each county (FIPS)

Table 2: Optimal number of cluster selection for k-means model.

Number Clusters	**Silhouette Score**	**Davies-Bouldin Score**
5	0.62	0.43
6	0.60	0.45
7	0.60	0.39
8	0.57	0.42
9	0.51	0.46
10	0.51	0.49
11	0.48	0.54
12	0.49	0.51
13	0.49	0.51

The hierarchy of the pre-clusters is illustrated in Figure within Appendix A. Subsequent hierarchical clustering resulted in 4 clusters. The silhouette score and Davies-Bouldin index are detailed in the accompanying table. While the Davies-Bouldin index decreased to 0.34, the silhouette score improved to 0.71. Additionally, trial and error indicated minimal variation in correlations, leading to the final selection of 4 clusters. The establishment of the final 4 classes opens opportunities to conduct a deeper comparative analysis of various factors among similar data points within each class.

Table 3: Optimal number of cluster selection for hierarchical agglomerative clustering.

Number Clusters	Silhouette Score	Davies-Bouldin Score
2	0.85	0.43
3	0.70	0.50
4	0.71	0.34

4 RESULT AND DISCUSSION

This section presents the findings from the two-step cluster analysis. In Sub-section 4.1, we explore the distribution of traffic, climate, and social features according to the International Roughness Index (IRI) within the pre-clusters generated by the KMeans model in the first step, as well as the final clusters formed through hierarchical agglomerative clustering in the second step. Sub-section 4.2 then examines how each feature's pattern correlates with IRI quantiles across the final four clusters. Lastly, sub-section 4.3 details the distribution of rural, small urban, and urban areas across IRI quintiles.

4.1 Features pattern in the Pre-clusters and Final Clusters

After the initial k-means clustering, climate, traffic, and social features distinctly divided the clusters based on IRI values, as depicted in Figure 2. Notably, socioeconomic variables like total population, median age, median income, and the distribution of racial groups including Black or African American Alone, Asian Alone, and Other Races demonstrated significantly differentiated patterns across the 10 clusters (Figure 2). Additionally, climate attributes like relative humidity and precipitation exhibited variations in the distribution of data points across each cluster.

The second step of clustering yielded more detailed results. The dendrogram, displayed in Appendix A, revealed that pre-clusters (2,8), (0,5), and (4,9) combined into a single cluster. Conversely, pre-clusters 3 and 7 did not merge with any other groups, remaining as two distinct clusters in all dendrogram methods (center, average, ward, and complete). Ultimately, cluster (1,6) also amalgamated into a final cluster during the second step. Consequently, Figure 3 illustrates the outcomes of the second clustering phase, where the initial ten pre-clusters have been consolidated into four definitive clusters. The results of the final clustering presented in Figure 3 indicate that population and racial composition, including the percentages of White alone and Black or African American alone, are among the features with significant differences across the four final clusters. Further analysis of these patterns will be explored in the subsequent section.

4.2 Traffic, Climate and Urban Features in the Final Clusters

The delineation into four final clusters allows us to observe variations in the determinants of pavement condition with other variables held constant. For instance, Clusters 0 and 2 predominantly encompass areas with a lower total population, whereas Cluster 1 includes regions with a higher population density (urban areas). Contrary to expectations, areas with higher AADT often had higher IRI values, indicating poorer pavement conditions. Upon analyzing socioeconomic factors across all four clusters, a noticeable trend emerged: an increase in the Black or African American population correlated with higher IRI values and deteriorating pavement conditions. A similar pattern was evident for the Asian alone and other race categories, particularly in Cluster 3, where the rise in IRI and these demographic factors was more pronounced compared to the other clusters (Figure 4).

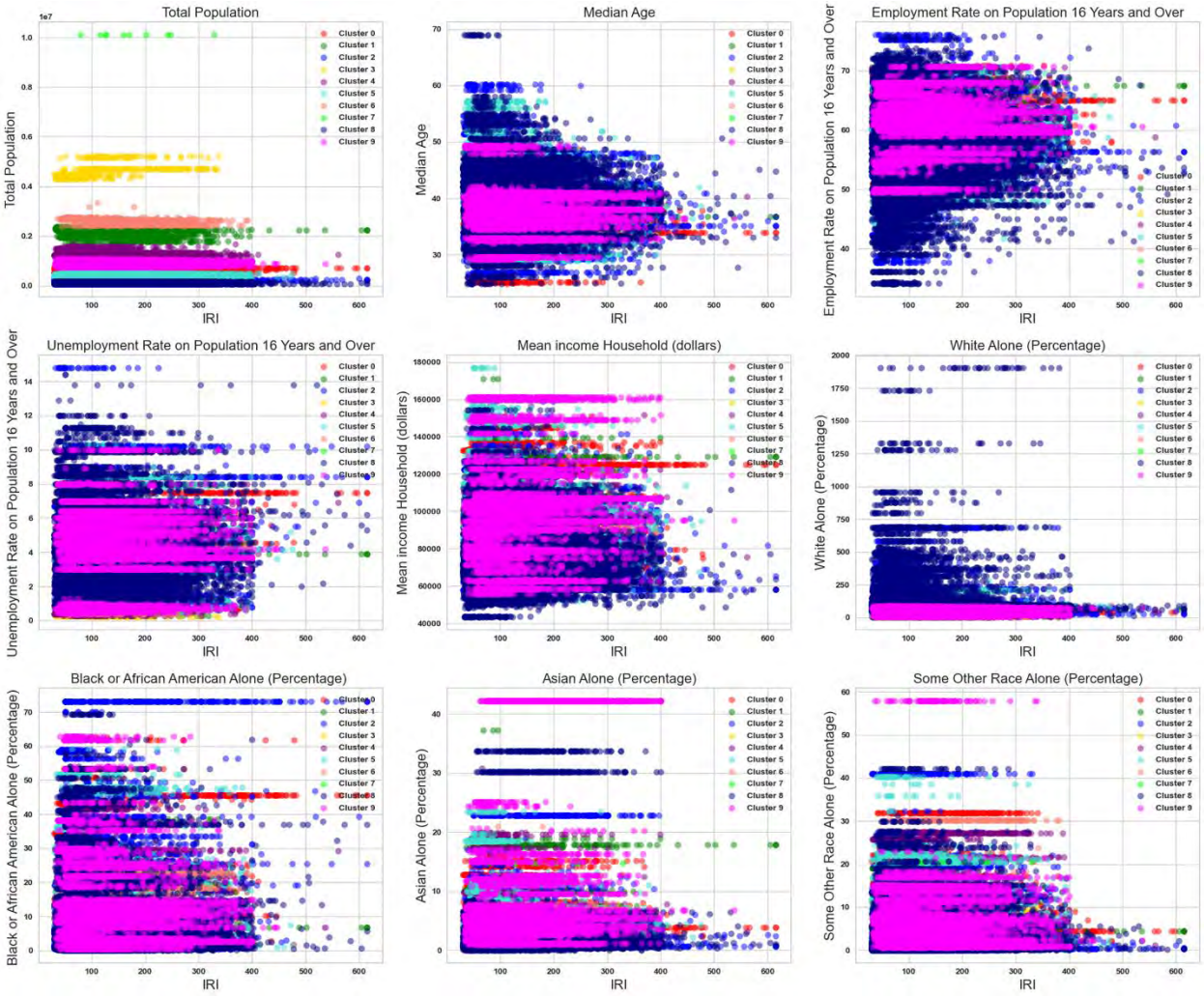

Figure 2: Features against the IRIs in the pre-clusters created by KMeans.

In contrast, other socioeconomic indicators such as the percentage of White alone, median age, or unemployment rate did not exhibit distinct patterns across the different IRI quartiles nationally. However, there was a slight trend where areas with lower unemployment rates experienced better pavement conditions, aligning with a lesser degree compared to the previous observations. An intriguing finding, particularly in Clusters 1 and 3, is the relationship between climate factors like humidity and precipitation and their patterns with the IRI. Specifically, areas with higher precipitation and humidity tend to have poorer pavement conditions, highlighting a notable correlation between these climate features and pavement quality. Conducting cluster analysis offers a valuable opportunity to monitor various attributes alongside pavement quality, significantly aiding regions within each cluster to account for these crucial attributes in their future asset management strategies.

4.3 Urbanity and Rurality in the Final Clusters

The data analysis delineated four primary clusters, effectively showing the rural, small urban, and urban count in the three quantiles of IRI values. This count was pivotal in understanding the impact of urbanity and rurality on different pavement conditions that lead to social inequity. As mentioned before, According to HPMS data structure, when the urban code is 9999, the region is rural; when the urban code is 9998, it is a small urban area, and the rest of the urban codes are urban areas (Figure 5). Notably, the findings from this segmentation highlighted a significant disparity in the pavement conditions of urban areas. This

revelation draws attention to the higher AADT, mobility, and lifestyle in urban areas, which may cause more damage to the pavement sections. This feature must be explored to understand the impact of urban, small urban, and rural areas in good, acceptable, and good pavement conditions.

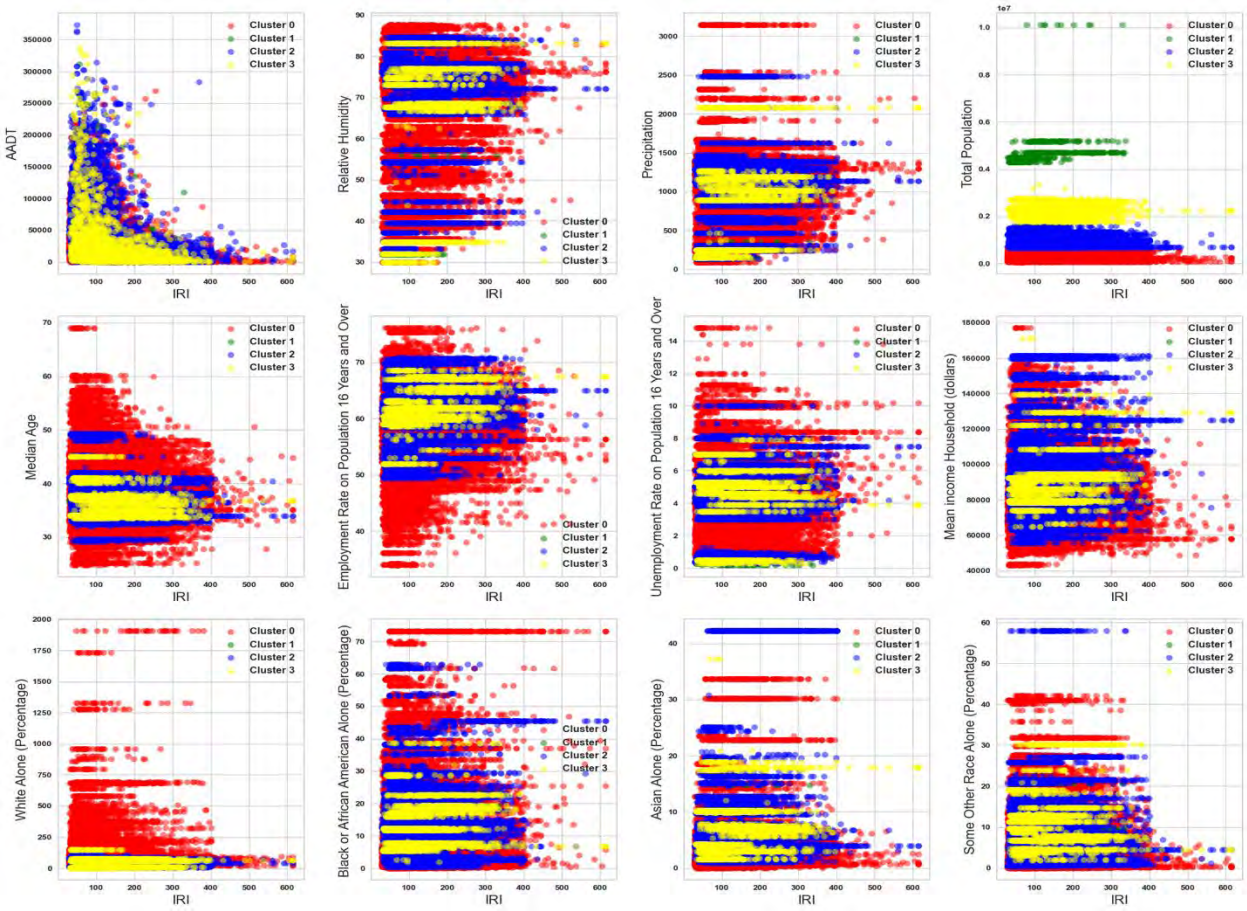

Figure 3: The traffic, climate, and social features after the final step.

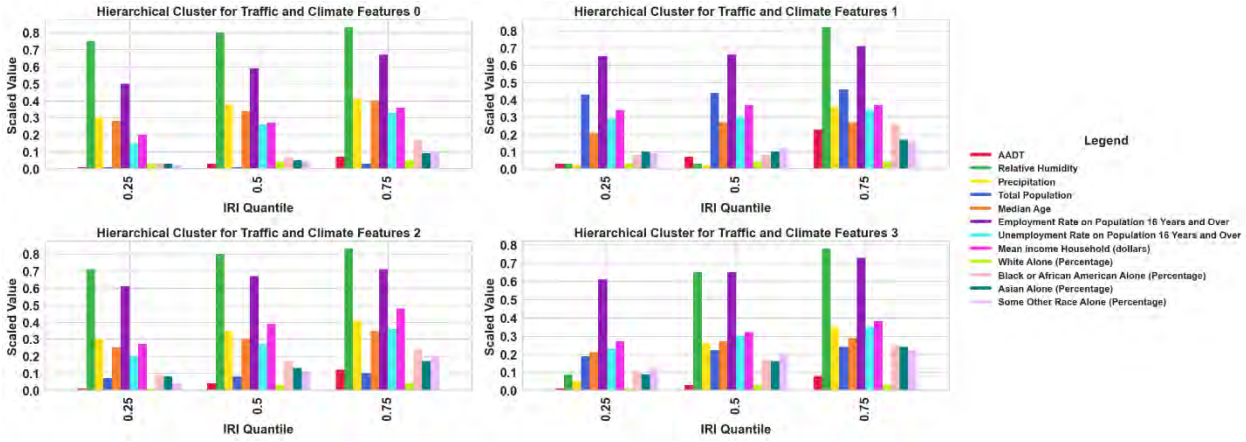

Figure 4: The traffic, climate, and social features after the final step.

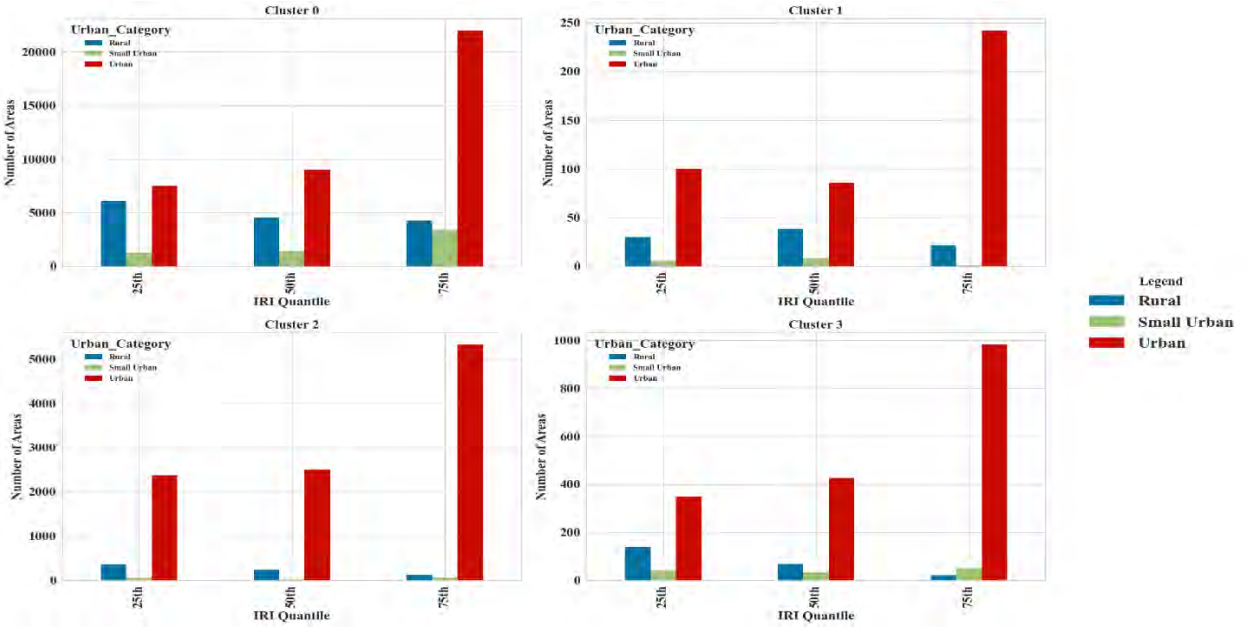

Figure 5: The urbanity and rurality the final step.

5 CONCLUSIONS

Utilizing a two-step clustering approach on the extensive HPMS dataset reveals the significant impact of socioeconomic factors on pavement conditions, correlating with the level of rehabilitation and maintenance that highways and roads receive. By including technical, traffic, climate, environmental, and social factors in our analysis, we gain a holistic understanding of how these variables fluctuate in relation to pavement quality as measured by the IRI index. In particular, the results indicated that urban regions experiencing higher traffic volumes, as well as areas with a larger Black or African American, Asian, and other racial population, tend to have reduced access to high-quality pavement. This observation holds true even when accounting for other factors, including technical specifications, environmental conditions, and climate influences. The influence of precipitation and relative humidity on pavement deterioration, as was the climatic influence on pavement conditions, was also evident. This trend underscores the pavement condition evaluation by social equity and highlights the necessity for targeted interventions to address the disproportionate impact of urbanization on pavement conditions.

The findings call for a nuanced understanding of the factors contributing to pavement degradation and the adoption of strategic approaches to infrastructure maintenance and improvement, particularly in densely populated urban areas. Most current maintenance simulations consider road network deterioration models with budget constraints, aiming to maximize the overall network condition within the available budget. While the entire network, or society, receives the highest benefit, it is important to consider how resources are allocated to different communities and regions. Future research should explore various metrics for identifying communities and optimize the road network to maximize the overall condition while addressing disparities between disadvantaged and non-disadvantaged communities. Our findings are constrained by the limited timeframe covered and the specific context of the US. First and foremost, this study considered only the IRI as the pavement condition metric, while other metrics are available. Evaluating socio-economic factors with multiple metrics would provide a more comprehensive assessment of social equity in pavement management systems. Another limitation was the data coverage duration; future research should include more data points. Lastly, this research focused solely on structured big data. Future studies could extend the evaluation of social equity in pavement conditions by incorporating unstructured data such as images, texts, or audio analysis.

A APPENDICES

NUMBER OF CLUSTER SELECTION IN AGGLOMERATIVE HIERARCHICAL CLUSTERING

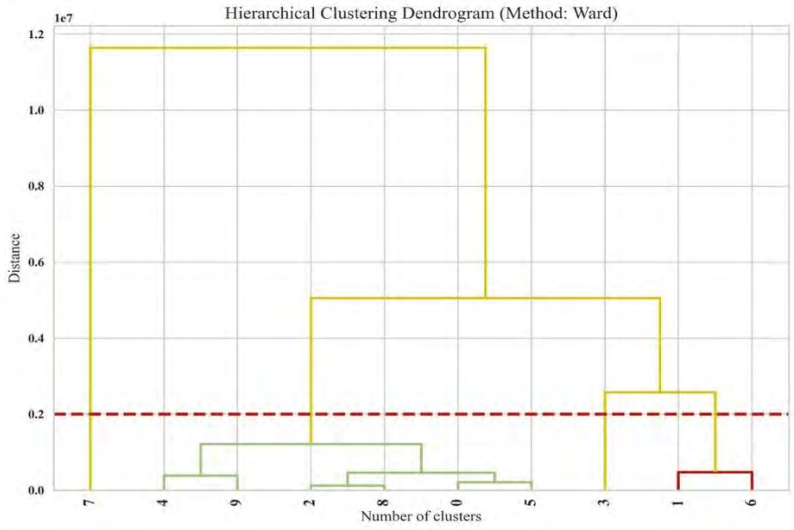

Figure 6: Dendrogram for optimal hierarchical clustering.

REFERENCES

Abdulhafedh, A. 2021. "Incorporating k-means, hierarchical clustering and pca in customer segmentation". *Journal of City and Development* 3(1), 12-30.

Cavallaro, F., Bruzzone, F., and Nocera, S. 2020. "Spatial and social equity implications for High-Speed Railway lines in Northern Italy". *Transportation Research Part A: Policy and Practice* 135, 327-340.

Currie, G., and A. Delbosc. 2010. "Modelling the social and psychological impacts of transport disadvantage". *Transportation* 37 953-966.

Corley-Lay, J. 2014. "Pavement performance measures: How states see good, fair, and poor." *Transportation Research Record*, 2431(1), 1-5.

Dabbura, I. 2018. K-means Clustering: Algorithm, Applications, Evaluation Methods, and Drawbacks. Towards Data Science. https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a, accessed 30th March 2024.

Gandy, C. A., Armanios, D. E., & Samaras, C. 2023. "Social equity of bridge management". *Journal of Management in Engineering*, 39(5), 04023027.

Erfani, A., J. Mahmoudi, & Q. Cui. 2024. "Measuring Social Equity in Pavement Conditions Using Big Data". In *Construction Research Congress* 2024 pp. 23-32.

Erfani, A., & Frias-Martinez, V. 2023. "A fairness assessment of mobility-based COVID-19 case prediction models". Plos one, 18(10), e0292090.

Evan, F. 2023. How Interstate Highways Gutted Communities—and Reinforced Segregation. History. https://www.history.com/news/interstate-highway-system-infrastructure-construction-segregation, accessed 29th March 2024.

Federal Highway Administration (FHWA). 2016. "Highway Performance Monitoring System Field Manual". Control No. 2125-0028. Office of Management & Budget (OMB).

Federal Highway Administration (FHWA). 2007. "Overview | Asset Management". FHWA-IF-08-008. Office of Asset Management.

IBM. 2024. "Two Step cluster node". https://www.ibm.com/docs/en/cloud-paks/cp-data/4.8.x?topic=modeling-twostep-cluster-node, accessed 29th March 2024.

Jain, A. 2019. Breaking Down the Agglomerative Clustering Process. Towards Data Science. https://towardsdatascience.com/breaking-down-the-agglomerative-clustering-process-1c367f74c7c2, accessed 30th March 2024.

Kassambara, A. Agglomerative Hierarchical Clustering. Datanovia. https://www.datanovia.com/en/lessons/agglomerative-hierarchical-clustering/, accessed 30th March 2024.

Khalife, F. G., Arneson, E. E., Atadero, R. A., & Ozbek, M. E. 2023. "Assessing social equity considerations within transportation asset management". *Transport Economics and Management*, 1, 160-167.

Kothari, C., J. France-Mensah, and O'Brien, W. J. 2022. "Developing a sustainable pavement management plan: Economics, environment, and social equity". *Journal of Infrastructure Systems* 28(2), 04022009.

Moscoso, Y. F., Rincón, L. F., Leiva-Maldonado, S. L., & ASC Campos e Matos, J. 2024. "Bridge deterioration models for different superstructure types using Markov chains and two-step cluster analysis". *Structure and Infrastructure Engineering*, 20(6), 791-801.

Murtagh, F., & Contreras, P. 2012. "Algorithms for hierarchical clustering: an overview". *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2(1), 86-97.

Ortega, E., E. López, and A. Monzón. 2012. "Territorial cohesion impacts of high-speed rail at different planning levels." *Journal of Transport Geography*, 24, 130-141.

Radovic, M., O. Ghonima, and T. Schumacher. 2017. "Data mining of bridge concrete deck parameters in the national bridge inventory by two-step cluster analysis." *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering* 3(2), F4016004.

Ramjerdi, F. 2006. "Equity measures and their performance in transportation". *Transportation Research Record*, 1983(1), 67-74.

Seyedrezaei, M., B. Becerik-Gerber, M. Awada, S. Contreras, and G. Boeing. 2023. "Equity in the built environment: A systematic review." *Building and Environment* 110827

Townsend, C. E. 2022. "Leveraging AI to assess inequities in pavement maintenance and rehabilitation strategies". Doctoral dissertation, University of Missouri, Columbia, Missouri.

U. S. Government Accountability Office (GAO). 2022. "National Highways: Analysis of Available Data Could Better Ensure Equitable Pavement Condition". GAO-22-104578

United States Census Bureau. 2023. https://www.census.gov/data.html, accessed 30th March 2024.

Welch, T. F., and S. Mishra. 2013. "A measure of equity for public transit connectivity." *Journal of Transport Geography*,33, 29-41.

AUTHOR BIOGRAPHIES

TAMIM ADNAN is a PhD student in the Department of Civil and Environmental Engineering at Michigan Technological University in Houghton, Michigan. His research interests include infrastructure management with artificial intelligence, intelligent systems, and big data. He completed his master's in construction and Facilities Management with a thesis on "Classification and Pixel-Level Segmentation of Asphalt Pavement Cracks Using Convolutional Neural Networks". His email address is tadnan@mtu.edu and his linkedIn address is https://www.linkedin.com/in/tamim-adnan-a740ab163/

ABDOLMAJID ERFANI is an Assistant Professor in the Department of Civil and Environmental Engineering at Michigan Technological University in Houghton, Michigan. His research interests include equity and diversity in transportation and workforce developments. His research expertise also extends to Data-driven Infrastructure Management, Project Delivery and Procurement, Text Analytics and Natural Language Processing, and Artificial Intelligence Modeling. He has authored numerous scientific research articles in his field on equity, natural language processing, and artificial intelligence. His email address is aerfani@mtu.edu and his website is https://www.mtu.edu/cege/people/faculty-staff/faculty/erfani/