Identifying Homeless Shelter Archetypes via K-Means Clustering to Inform State Responsibility

Hendro Budiyanto[1,*], Chininta Rizka Angelia[2]

[1]Faculty of Business Universitas Multimedia Nusantara, Scientia Boulevard, Tangereng, Banten, Indonesia

[2]Faculty of Communication Science, Universitas Multimedia Nusantara, Jakarta, Indonesia

ABSTRACT

Public policies designed to address homelessness often rely on broad, system-wide metrics that mask significant operational heterogeneity, leading to inefficient "one-size-fits-all" strategies. This paper argues that the availability of granular data and accessible machine learning techniques creates a new standard for state responsibility under international human rights law. We employ an unsupervised machine learning approach, specifically K-Means clustering, to analyze a dataset of daily operational metrics from a network of homeless shelters. The analysis is based on four key features: total capacity, occupancy rate, average age, and the percentage of male occupants. The clustering algorithm successfully identified four distinct and interpretable shelter archetypes, revealing a hidden typology within the system. The most critical finding is the emergence of a "Strained Mid-Sized Shelter" archetype, characterized by moderate capacity and the highest average occupancy rate, providing empirical evidence of a recurring state of systemic stress. The existence of these data-defined archetypes transforms the abstract risk of shelter failure into a concrete and foreseeable event. We conclude that this data-driven foreseeability elevates the state's duty to act under the Right to Adequate Housing (UDHR, Article 25). The failure to use available analytical methods to identify and respond to predictable patterns of strain can be construed as a breach of this duty. This study provides a novel framework for linking data science to legal accountability, advocating for the adoption of evidence-based, targeted policies that reflect the nuanced realities of social service provision. This approach offers a new paradigm for holding states accountable for protecting the rights of vulnerable populations in the digital age.

Keywords Clustering, Foreseeability, Homelessness, Human Rights, Public Policy

Submitted 26 July 2025
Accepted 12 August 2025
Published 1 September 2025

*Corresponding author
Hendro Budiyanto,
hendro.budiyanto@umn.ac.id

Additional Information and Declarations can be found on page 243

DOI: 10.63913/jcl.v1i3.13

Introduction

One prominent aspect of the integration of big data into public governance is its ability to enhance organizational efficiency and effectiveness. Research indicates that big data facilitates better decision support and fosters more informed policymaking by providing richer insights into ongoing social issues [1]. Big data analytics empowers policymakers to develop a nuanced understanding of citizen needs, enabling tailored solutions that enhance public welfare [2]. The emphasis on agile analytics reflects a shift towards real-time, data-driven governance models that prioritize continual learning and rapid adjustment of strategies to better serve the populace [3].

Furthermore, the role of data governance frameworks cannot be overstated. Such frameworks establish a structured approach to managing data quality, security, and compliance—essential elements as governance increasingly relies on cloud-based analytics [4]. The presence of comprehensive data governance mechanisms is crucial in mitigating potential risks and enhancing the efficacy of

data-focused initiatives [5][6]. Successful implementations of data governance frameworks in local government contexts highlight the importance of inter-agency collaboration and stakeholder participation in data analytics endeavors, further extending the potential for social welfare improvements [7].

In the case of homelessness, specific applications of big data can illuminate underlying trends and foster more effective interventions. By analyzing data from various sources—including social services, healthcare, and housing statistics—governments can identify patterns contributing to homelessness and deploy resources strategically [8]. Effective data analytics can ultimately enhance the targeting of services, allowing for timely interventions that address the immediate and long-term needs of vulnerable populations [9]. This approach aligns with the overarching goal of leveraging big data not merely for operational efficiencies but to create societal value through improved service delivery [2].

Moreover, the rise of open data initiatives has added another layer to the discourse on governance and social welfare. Open data platforms foster transparency and promote citizen engagement, granting stakeholders access to valuable datasets that can drive innovation in public services [3]. However, the effectiveness of such platforms is contingent upon the political and organizational support they receive, indicating the necessity of addressing existing tensions within governance structures to maximize their potential impact on democratic processes [8]. The reduction of silos in data management can facilitate smoother data sharing and enrich public policy discussions, leading to responses that are better aligned with community needs [10].

As organizations seek to implement big data initiatives, maturity models can provide a structured pathway for public sector entities to evolve in their use of analytics. These models elucidate key attributes of organizational readiness, ranging from technological capabilities to human resource competencies, which are critical for harnessing the full potential of big data in governance and policymaking [11]. Such frameworks enable public entities to assess their current position and target areas for development, ultimately cultivating environments where data-driven decision-making becomes ingrained in the organizational culture.

The methodological diversification provided by big data analytics tools also brings forth innovative approaches for problem-solving in public policy. For example, predictive analytics can be utilized to forecast instances of homelessness based on various societal indicators, allowing preemptive measures to be enacted [12]. The incorporation of AI alongside big data enhances this capability, as machine learning facilitates sophisticated modeling techniques that can identify complex interrelationships within the data that may not be readily apparent through traditional analytical methods [13]. Such advancements position big data not merely as a transformative force in public governance but also as a necessary instrument for addressing multifaceted social challenges.

However, challenges remain in the implementation of big data within public policy frameworks. Institutional barriers, such as rigid hierarchical structures and a lack of data literacy among staff, can impede the effectiveness of data-driven initiatives [14]. Continuous education and training are essential to cultivate a workforce equipped to utilize big data effectively [15]. This investment in human capital is vital for developing an analytically mature public sector that can leverage data insights to inform policy decisions and improve service outcomes.

Ultimately, the interplay between big data analytics, governance practices, and social welfare exemplifies a paradigm shift in public administration, where policy decisions are increasingly informed by empirical evidence and rigorous analysis. As governments embrace these methodologies, they are better positioned to formulate policies that respond to immediate social crises, such as homelessness, and foster long-term sustainability and resilience within communities [16]. As this field continues to evolve, ongoing research and adaptation will be necessary to address the complexities and ethical considerations associated with data use in governance.

A compelling argument for the ineffectiveness of uniform homelessness policies is highlighted in the literature concerning specific demographics, such as older adults and women. Research indicates that traditional policies have predominantly centered around the needs of male populations, often overlooking the unique challenges faced by women experiencing homelessness [17]. A more effective approach necessitates a systems framework that accommodates varied experiences and aligns with the specific support needs of diverse subpopulations [18]. The failure to implement such tailored interventions can contribute to the continued marginalization of those most at risk, exacerbating cycles of poverty and vulnerability [19].

Furthermore, findings indicate that public attitudes towards homelessness can significantly shape policy responses, impacting the effectiveness of intervention strategies [19]. Historical shifts in public perception are correlated with policy changes, highlighting the need for interventions that resonate with community sentiments while being informed by social science. This can aid in promoting data-driven policies that address the root causes of homelessness rather than merely responding to its symptoms [20]. The misconception that blanket policies can resolve the multifaceted aspects of homelessness often leads to neglect of fundamental issues, such as systemic barriers to adequate housing and social services [17].

In examining evidence-based approaches to resource allocation, the "Housing First" model emerges as a leading example where nuanced strategies have shown efficacy in tackling chronic homelessness. Advocates for the Housing First approach stress its focus on providing stable housing without preconditions, addressing the immediate need for shelter while facilitating access to essential social services [18]. However, challenges persist, as implementation frequently encounters barriers related to inconsistent policy frameworks and inadequate funding for supportive services [21]. Ongoing evaluation and adaptation of such programs are crucial to maximize their potential across diverse contexts [22].

Another important aspect is the impact of social stigma and public framing of homelessness on governmental responses. Studies have shown that effective communication and advocacy are vital for reshaping public perceptions, which can translate into more supportive legislation and funding for vulnerable populations [22][20]. Addressing these representations through public discourse is necessary to promote policies that are not only well-received but are also adequately equipped to tackle the chronicity of homelessness and its underlying causes [22].

Racial disparities in homelessness highlight the necessity for targeted policies that confront systemic inequities impacting marginalized groups [23] [24]. The higher rates of homelessness among Black and Indigenous populations reveal

the nuances that generic policies often overlook, emphasizing the urgent need for racially informed approaches to social welfare [23]. Developing responsive, culturally competent strategies necessitates extensive engagement with affected communities, ensuring that their voices inform the policymaking process [24].

This paper proposes a new paradigm for legal accountability in social welfare, arguing that unsupervised artificial intelligence can be leveraged to create a more robust and evidence-based standard for state responsibility. We demonstrate the application of machine learning clustering to a dataset of homeless shelter operational metrics, a method that moves beyond traditional aggregate statistics. This approach allows for the identification of distinct, data-defined archetypes of shelter situations, revealing a hidden typology of operational realities that are often obscured by system-wide averages. By segmenting the data into these naturally occurring groups, we can pinpoint specific, recurring scenarios of systemic stress and need that demand a more nuanced policy response.

The central argument of this thesis is that these identified archetypes establish a new, higher standard of foreseeability, which legally obligates the state to develop targeted policies under international human rights law. When a specific, high-risk archetype—such as a "strained shelter"—can be empirically identified and defined, the failure of that part of the system is no longer a random or unforeseeable event. It becomes a predictable outcome that the state has a duty to mitigate. This data-driven foreseeability strengthens the legal basis for demanding that states move away from inefficient, "one-size-fits-all" strategies and instead implement precise, evidence-based interventions tailored to the specific needs of the most vulnerable, as defined by the data itself.

Literature Review

Data Analytics in Public Administration

The application of data analytics in public administration, particularly within the social services sector, has garnered increasing attention in recent years through predictive and analytical models. This evolving landscape unveils both the potential benefits of enhanced decision-making and serious limitations, ethical challenges, and criticisms that accompany these data-driven approaches. By reviewing existing scholarship, we can better understand the capabilities and constraints of predictive analytics, particularly in contrast with unsupervised learning methodologies.

Predictive analytics has emerged as an influential tool for public administration, playing a crucial role in identifying trends and assisting resource allocation within the social services domain. The utilization of predictive models enables public agencies to forecast demand and better direct resources to areas of greatest need, leveraging large datasets that capture various socioeconomic indicators [25][26].For instance, through robust analytics, predictive models can identify populations at risk of homelessness, potential fraudulent activities, or early indicators of social distress [27]. In this regard, data analytics can significantly enhance the efficacy of social service programs by anticipating needs rather than merely responding to crises as they arise (-, 2024).

However, the rise of predictive policing models rooted in data analytics has sparked substantial debate regarding their ethical implications and limitations.

One of the primary concerns relates to the inherent biases present within the datasets utilized. Predictive models often reflect historical inequalities, and when these biases remain unaddressed, they can perpetuate discriminatory practices [28]. For instance, data-driven systems used in policing may disproportionately target marginalized communities, leading to over-policing and systemic injustices [28][29]. This raises essential questions about fairness and accountability in algorithmic decision-making processes within governance structures, which necessitates rigorous scrutiny and ongoing evaluation [30].

Moreover, predictive analytics' reliance on historical data can be problematic when predicting future outcomes in rapidly changing environments. Critics argue that static models may fail to adapt to new variables or shifts in social behavior, ultimately leading to misguided interventions [31]. This concern is particularly relevant in social contexts impacted by dynamic and evolving conditions, such as public health crises or economic downturns. The potential misalignment between predictive outputs and real-world complexities underscores the limitations of strictly relying on predictive analytics for critical policymaking decisions.

Contrastingly, unsupervised learning techniques provide a fundamentally different approach to data analysis, allowing for exploratory insights that are not constrained by predefined categories or assumptions. Unsupervised learning can uncover hidden patterns within data that predictive models may overlook due to their reliance on historical trends [32]. For example, clustering algorithms can identify distinct groups within homeless populations, revealing unique needs and potential interventions tailored to those groups. This method supports a more nuanced understanding of complex social issues, conducive to developing specialized programs that are responsive to varied circumstances [33].

While both predictive and unsupervised learning contribute valuable perspectives to understanding and addressing social service needs, the divergent nature of these methodologies highlights the imperative for comprehensive ethical considerations in their application. Important ethical challenges include transparency, interpretability, and the potential for misuse of data in decision-making processes [30][34]. Policymakers must be aware of these concerns and actively seek to cultivate frameworks that emphasize accountability and ethical data stewardship, ensuring that technology serves to enhance, rather than hinder, social equity.

Data privacy concerns also emerge as a critical issue in the implementation of predictive and analytical models. The extent of data collection required for nuanced analytics raises questions about individuals' rights and the potential for exploitation of personal information [35][36]. Safeguarding sensitive data while maximizing the utility of analytics requires robust legal frameworks and governance structures that align with ethical standards, thus maintaining public trust in the systems of governance [27][37].

Further, as the field continues to evolve, it becomes increasingly important for public institutions to adopt inclusive strategies that incorporate stakeholder input into the design of analytical models. Engaging communities can foster a deeper understanding of data needs and usage while ensuring that affected populations are considered in decision-making processes [23][38]. Collaboration with local organizations and advocacy groups can aid in addressing the diversity of needs present among various demographic segments, ultimately leading to more effective resource allocation and social services.

Technology, Foreseeability, and State Duty in Cyber Law

The evolving landscape of cyber law is intricately connected to the state's duty to utilize available technological means in protecting fundamental rights, particularly under human rights frameworks. Legal scholarship surrounding this theme has increasingly focused on the implications of data-driven foreseeability, especially in terms of how states identify and manage high-risk groups. This analysis delves into the intersection of technology, foreseeability, and the resultant legal standards for negligence, illustrating the complexities and responsibilities faced by governments in a digitally connected world.

As cyber threats grow more sophisticated, the obligations of states to ensure cybersecurity and protect citizens' rights have become paramount. Faga argues that the rise of transnational cyber threats necessitates a reassessment of existing legal frameworks, indicating that the traditional distinctions between cybercrime, cyber-attacks, and cyber warfare require nuanced re-evaluation to safeguard international humanitarian law (IHL) [39]. This evolving standard compels states to proactively address cyber vulnerabilities, placing heavier responsibilities upon them to effectively employ available technological tools [39]. Such an obligation includes adapting laws to provide robust protection against entrenched cyber violations that threaten fundamental human rights.

The concept of ‘cyber due diligence’ is emerging as a critical discourse in determining the responsibilities of states regarding cybersecurity. Coco and Dias contend that the patchwork of protective obligations within international law prompts a shift from a binary perspective on state duties to a more nuanced understanding of the varying responsibilities to safeguard citizens and other states from cyber harm [40]. The adoption of this framework underscores the necessity for states to take affirmative steps to utilize the best available technology and practices to predict and respond to cybersecurity threats, thereby preventing significant harm.

In legal contexts, foreseeability refers to the ability to anticipate potential risks based on available data [40]. In cybersecurity, identifying high-risk groups through data analytics necessitates a heightened level of diligence from state authorities. If a government fails to predict and protect against foreseeable harm to these vulnerable groups due to cybersecurity breaches, it may face negligence claims and heightened liability under human rights frameworks [41]. Thus, public authorities must harness predictive analytics responsibly to ensure proactive measures are in place, fulfilling their legal and moral duties to protect citizens.

Method

This study employed an unsupervised machine learning approach to identify naturally occurring groups or archetypes within a dataset of homeless shelter operational metrics. The objective was to move beyond a monolithic view of the shelter system by using an exploratory, data-driven methodology to reveal underlying heterogeneity. Unsupervised learning is uniquely suited for this task, as it discovers inherent patterns and structures within data without preconceived labels or target outcomes. The analysis was conducted using Python, leveraging the scikit-learn library for machine learning, pandas for data manipulation, and matplotlib with seaborn for visualization.

Data Preparation and Feature Selection

The analysis was performed on an anonymous dataset comprising daily operational records from a network of homeless shelters. From this dataset, four key quantitative features were selected to serve as the basis for clustering: total_capacity, occupancy_rate, average_age, and male_percentage. The selection of these features was deliberate. total_capacity serves as a proxy for a shelter's scale and resource level, while occupancy_rate provides a direct measure of the demand-supply dynamic, acting as a critical indicator of operational strain. The demographic features, average_age and male_percentage, were included as they are fundamental descriptors of the population being served, which can inform the specific types of services and support structures that may be required. A critical preprocessing step was performed to ensure that each feature contributed equally to the clustering process. The data was standardized using the StandardScaler function from scikit-learn. This function transforms each feature to have a mean of zero and a standard deviation of one, a process known as z-score normalization. This step is essential for distance-based algorithms like K-Means, which are highly sensitive to the scale of input variables. Without normalization, features with larger numerical ranges (e.g., total_capacity) would disproportionately influence the Euclidean distance calculations at the core of the algorithm, effectively overshadowing the contributions of features with smaller ranges (e.g., occupancy_rate). Standardization ensures that each feature is on a common scale, allowing the algorithm to discern patterns based on the relative relationships within the data, not arbitrary measurement units.

Clustering Algorithm and Parameter Tuning

The primary methodology utilized was K-Means clustering, an iterative, centroid-based algorithm that partitions data into a predetermined number of distinct, non-overlapping subgroups. The algorithm's objective is to minimize the within-cluster sum of squares (WCSS), also known as inertia, thereby creating clusters that are as internally coherent and externally separated as possible. A crucial parameter in this method is the number of clusters (K). To empirically determine the optimal value for K, the Elbow Method was applied. This heuristic involves executing the K-Means algorithm on the scaled dataset for a range of K values, in this case from one to ten. For each iteration, the model's inertia was calculated and plotted. The resulting graph visualizes the trade-off between the number of clusters and the total inertia. As K increases, inertia naturally decreases, but the rate of this decrease slows. The "elbow" point on the plot represents the point of diminishing returns—where the marginal gain in explanatory power from adding another cluster is no longer worth the cost of increased model complexity. The analysis of the inertia plot revealed a distinct "elbow" at K=4, indicating that four clusters provided the most meaningful and parsimonious grouping of the data. Based on this empirical evidence, an optimal K of four was selected for the final model.

Cluster Identification and Visualization

With the optimal number of clusters established, the final K-Means algorithm was fitted to the scaled dataset. The model was configured with n_init=10, a parameter that instructs the algorithm to run ten times with different random centroid initializations and select the run that yields the lowest inertia. This approach mitigates the risk of settling on a suboptimal local minimum, which is

a known sensitivity of the K-Means algorithm, thereby enhancing the stability and validity of the final cluster assignments. For reproducibility of the results, a random_state of 42 was used to ensure that the same pseudo-random centroid initializations are used every time the code is executed, a critical component for academic verification. This process assigned each data point, representing a unique shelter-day record, to one of the four identified clusters. Finally, to facilitate the interpretation and visualization of these multi-dimensional clusters, Principal Component Analysis (PCA) was employed. It is important to note that PCA was used solely for visualization and not for the clustering itself, which was performed on the original four-dimensional feature space. PCA is a dimensionality reduction technique that transforms the data into a new coordinate system of orthogonal axes, or principal components, that capture the maximum possible variance. By configuring PCA with n_components=2, the four selected features were reduced into two principal components. Projecting the data onto these two components allowed the distinct clusters to be plotted and visually inspected on a two-dimensional scatter plot, providing an intuitive confirmation of their separation and coherence.

Result and Discussion

Results of Exploratory Data Analysis

Prior to clustering, an exploratory data analysis (EDA) was conducted to understand the fundamental characteristics and distributions of the dataset. This initial analysis provided crucial context for the subsequent machine learning application and revealed key trends within the shelter operational data.

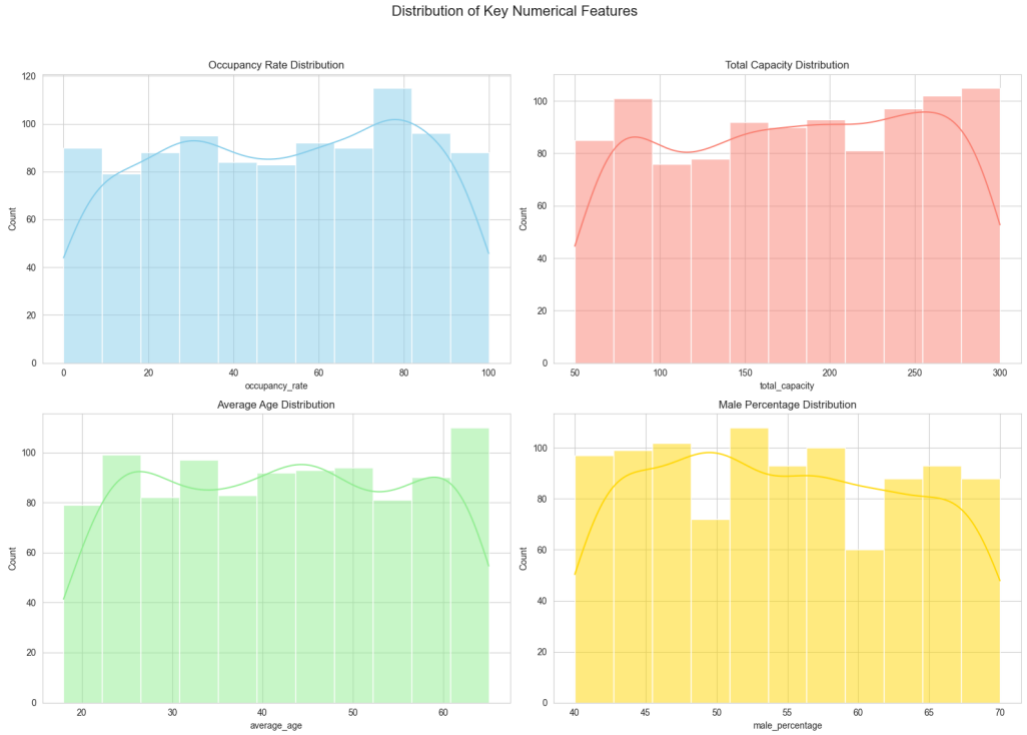

Figure 1 Distribution of Key Numerical Features

An examination of the four primary numerical features selected for clustering reveals distinct distributions, as shown in figure 1. The occupancy_rate displays

a relatively uniform distribution with a slight peak between 70-90%, indicating that while shelters experience a wide range of occupancy levels, days of high occupancy are common. The total_capacity is bimodal, with significant concentrations of smaller shelters (50-100 capacity) and larger shelters (250-300 capacity), suggesting two common scales of operation. The average_age of shelter occupants is spread fairly evenly from 20 to 65, while the male_percentage shows a left skew, with most shelter-day records reporting a male population between 45% and 65%.

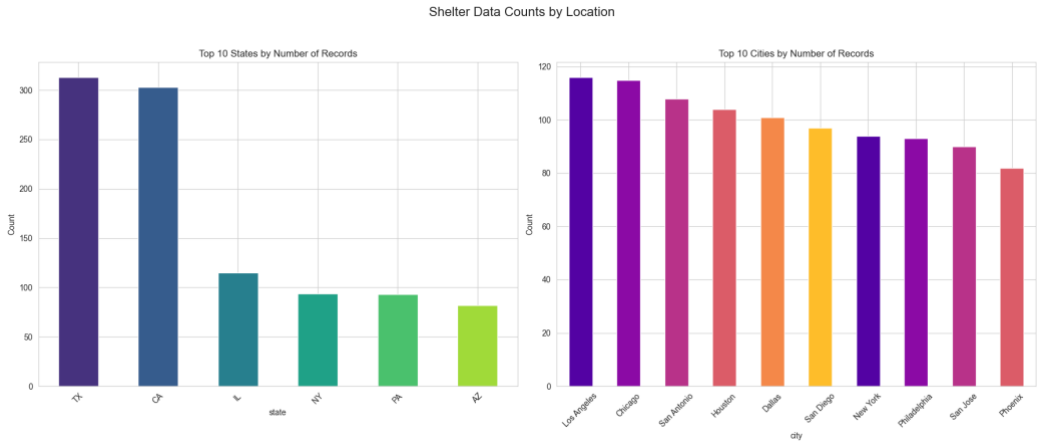

Figure 2 Geographic Distribution of Data Records

The geographic distribution of the data records, visualized in figure 2, highlights the primary locations represented in the dataset. The analysis shows a significant concentration of records from Texas (TX) and California (CA), which together account for a majority of the data points. The city-level data is more evenly distributed among the top 10, with Los Angeles and Chicago having the highest number of records. This geographic context is important for understanding the potential regional influences on the operational patterns observed in the data.

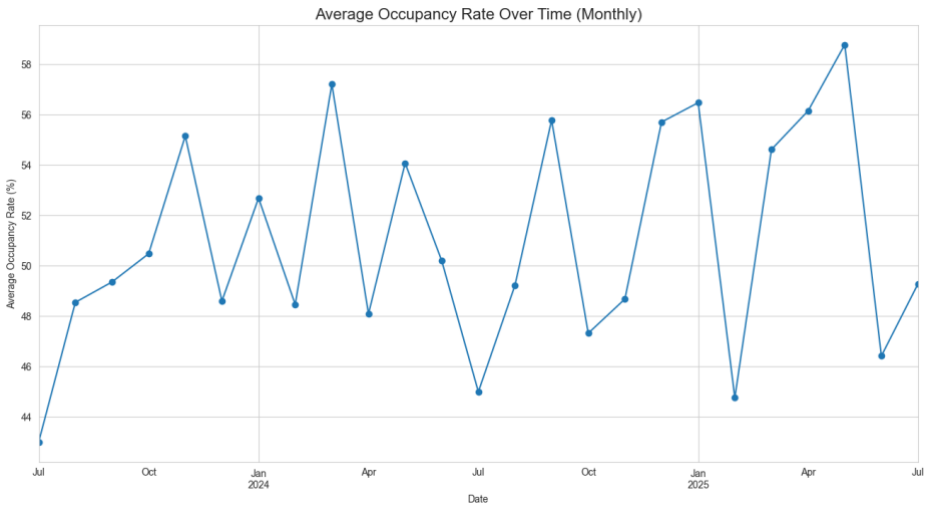

Figure 3 Average Occupancy Rate Over Time (Monthly)

A time-series analysis of the average monthly occupancy rate from July 2023 to July 2025 (figure 3) reveals significant volatility and potential seasonal patterns.

The plot shows sharp peaks and troughs, with notable increases in occupancy during the fall and spring months and dips during the summer. For instance, occupancy rates consistently peak around October and April, while dropping in months like July. This cyclical pattern underscores the dynamic nature of shelter demand over time.

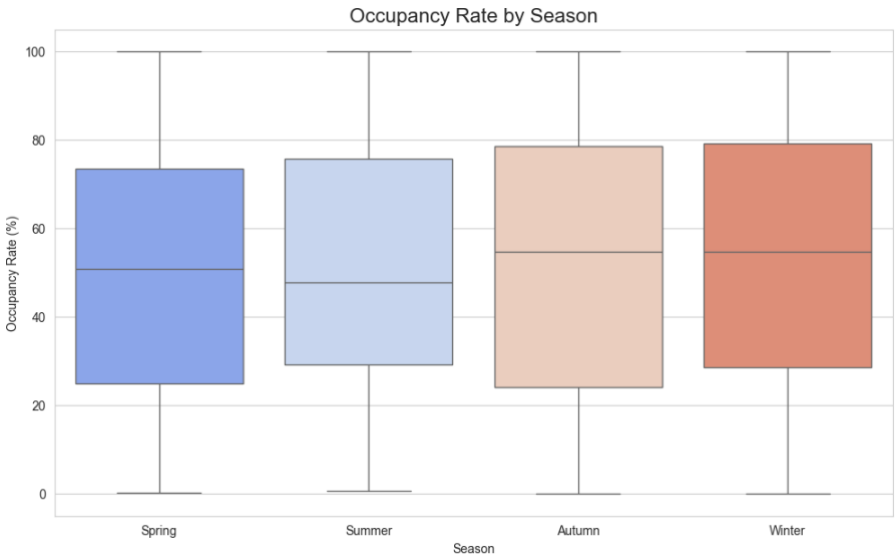

Figure 4 Occupancy Rate by Season

To further investigate the seasonal trends observed in the time-series plot, the occupancy rates were grouped by season. The boxplot in figure 4 confirms that shelter occupancy varies significantly with the seasons. The median occupancy rate is highest during the Autumn and Winter months, which also exhibit a wider range of occupancy values, indicating greater variability. Conversely, Spring and Summer show lower median occupancy rates. This finding aligns with common understandings of homelessness, where demand for shelter increases during colder weather, but it provides a quantitative validation of this trend within the dataset.

Clustering Results

The analytical process began with determining the optimal number of clusters (K) to partition the dataset. To achieve this empirically, the Elbow Method was employed, with the results visualized in Figure 1. The plot shows the model's inertia (the within-cluster sum of squares) calculated for a range of K values from 1 to 10. A distinct "elbow" is visible at K=4, where the rate of decrease in inertia slows considerably. This point indicates that adding more clusters beyond four yields diminishing returns in explaining the data's variance. Based on this graphical evidence, K=4 was selected as the optimal number of clusters for the final K-Means analysis, providing a methodologically sound and parsimonious model.

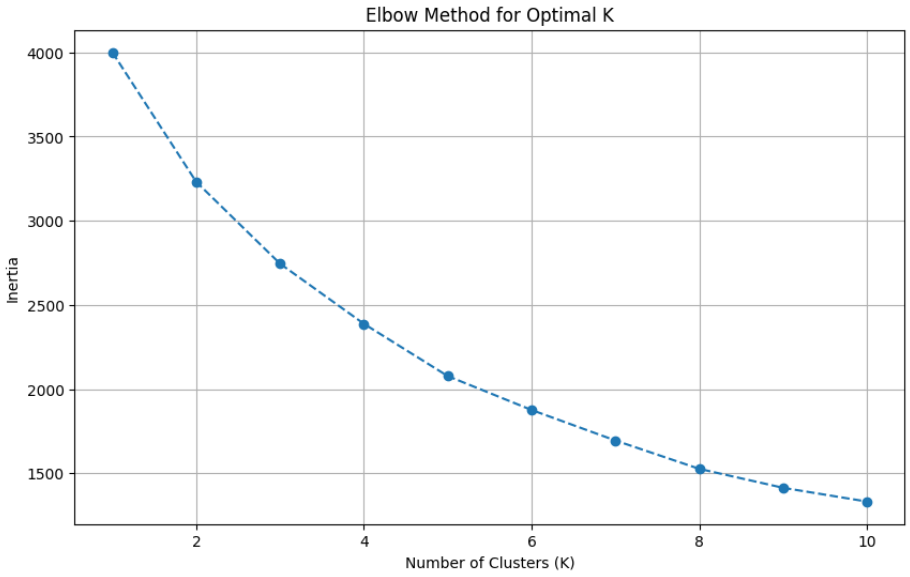

Figure 5 Elbow Method for Optimal K

Following the execution of the K-Means algorithm with K=4, the resulting clusters were visualized to confirm their distinctness and separation. As the clustering was performed on four features, Principal Component Analysis (PCA) was used to reduce the data's dimensionality to two principal components for plotting. Figure 6 presents a scatter plot of these components, with each point representing a shelter-day record and colored according to its assigned cluster. The plot clearly shows four distinct groupings of data points, corresponding to the four identified archetypes. The visual separation between the clusters—for instance, the concentration of purple points (Cluster 0) in the upper left and yellow points (Cluster 3) in the lower right—provides strong visual confirmation that the algorithm successfully identified meaningful and coherent patterns within the operational data.

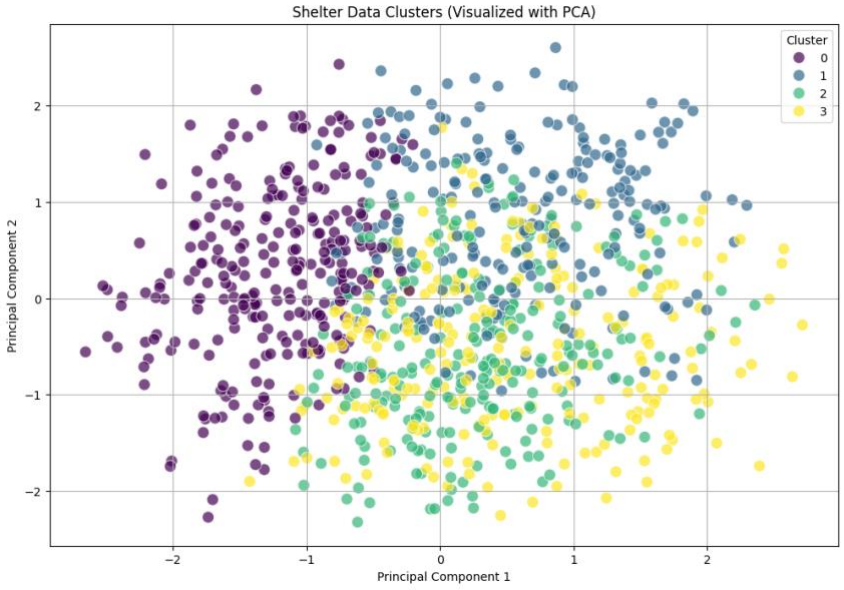

Figure 6 Shelter Data Clusters (Visualized with PCA)

The table 1 presents the quantitative profiles of the four identified clusters, detailing the mean values for each feature and the descriptive archetype name assigned to each group.

Table 1 Quantitative Profiles of Identified Clusters

Cluster	total_capacity	occupancy_rate	average_age	male_percentage	Archetype Name
0	235.64	68.10%	40.18	61.64%	Large, Busy Male-Majority Shelters
1	116.77	76.31%	42.17	50.50%	Strained Mid-Sized Shelters
2	233.79	34.95%	45.30	47.81%	High-Capacity, Low-Occupancy Shelters
3	119.91	24.50%	40.68	57.69%	Underutilized Mid-Sized Shelters

Identification of Four Distinct Shelter Archetypes

The quantitative profiles of the four clusters reveal significant heterogeneity within the homeless shelter system, challenging any monolithic policy approach. Cluster 0, "Large, Busy Male-Majority Shelters," represents the large-scale, traditional urban shelter model. These facilities operate with a consistently high number of occupants relative to their substantial capacity and serve a predominantly male population. The operational focus here is likely on logistics, crowd management, and providing basic necessities at scale. In stark contrast, Cluster 2, "High-Capacity, Low-Occupancy Shelters," describes facilities of a similar large scale but with an average occupancy rate of only 35%. This low utilization, coupled with a slightly older and more gender-balanced population, suggests a potential misalignment between the services offered and the needs of the local homeless population. These may be facilities with barriers to entry, a poor reputation, or a service model that does not cater to the specific demographic, such as a lack of resources for older adults or families.

Similarly, the two smaller-capacity archetypes also show a critical divergence in operational reality. Cluster 3, "Underutilized Mid-Sized Shelters," operates with an extremely low average occupancy rate of just 24.5%. This indicates a significant surplus of available beds relative to daily demand, which could point to various factors such as geographic isolation, restrictive intake policies, or being located in an area with lower overall need. The most critical finding for this study, however, is the emergence of Cluster 1, the "Strained Mid-Sized Shelters." This archetype is defined by a potent combination of moderate capacity and the highest average occupancy rate (76.31%) across all clusters. These shelters are consistently operating near or at their functional capacity, creating a high-stress environment for both staff and residents. Serving a slightly

older population with an even gender distribution, these facilities are likely the frontline crisis response centers in their communities. The existence of this cluster provides robust empirical evidence of a specific, recurring state of systemic stress within the shelter network that demands a targeted policy response.

Legal Implications for the Right to Adequate Housing

The data-driven identification of these archetypes, particularly the "Strained Mid-Sized Shelters" (Cluster 1), has profound implications for the legal concept of state responsibility. In legal doctrine, foreseeability is a cornerstone of duty and negligence. An outcome is considered foreseeable if a reasonable person—or in this case, a state entity equipped with modern analytical tools—should have anticipated it. This analysis transforms the abstract, generalized risk of shelter overcrowding into a concrete, data-defined, and therefore foreseeable event. The state is no longer merely aware of a general problem; it now possesses, through this methodology, the technological means to identify a specific, recurring profile of systemic failure. This moves the issue from the realm of unfortunate circumstance to one of predictable, and thus preventable, harm.

This newfound level of data-driven foreseeability elevates the state's duty under international human rights law, specifically the Right to Adequate Housing as articulated in Article 25 of the Universal Declaration of Human Rights. The simple provision of a uniform, "one-size-fits-all" funding model or policy for all shelters is rendered insufficient and arguably negligent when evidence clearly shows that distinct, predictable situations of strain exist. For example, a policy that allocates funding based solely on total capacity would treat Cluster 0 and Cluster 2 identically, ignoring the fact that Cluster 2 is chronically underutilized while shelters in Cluster 1 are perpetually strained. The identification of Cluster 1 creates a legal and ethical imperative for the state to develop and implement targeted policies. Such policies could include proactive resource allocation to increase staffing, specialized support for case managers dealing with high-stress environments, or the establishment of clear overflow planning and partnerships specifically for shelters that match the "Strained" archetype. The failure to implement such targeted interventions could now be construed as a foreseeable and preventable breach of the state's obligation to ensure adequate housing. This unsupervised AI approach thus creates a new paradigm for accountability, where the state's duty is not just to act, but to act with the nuance and precision that modern data analytics makes possible.

Comparison with Previous Research

This study contributes to and diverges from existing scholarship in two key domains. First, within public administration and social welfare research, studies of homelessness have traditionally relied on aggregate descriptive statistics—such as city-wide point-in-time counts or system-wide occupancy rates—to characterize the problem. While essential for macro-level planning and securing federal funding, these approaches often obscure the operational heterogeneity of the system due to an "averaging out" effect. A system-wide 75% occupancy rate, for instance, might mask a dangerous reality where half the shelters are perpetually at 100% capacity while the other half are half-empty. This research departs from that tradition by applying an unsupervised machine learning lens to reveal these granular, data-driven personas of shelter states. It provides a meso-level analysis that is more actionable for operational decision-making.

Instead of asking "how full is the system?," we ask "what distinct types of situations exist within the system?," a question that is far more relevant for targeted resource allocation and policy design.

Second, in the legal domain, scholarship on the Right to Adequate Housing has typically focused on broad principles of state obligation, often relying on qualitative evidence of system failure that can be dismissed as anecdotal. This paper builds a novel bridge between data science and legal duty by introducing a quantitative, replicable methodology to the conversation. It argues that the outputs of clustering algorithms can create a new, more stringent standard of foreseeability, thereby operationalizing the state's responsibility in a technologically-informed manner. The central argument is that foreseeability can now be tied to a state's capacity to perform such an analysis. If a state possesses the data and the accessible analytical tools to identify a recurring "Strained Shelter" archetype but fails to do so, its claim of ignorance is significantly weakened. This approach moves the legal argument from a purely philosophical debate to an evidence-based one, creating a testable benchmark for what a state should have known and strengthening the case for accountability under international human rights frameworks.

Limitations

Several limitations should be acknowledged. First, the findings are based on a dataset from a single shelter network, and their generalizability to other jurisdictions with different demographic profiles, funding models, or policy environments remains to be tested. Second, the analysis is constrained by the available features. The identified archetypes are purely operational and do not include crucial client-level data such as mental health status, reasons for homelessness, or outcomes after exiting the shelter. The inclusion of such data would undoubtedly yield a more nuanced and holistic typology. Finally, the K-Means algorithm itself imposes certain limitations; it creates hard, discrete boundaries between clusters, whereas in reality, a shelter may fluidly transition between states or exhibit characteristics of multiple archetypes.

Future Research Suggestions

This study opens several avenues for future research. A clear next step is to apply this clustering methodology to datasets from different cities and countries to test the robustness and universality of the identified archetypes. Future work should also seek to enrich the analysis by incorporating a wider array of features, including client-level outcomes, staffing ratios, and qualitative data from shelter managers, to move from operational archetypes to more comprehensive socio-ecological models. A longitudinal analysis, tracking how individual shelters move between these clusters over time (e.g., in response to seasonal changes or policy interventions), could provide invaluable insights into system dynamics. Finally, this data-driven framework could serve as the foundation for developing predictive models to identify shelters at high risk of becoming "Strained," enabling preemptive rather than reactive resource allocation and support.

Conclusion

This study successfully demonstrated the utility of unsupervised machine learning in deconstructing the complex operational landscape of homeless shelter systems. By applying K-Means clustering to daily operational data, we

have moved beyond monolithic, system-wide averages to identify four distinct, data-driven archetypes of shelter situations, including the critically important "Strained Mid-Sized Shelter." This core finding provides empirical evidence that the challenges within the homelessness sector are not uniform; rather, they are a collection of specific, recurring, and identifiable scenarios. The research contributes a new, nuanced framework for understanding risk and need, offering a replicable methodology that public bodies can use to gain a more granular and actionable understanding of the social services they administer. Ultimately, this paper argues for a paradigm shift in how we conceptualize state responsibility in the digital age. The ability to identify a foreseeable pattern of systemic strain, such as the "Strained Shelter" archetype, is not merely an analytical exercise; it creates a new legal and ethical imperative to act with precision. We contend that under international human rights law, this data-driven foreseeability obligates the state to move beyond inefficient, uniform policies and develop targeted interventions tailored to the specific, evidence-based needs of each identified group. We recommend that policymakers adopt similar clustering methodologies to inform a more just and effective allocation of resources, and we call for further legal scholarship to define the standards by which a state's duty to use available data to protect fundamental rights should be measured and enforced.

Declarations

Author Contributions

Conceptualization: H.B.; Methodology: C.R.A.; Software: C.R.A.; Validation: C.R.A.; Formal Analysis: H.B.; Investigation: C.R.A.; Resources: H.B.; Data Curation: C.R.A.; Writing Original Draft Preparation: H.B.; Writing Review and Editing: H.B.; Visualization: C.R.A.; All authors have read and agreed to the published version of the manuscript.

Data Availability Statement

The data presented in this study are available on request from the corresponding author.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] B. Klievink, B.-J. Romijn, S. W. Cunningham, and H. de Bruijn, “Big Data in the Public Sector: Uncertainties and Readiness,” *Inf. Syst. Front.*, 2016, doi: 10.1007/s10796-016-9686-2.
[2] P. B. Putera and R. Pasciana, “Big Data for Public Domain: A Bibliometric and Visualized Study of the Scientific Discourse During 2000–2020,” *Policy Gov. Rev.*, 2021, doi: 10.30589/pgr.v5i3.440.
[3] E. Ruijer, S. Grimmelikhuijsen, and A. Meijer, “Open Data for Democracy: Developing a Theoretical Framework for Open Data Use,” *Gov. Inf. Q.*, 2017, doi: 10.1016/j.giq.2017.01.001.
[4] A. Islam, “Data Governance and Compliance in Cloud-Based Big Data Analytics: A Database-Centric Review,” *Ajsteme*, 2024, doi: 10.69593/ajieet.v1i01.122.
[5] J. Baijens, T. Huygh, and R. Helms, “Establishing and Theorising Data Analytics Governance: A Descriptive Framework and a VSM-based View,” *J. Bus. Anal.*, 2021, doi: 10.1080/2573234x.2021.1955021.
[6] M. A. Hossin, J. Du, L. Mu, and I. O. Asante, “Big Data-Driven Public Policy Decisions: Transformation Toward Smart Governance,” *Sage Open*, 2023, doi: 10.1177/21582440231215123.
[7] F. Cronemberger and J. R. Gil-García, “Characterizing Stewardship and Stakeholder Inclusion in Data Analytics Efforts: The Collaborative Approach of Kansas City, Missouri,” *Transform. Gov. People Process Policy*, 2022, doi: 10.1108/tg-05-2022-0065.
[8] Y. Li and Y. Zhang, “Research on the Current Application of Big Data in Public Governance,” *Adv. Econ. Manag. Polit. Sci.*, 2024, doi: 10.54254/2754-1169/80/20241808.
[9] Y. He, “Public Health Governance Policy Optimization Based on Health Big Data,” *Acad. J. Manag. Soc. Sci.*, 2023, doi: 10.54097/u30h8z3l.
[10] A. Suominen and A. Hajikhani, “Research Themes in Big Data Analytics for Policymaking: Insights From a Mixed-methods Systematic Literature Review,” *Policy Internet*, 2021, doi: 10.1002/poi3.258.
[11] I. Pencheva, M. Estève, and S. Mikhaylov, “Big Data and AI – A Transformational Shift for Government: So, What Next for Research?,” *Public Policy Adm.*, 2018, doi: 10.1177/0952076718780537.
[12] A. Okuyucu and N. Yavuz, “Big Data Maturity Models for the Public Sector: A Review of State and Organizational Level Models,” *Transform. Gov. People Process Policy*, 2020, doi: 10.1108/tg-09-2019-0085.
[13] S. Huang, “Strategies for the Application of Big Data in Preventive Medicine in the Field of Public Health,” *Acad. J. Sci. Technol.*, 2024, doi: 10.54097/a4ekws45.
[14] E. Supriyanto and J. Saputra, “Big Data and Artificial Intelligence in Policy Making: A Mini-Review Approach,” *Int. J. Adv. Soc. Sci. Humanit.*, 2022, doi: 10.56225/ijassh.v1i2.40.
[15] S. Giest, “Big Data for Policymaking: Fad or Fasttrack?,” *Policy Sci.*, 2017, doi: 10.1007/s11077-017-9293-1.
[16] B. Levy *et al.*, “The Role of Analytics Governance to Promote Health Care Transformation,” *Aci Open*, 2022, doi: 10.1055/s-0042-1744383.
[17] A. Galán-Sanantonio, M. Botija, and Á. Carbonell, “A Systematic Integrative Review of International Policies for Women Experiencing Homelessness,” *Int. J. Soc. Welf.*, 2025, doi: 10.1111/ijsw.70005.
[18] J. Song, Y. Deng, Y. Yang, L. P. Gleason, and A. Kho, “Change in Address in Electronic Health Records as an Early Marker of Homelessness,” *Plos One*, 2025, doi: 10.1371/journal.pone.0318552.
[19] J. Tsai, C. Y. See Lee, T. Byrne, R. H. Pietrzak, and S. M. Southwick, “Changes in Public Attitudes and Perceptions About Homelessness Between 1990 and 2016,” *Am. J. Community Psychol.*, 2017, doi: 10.1002/ajcp.12198.
[20] J. Tsai, C. Y. S. Lee, J. Shen, S. M. Southwick, and R. H. Pietrzak, “Public Exposure and Attitudes About Homelessness,” *J. Community Psychol.*, 2018, doi: 10.1002/jcop.22100.
[21] C. Parsell, A. Clarke, and E. Kuskoff, “Understanding Responses to

Homelessness During COVID-19: An Examination of Australia," *Hous. Stud.*, 2020, doi: 10.1080/02673037.2020.1829564.

[22] L. S. Reeves, A. Clarke, E. Kuskoff, and C. Parsell, "Fulfilling and Desperately Needed: Australian Media Representations of Responses to Homelessness," *Aust. J. Soc. Issues*, 2022, doi: 10.1002/ajs4.201.

[23] E. J. Edwards, "Who Are the Homeless? Centering Anti-Black Racism and the Consequences of Colorblind Homeless Policies," *Soc. Sci.*, 2021, doi: 10.3390/socsci10090340.

[24] J. Olivet *et al.*, "Racial Inequity and Homelessness: Findings From the SPARC Study," *Ann. Am. Acad. Pol. Soc. Sci.*, 2021, doi: 10.1177/0002716221991040.

[25] A. Shah, D. Shah, D. Shah, D. Chordiya, N. Doshi, and R. Dwivedi, "Blood Bank Management and Inventory Control Database Management System," *Procedia Comput. Sci.*, vol. 198, pp. 404–409, 2021, doi: 10.1016/j.procs.2021.12.261.

[26] H. A. Shah, "Mapping Loneliness Through Comparative Analysis of USA and India Using Social Intelligence Analysis," *BMC Public Health*, vol. 24, no. 1, 2024, doi: 10.1186/s12889-023-17630-3.

[27] N. Novita and A. I. Nanda Anissa, "The Role of Data Analytics for Detecting Indications of Fraud in the Public Sector," *Int. J. Res. Bus. Soc. Sci. 2147-4478*, 2022, doi: 10.20525/ijrbs.v11i7.2113.

[28] I. Lauria, T. E. McEwan, S. Luebbers, M. Simmons, and J. R. P. Ogloff, "Evaluating the Ontario Domestic Assault Risk Assessment in an Australian Frontline Police Setting," *Crim. Justice Behav.*, 2017, doi: 10.1177/0093854817738280.

[29] J. Verrey, B. Ariel, V. Harinam, and L. Dillon, "Using Machine Learning to Forecast Domestic Homicide via Police Data and Super Learning," *Sci. Rep.*, 2023, doi: 10.1038/s41598-023-50274-2.

[30] D. Schiff, K. J. Schiff, and P. Pierson, "Assessing Public Value Failure in Government Adoption of <scp>artificial Intelligence</Scp>," *Public Adm.*, 2021, doi: 10.1111/padm.12742.

[31] D. Rogger and C. Schuster, "How Scholars Can Support Government Analytics: Combining Employee Surveys With More Administrative Data Sources Towards a Better Understanding of How Government Functions," *Public Adm. Rev.*, 2024, doi: 10.1111/puar.13894.

[32] Н. Шевченко, О. Марухленко, O. Trach, P. Shvedenko, and O. Dubovych, "The Use of Data Analytics in Public Administration for Corruption Prevention During Hybrid Warfare," *PJC*, 2024, doi: 10.62271/pjc.16.2.943.958.

[33] D.-S. Branet and C.-D. Haţegan, "Bibliometric Framing of Research Trends Regarding Public Sector Auditing to Fight Corruption and Prevent Fraud," *J. Risk Financ. Manag.*, 2024, doi: 10.3390/jrfm17030094.

[34] T. Molobela and D. E. Uwizeyimana, "E-Governance as a New Public Administration Paradigm: A Rhetoric or Reality?," *Int. J. Res. Bus. Soc. Sci. 2147-4478*, 2023, doi: 10.20525/ijrbs.v12i8.2931.

[35] A. Simonofski, T. Tombal, C. de Terwangne, P. Willem, B. Frénay, and M. Janssen, "Balancing Fraud Analytics With Legal Requirements: Governance Practices and Trade-Offs in Public Administrations," *Data Policy*, 2022, doi: 10.1017/dap.2022.6.

[36] L. Judijanto, T. Taufiqurokhman, S. A. Hendrawan, and H. Herwanto, "Strategies for Utilizing AI and Data Analytics to Improve the Effectiveness of Public Services in Indonesia: A Local Government Level Approach," *West Sci. Bus. Manag.*, 2023, doi: 10.58812/wsbm.v1i05.470.

[37] H. Broomfield and L. Reutter, "Towards a Data-Driven Public Administration: An Empirical Analysis of Nascent Phase Implementation," *Scand. J. Public Adm.*, 2021, doi: 10.58235/sjpa.v25i2.7117.

[38] Y. Kalnysh, "Logic and Methodology of Problem Analysis in Public Administration," *Věda Perspekt.*, 2021, doi: 10.52058/2695-1584-2021-2(2)-7-16.

[39] H. P. Faga, "The Implications of Transnational Cyber Threats in International Humanitarian Law: Analysing the Distinction Between Cybercrime, Cyber Attack,

and Cyber Warfare in the 21stCentury," *Balt. J. Law Polit.*, 2017, doi: 10.1515/bjlp-2017-0001.

[40] A. Coco and T. Dias, "'Cyber Due Diligence': A Patchwork of Protective Obligations in International Law," *Eur. J. Int. Law*, 2021, doi: 10.1093/ejil/chab056.

[41] S. Haataja, "Cyber Operations Against Critical Infrastructure Under Norms of Responsible State Behaviour and International Law," *Int. J. Law Inf. Technol.*, 2022, doi: 10.1093/ijlit/eaad006.