UDC 001.51
IRSTI 28.23.29

https://doi.org/10.55452/1998-6688-2025-22-2-94-109

[1]**Talgatuly Zh.,**
Researcher, ORCID ID: 0009-0004-1537-7975,
e-mail: zforzesteam@gmail.com
[1]***Amirgaliyev B.Ye.,**
PhD, Professor, ORCID ID: 0000-0003-0355-5856,
*e-mail: beibut.amirgaliyev@astanait.edu.kz
[1]**Yedilkhan D.,**
PhD, Associate Professor, ORCID ID: 0000-0002-6343-5277,
e-mail: d.yedilkhan@astanait.edu.kz
[1]**Turginbekov A.,**
Researcher, ORCID ID: 0000-0002-2742-6779,
e-mail: turginbekovalmaz@gmail.com
[2]**Gadaborshev Kh. S.,**
MBA, ORCID ID: 0009-0001-3868-3594,
e-mail: khavazh@rmggroup.kz

[1]Astana IT University, Astana, Kazakhstan
[2]Almaty Management University, Almaty, Kazakhstan

CLUSTERING-BASED METHODS FOR DATA-DRIVEN OPTIMIZATION IN URBAN COURIER LOGISTICS

Abstract

With the rapid development of cities and their infrastructure, the demand for high-quality urban deliveries is increasing at the same rate. This work explores the possibilities of dynamically allocating delivery zones for courier deliveries based on data provided by the courier company. Traditional manually created delivery zones often do not ensure that the picture is relevant to the real situation in the city (weather, traffic, roads, etc.). This study presents the results of how K-Means and DBSCAN clustering algorithms can contribute to the dynamic distribution of delivery zones in clusters. The comparative analysis includes consideration of such indicators as Silhouette value and computational complexity of Big-O Notation. The results show that the K-Means algorithm creates structured and uniform clusters, while DBSCAN shows results in defining flexible clusters based on the density of data in the region. Multi-level DBSCAN provides an opportunity to reduce the concentration of "noise", thereby increasing the coverage of all delivery points. The results obtained highlight the advantages of using clustering algorithms in creating dynamic delivery zones to improve the distribution of orders between couriers and reduce operating costs. Further research should include obtaining continuous real-time data flow to monitor the operation of algorithms in a dynamic environment.

Keywords: urban logistics; clustering algorithms; courier optimization; smart city; delivery zones; machine learning.

Introduction

Sustainable city growth often requires adapting infrastructure and resources with rapid population growth and urbanization. The "Smart city" concept addresses this issue, providing the use of advanced technologies, data analysis, and industrial automation to improve life quality, environmental sustainability, and safety of citizens [1]. The work of couriers in an urban environment is one of the key elements of a modern logistics system, especially in growing demand for fast and efficient delivery of goods and services [2]. Research shows that the use of big data analytics and machine learning algorithms to model optimal routes can reduce delivery time, reduce fuel costs, and increase customer satisfaction [3]. It's worth mentioning that the COVID-19 pandemic has highlighted the critical role of resilient logistics systems in ensuring the continuous supply of goods, especially in urban areas. With the shift towards online shopping and the increasing demand for contactless delivery options, courier services must adapt by adopting more efficient and automated systems [4–6]. This research addresses these needs by exploring innovative solutions for optimizing delivery operations using spatial algorithms, data clustering, and automated courier assignments.

The primary goal of this research is to develop and evaluate data-driven methods and models that enhance the efficiency of courier services within a smart city framework. The research object is the optimization of courier delivery systems in urban environments, specifically within the smart city ecosystem. The novelty of this work lies in integrating automated clustering algorithms with spatial computing techniques to optimize delivery zones in real-time. Unlike traditional manual zoning approaches, which are labor-intensive, this study applies a dynamic approach forming clusters of urban zones based on data-driven analysis.

Review of existing studies includes systematic approach and selects specific works related to the subject of the study. The main source of information was peer-reviewed journals and conferences available on reputable platforms such as IEEE Xplore, SpringerLink, ScienceDirect and Web of Science. The search period was limited to publications from 2010 to 2024, which allows us to take into account current trends and achievements in the field of optimization of logistics processes and the introduction of smart city technologies. The literature selection process included several stages. At the first stage, publications were searched for keywords: "courier", "transportation", "smart city", "urban delivery", "optimization", "clustering", "logistics", "routing", "data analysis", "machine learning". At the second stage, filters were applied that limit the results by date of publication (since 2010) and type of documents (articles from peer-reviewed journals and conference proceedings). The following criteria were used to select relevant publications: topic relevance with courier deliveries and logistic optimization, the use of modern data analysis frameworks and application of machine learning algorithms, as well as the practical focus of research on improving transport processes in smart city ecosystems. The selected articles were analyzed for the availability of methodological approaches applicable to optimize the work of courier services and to consider innovative solutions in the field of logistics.

Optimizing logistics hub locations using a hybrid approach that combines K-Means clustering and the P-Median model presents a significant advancement in addressing the inefficiencies of last-mile delivery. Prioritizing road network distances over traditional Euclidean measures provides a more accurate representation of real-world travel conditions, leading to a measurable reduction in delivery distances and associated costs [7]. The integration of delivery volumes and population density enhances the model's adaptability, making it suitable for diverse urban logistics scenarios. The process effectively refines cluster formations through iterative optimization and road network-based calculations, ensuring minimal travel distances and improved operational efficiency. Despite these strengths, the reliance on simulated data and static optimization processes limits the approach's ability to account for the dynamic nature of real-world logistics challenges. Additionally, computational intensity and dependence on accurate road network data raise concerns about scalability in larger, more complex urban environments. These considerations highlight both the potential and the areas for improvement in applying road network-based optimization to urban logistics systems. Building on the exploration of alternative delivery mechanisms, the optimization of logistics hub locations plays a crucial role in addressing last-mile delivery inefficiencies. By strategically situating hubs based on real-world travel conditions, these systems can enhance travel distances reduction, air quality improvement and operational cost optimizations. Therefore, it might bring significant contribution to public transport integration into delivery logistics. This approach, as demonstrated through advanced clustering and optimization techniques, highlights the importance of leveraging data-driven methodologies to enhance urban logistics efficiency.

Optimizing courier deliveries in urban environments has become an urgent task considering the rapid growth of e-commerce and related problems such as increased traffic jams, emissions and the burden on transport infrastructure [8, 9]. A study examines the possibility of using public transport systems to deliver goods [10]. One of the key ideas is to use the spare capacity of public transport, such as buses, trams and trains, to transport parcels during low-load hours. This brings an opportunity to reduce emissions of pollutants by reusing public transport for delivery purposes. However, the applicability of the proposed approach strongly depends on the level of development of the public transport system in a particular city. In cities with underdeveloped infrastructure or irregular public transport, the effectiveness of the proposed model may decrease significantly. It is also worth noting that the vehicles used to transport passengers have limited space for placing parcels, which may limit the delivery volume during peak hours when public transport is already loaded with passengers. Its implementation may require significant resources to restructure logistics, which may not be suitable for most companies and cities. A study identifies the critical role of applying time series data to optimize demand forecasting in delivery terms [11]. Historical data brings significant contribution to the accuracy improvement of forecasting models. In study analysis, classic time-series models like exponential smoothing and ARIMA perform optimally when a relatively large amount of data (at least two months) is provided. The study also highlights an application of machine learning algorithms such as Random Forest and Support Vector Regression for conditions with time-limited data. These algorithms demonstrated great performance in such cases as they can effectively identify patterns in a limited amount of input (data). The accuracy of forecasting models is highly dependent on data quality. Whenever there is an insufficient amount of data, most forecasting algorithms can degrade significantly. This creates challenges to generate reliable forecasts for delivery demands in regions with limited data. Changes in consumer behavior due to holidays, promotions, or unexpected events (like weather changes) can introduce irregularities that the models may not be able to capture accurately without further tuning.

In recent years, innovative last-mile delivery models have gained significant attention as cities seek to enhance efficiency, reduce costs, and minimize the environmental impact of urban logistics. Among these models, crowdshipping has emerged as a promising solution. Crowdshipping leverages the concept of sharing economy, utilizing the unused capacity of individuals' vehicles or their existing travel routes to deliver packages. A study presents a comprehensive study on optimizing last-mile delivery using a multi-criteria approach that integrates automated smart lockers, capillary distribution, and crowdshipping [12]. The study emphasizes that crowdshipping can significantly enhance the efficiency of urban logistics by tapping into a network of volunteers who deliver packages to smart lockers or directly to customers. This approach optimizes delivery times and costs making it flexible within logistic processes. However, the study also identifies limitations associated with crowdshipping, particularly in terms of coordination and reliability. While crowdshipping reduces delivery costs and leverages underutilized resources, it introduces complexities in managing a network of decentralized delivery agents. Ensuring timely deliveries and maintaining service quality can be challenging due to the variability in participants' availability and commitment. Incorporating these methods into urban logistics systems presents opportunities for cities to enhance the efficiency of last-mile deliveries. It is also worth noting that the reliance on crowd participation and the infrastructure required for automated lockers highlight the need for further research and investment to scale these models effectively.

Materials and Methods

The methodology used in this study is both exploratory and explanatory. The exploratory aspect involves investigating the potential of spatial algorithms and optimization techniques to enhance courier delivery efficiency. Meanwhile, the explanatory component focuses on understanding how these methods impact the logistics processes within predefined urban delivery zones. The research design is cross-sectional, capturing data and insights at a specific time to provide a comprehensive analysis of the current delivery system and optimization methods.

To identify the most effective clustering algorithms for optimizing courier delivery zones, the study employed a dual evaluation framework: qualitative analysis, focusing on conceptual suitability and practical strengths, and quantitative analysis, emphasizing performance metrics and computational efficiency.

The methodology evaluates algorithms utilizing Big O notation to assess performance at the computational expense level.

K-Means Clustering. K-Means was chosen as a baseline method due to its simplicity and efficiency in handling large datasets [13]. Its ability to partition data into a fixed number of clusters makes it well-suited for scenarios where delivery zones are expected to maintain consistent sizes and densities as shown in Equation 1.

$$WCSS = \sum_{i=1}^{k} \sum_{x \in C_i} \left|\left| x - \mu_i \right|\right|^2 \tag{1}$$

where C_i represents the cluster of iteration i-th, x is a data (geographical point) and μ_i is the centroid of a cluster (C_i). The algorithm advances thoroughly until centroids show no large difference between iterations, indicating that data points within a cluster have been collected into a cluster. In terms of computational expenses, the algorithm demonstrates the time complexity (O notation) in Equation 2 [13].

$$O(InKd) \tag{2}$$

where $O(nKd)$ evaluates computing distances for all n points in d-dimensional space, $O(nd)$ computing new centroids and I indicates total complexity (iterations).

DBSCAN Clustering. DBSCAN (Density-Based Spatial Clustering of Applications with Noise) was selected for its capacity to identify clusters of varying densities and exclude noise, making it ideal for urban settings with irregular demand patterns [14]. Noise points (points without clusters) are an essential part of the algorithm that determine outsider points that are left without clusters. It depends on the minimum neighbor amount and distance between neighbor parameters. To evaluate algorithm's performance, a noise points percentage is calculated as shown in Equation 3:

$$Noise\ Percentage = \left(\frac{Noise\ Points}{Total\ Points}\right) \times 100 \tag{3}$$

The algorithm is highly dependent on the following parameters:

1) ε (Epsilon radius) maximum distance between two points to be considered neighbors. It is the radius within which the algorithm counts neighbors, so it determines what is "close enough." If ε is set too small, even natural neighborhoods look empty and most points become noise; if it is too large, distinct zones blur together into one oversized cluster;

2) $MinPts$ (Minimum Points) minimum number of points required to form a dense region. Low $MinPts$ makes it easy for random alignments of addresses to masquerade as clusters, while a high $MinPts$ filters out those coincidences but can also hide legitimately small zones.

For given x (data point), the epsilon radius is defined as shown in Equation 4:

$$N_\varepsilon(x) = \{y \in X \mid d(x,y) \leq \varepsilon\} \tag{4}$$

where $d(x, y)$ is typically the euclidean distance, used for determining distance between data points (addresses). The group of data points is considered as a cluster by satisfying and classifying [15].

1) If it has at least $MinPts$ points (including itself) within its ε-epsilon radius (core point) (Equation 5).

$$|\ N_\varepsilon(x)\ | \geq MinPts \tag{5}$$

2) Lies within the ε-epsilon radius of a core point but does not have enough neighbors to be a core point (border point) (Equation 6).

$$N_\varepsilon(x) < MinPts \tag{6}$$

3) Neither a core nor a border point (outlier, noise) (Equation 7).

$$N_\varepsilon(x) < 1 \tag{7}$$

Overall time complexity fits within (Equation 8):

$$O(nlog_n) \tag{8}$$

where n stands for total dataset size.

Evaluation of clustering algorithms performance is conducted based on the Silhuotte Score (Equation 9). It is a metric used to evaluate the quality of clusters in a dataset, it provides a useful measure for evaluating clustering results, but its interpretation should consider the data's dimensionality and the shapes of the clusters [16, 17].

$$s(i) = \frac{b(i) - a(i)}{max(a(i), b(i))} \tag{9}$$

where $a(i)$ is distance to points in the same cluster (cohesion within cluster) and $b(i)$ a distance to the closest different cluster (cohesion of clusters). Lower $a(i)$ means better cohesion. Used alongside the Elbow Method to determine the optimal number of clusters (k) in K-Means clustering [18, 19]. It is based on the inertia (sum of squared distances) between data points and their assigned cluster centroids. Inertia is calculated using the same equation (1), and results can be interpreted in the following rules: 1) low WCSS, clusters are more compact; 2) high WCSS, clusters are spread out. The "Elbow" is where the rate of decrease in WCSS(k) slows down, meaning adding more clusters does not significantly improve clustering quality (Equation 10).

$$\Delta WCSS(k) = WCSS(k-1) - WCSS(k) \tag{10}$$

compute the second derivative to find the point where the curve flattens (Equation 11):

$$\Delta^2 WCSS(k) = \Delta WCSS(k-1) - \Delta WCSS(k) \tag{11}$$

The elbow point occurs at the value of k where the second derivative is maximized.

DBSCAN utilizes a different approach in evaluating Silhouette Score of clusters (Equation 12) [20], because: 1) it labels some points as noise (-1); 2) noise points must be excluded from the score computation.

$$S_{DBSCAN} = \frac{1}{N_{clustered}} \sum_{i \in clustered\ points} s(i) \tag{12}$$

where $N_{clustered}$ is the number of non-noise points (belonging to a cluster).

The Silhouette Score is interpreted in range of values (Table 1):

Table 1 – Silhouette score value classification

Silhouette Score value	Clustering quality
Near 1.0	Well-separated, compact clusters
0.5 - 0.7	Good clustering, but some overlap
0.2 - 0.4	Moderate clustering
Near 0.0	Poor separation, clusters are overlapping
Negative (s < 0)	Clusters may be incorrectly assigned

A study of clustering algorithm performance metrics on blood donors indicate that K-Means clusterization is most effective in cases where the data is well-separated and follows a spherical distribution (which fits well with address data points) [21]. The only limitations are strong reliance on initial centroid placement and mandatory specification of number of the clusters (K). It is worth mentioning that Hierarchical Clustering proves to be great for understanding correlation at different layers of data clusters. However, its computational requirements and sensitivity to data noise might be primary reasons for high maintenance costs. DBSCAN is well-suited for handling datasets with varying densities and noise, making it an excellent option for complex and irregular data distributions. However, its performance is highly dependent on the careful selection of parameters (ϵ and minPts), which, if not properly tuned, can lead to can lead to less effective grouping of data points.

The dataset used in this study comprises several key features, including geographical information (latitude and longitude) representing receiver locations, along with additional attributes such as delivery details and timestamps. Each record corresponds to a unique geographical point within an urban region, ensuring the dataset's relevance to practical applications like delivery zone optimization and courier assignment. The dataset went through preprocessing to address missing or invalid entries, resulting in a clean and structured format suitable for analysis. The data captures deliveries over a 2-month period, spanning from 30 October 2024 to 31 December 2024 resulting in ~900 delivery items with unique receiver addresses. Such a short period is due to the fact that the company's integration with geographic address detection systems (2GIS) falls at the end of October (Figure 1).

Given the proprietary nature of the data and processes used in this research, the majority of the content falls under a Non-Disclosure Agreement (NDA) with the company. Consequently, detailed information related to company-specific data cannot be disclosed. All research activities, including the collection and analysis of data, were conducted in compliance with confidentiality agreements and were fully approved by the company's management.

rder_item_create_dat	status	process	actual_delivery_date	ender_city	nder_stre	nder_build	ender_geo	eceiver_cit	ceiver_stre	eceiver_buildin	eceiver_geo
2024-11-13 12:07:05.	DELIVERED_ACCEPTE	ALSECO	2024-11-28 13:27:1	Алматы				Алматы	Байтурсы	58	{"lat":43.246555,"lon":76.927097}
2024-11-13 12:07:05.	DELIVERED_ACCEPTE	ALSECO	2024-11-28 13:27:1	Алматы				Алматы	Абая	79	{"lat":43.241582,"lon":76.928503}
2024-11-13 12:07:05.	DELIVERED_ACCEPTE	ALSECO	2024-11-28 13:27:3	Алматы				Алматы	Абая	89	{"lat":43.241301,"lon":76.92469}
2024-08-28 13:47:22.	DELIVERED_ACCEPTE	IN_CITY		Алматы	Бекмахан	2/6		Алматы	Алматы, у	38A	
2024-11-13 12:07:06.	DELIVERED_ACCEPTE	ALSECO	2024-11-22 06:09:5	Алматы				Алматы	Абылай х	121	{"lat":43.249238,"lon":76.941672}
2024-11-13 12:07:05.	DELIVERED_ACCEPTE	ALSECO	2024-11-22 07:00:17	Алматы				Алматы	Желтокса	144	{"lat":43.247941,"lon":76.940904}
2024-11-01 06:52:16.	DELIVERED_ACCEPTE	IN_CITY	2024-11-12 17:30:07	Алматы	Бекмахан	2/6		Алматы	Ратушно	78/1	
2024-12-13 07:54:39.	DELIVERED_ACCEPTE	ALSECO	2024-12-15 08:04:3	Алматы				Алматы	Шевченк	48	{"lat":43.246241,"lon":76.947116}
2024-08-28 10:38:46.	DELIVERED_ACCEPTE	IN_CITY	2024-10-31 09:14:21	Алматы	Бекмахан	2/6		Алматы	Жумабае	176/11	
2024-12-11 17:44:56.	DELIVERED_TO_WAR	IN_CITY		Алматы	Бекмахан	2/6		Алматы	Байзаков	225	{"lat":43.246877,"lon":76.911999}
2024-11-13 12:07:05.	DELIVERED_ACCEPTE	ALSECO	2024-11-22 15:09:03	Алматы				Алматы	Курманга	113	{"lat":43.243475,"lon":76.92553}
2024-12-11 17:45:42.	DELIVERED_TO_WAR	IN_CITY		Алматы	Бекмахан	2/6		Алматы	Ауэзова	33/35	{"lat":43.24729,"lon":76.902423}
2024-12-02 09:20:43.	DELIVERED_ACCEPTE	IN_CITY	2024-12-06 04:25:17	Алматы	Бекмахан	2/6		Алматы	микрорай	17Б	{"lat":43.233137,"lon":76.83557}
2024-12-13 08:27:43.	READY_FOR_DELIVER	IN_CITY		Алматы	Абылай х	53	{"lat":43.2	Алматы	БЦ Кокте		
2024-12-20 04:40:45.	READY_FOR_PICKUP	IN_CITY		Алматы	Уалихано	137/1	{"lat":43.2	Алматы	Суюнбай		
2024-12-17 11:30:43.	READY_FOR_DEPART	INTER_CITY		Алматы	Абылай х	53	{"lat":43.2	Рудный	Павла Ко	117	{"lat":52.96797,"lon":63.089711}
2024-12-16 09:25:57.	READY_FOR_DELIVER	IN_CITY		Алматы	Земнухов	35	{"lat":43.3	Алматы	микрорай	347/2	{"lat":43.206329,"lon":76.788501}
2024-10-25 16:04:24.	DELIVERED_ACCEPTE	IN_CITY	2024-10-31 09:11:03	Алматы	Бекмахан	2/6		Алматы	Жарокова	230	
2024-10-25 16:05:57.	DELIVERED_ACCEPTE	IN_CITY	2024-10-31 09:10:53	Алматы	Бекмахан	2/6		Алматы	Жансугур	462	
2024-10-25 16:06:23.	DELIVERED_ACCEPTE	IN_CITY	2024-10-31 09:10:44	Алматы	Бекмахан	2/6		Алматы	Ратушно	78/1	
2024-12-06 09:02:12.	READY_FOR_DELIVER	IN_CITY		Алматы	Абылай Х	53		Алматы	Ратушно	78/1	{"lat":43.231467,"lon":76.886416}
2024-12-14 08:02:06.	DELIVERY_COURIER_	IN_CITY		Алматы	Суюнбая	152г	{"lat":43.3	Алматы	Мынбаев	151	{"lat":43.236004,"lon":76.881615}
2024-12-20 04:26:28.	READY_FOR_PICKUP	IN_CITY		Алматы	Уалихано	137/1	{"lat":43.2	Алматы	Ауэзова		
2024-11-06 13:29:00.	DELIVERED_ACCEPTE	IN_CITY	2024-12-02 08:42:37	Алматы	Бекмахан	2/6		Алматы	Мкр Таст	13	
2024-11-06 13:29:36.	DELIVERED_ACCEPTE	IN_CITY	2024-12-02 08:42:2	Алматы	Бекмахан	2/6		Алматы	Ратушно	78/1	
2024-11-30 04:08:28.	DELIVERED_ACCEPTE	IN_CITY	2024-12-02 09:15:0	Алматы	Бекмахан	2/6		Алматы	Ратушно	78/1	
2024-11-22 18:14:27.	DELIVERED_ACCEPTE	IN_CITY	2024-12-02 09:15:1	Алматы	Бекмахан	2/6		Алматы	Гагарина	244A	{"lat":43.210566,"lon":76.898933}
2024-11-22 18:13:51.	DELIVERED_ACCEPTE	IN_CITY	2024-12-02 09:15:27	Алматы	Бекмахан	2/6		Алматы	микрорай	45	{"lat":43.213729,"lon":76.88116}
2024-11-19 15:44:35.	DELIVERED_ACCEPTE	IN_CITY	2024-12-02 09:17:04	Алматы	Бекмахан	2/6		Алматы	Курманга	104	{"lat":43.242997,"lon":76.92632}
2024-11-19 15:41:43.	DELIVERED_ACCEPTE	IN_CITY	2024-12-02 09:17:1	Алматы	Бекмахан	2/6		Алматы	Аль-Фара	116/4	{"lat":43.213632,"lon":76.941946}

Figure 1 – Dataset content (preview)

Results and Discussion

The company owning the dataset, described previously, utilizes their own delivery zones that are managed manually and rarely updated (Figure 2). From now on, these delivery zones are referred to as static delivery zones.

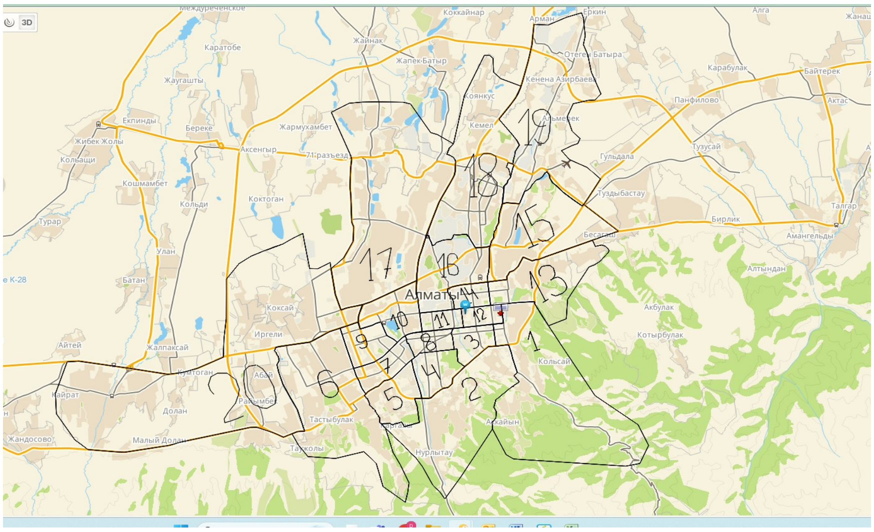

Figure 2 – Company's traditional static delivery zone design

The static delivery zones within Almaty are designed based on administrative and geographical boundaries. There are 19 zones numbered and distributed across the city, reflecting population density, urban infrastructure, and delivery demand. Key observations about these zones are provided further.

Major roads and highways traverse through the zones, ensuring connectivity and accessibility for couriers. Zones 5 and 18 are adjacent to primary transportation routes, enabling efficient movement. The static nature of these zones does not account for dynamic changes in delivery demands, such as seasonal fluctuations or time-specific orders. Additionally, irregular zone shapes and varying sizes can lead to inefficiencies in courier allocation and route optimization. To understand the real picture of delivery demands, Kernel Density Estimation (KDE) heatmap is applied to get deeper insights (Figure 3).

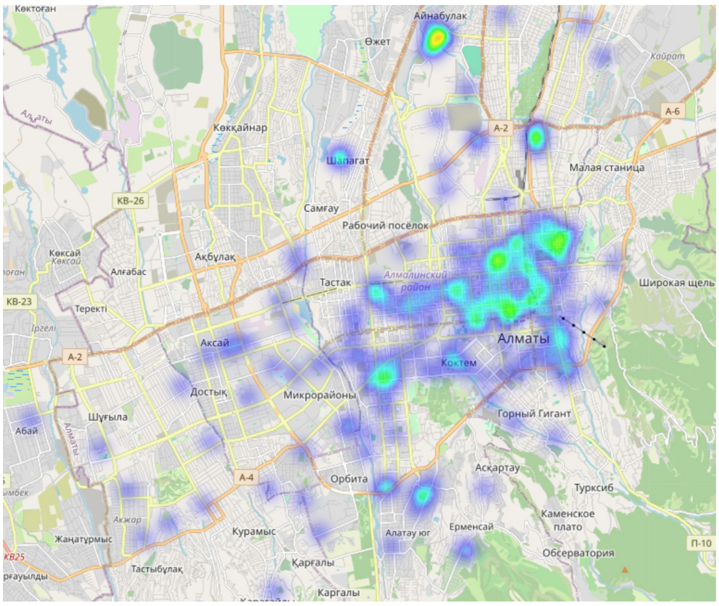

Figure 3 – Heatmap visualization of delivery address distribution

This heatmap shows insights into delivery demands based on the data provided. Combination of static zones and heatmap markers reveals the relevance of each static delivery zone in relation to real data. Only a few zones have an increased load (color saturation and volume) (Figure 4).

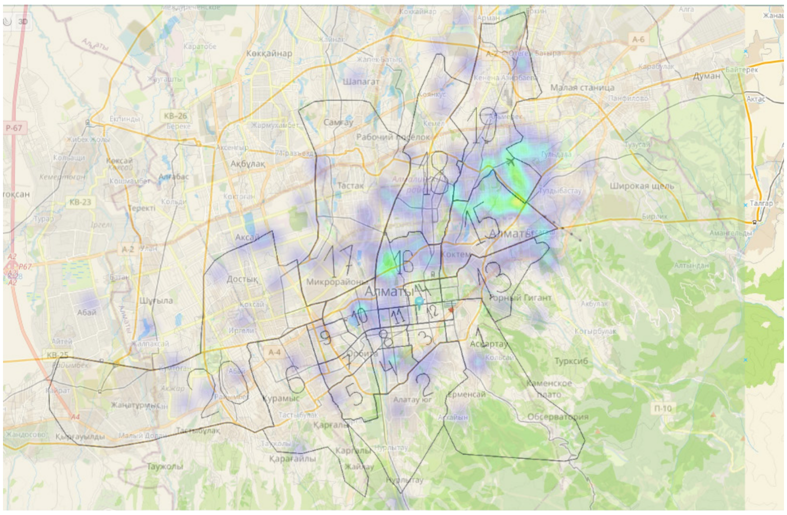

Figure 4 – Static zone map and heatmap figures merged

The visualization reveals critical insights for resource allocation, enabling courier services to distribute resources more effectively in high-demand zones.

K-Means powered delivery zones (clusters) are more uniform than the static zones, reflecting the algorithm's preference for even distribution (Figure 5). However, the shapes of the zones are geometric and may not perfectly align with natural urban boundaries.

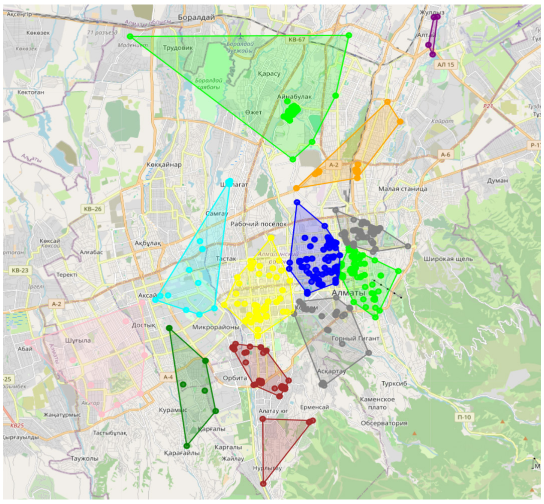

Figure 5 – K-Means clusterization of delivery address data

The K-Means clustering algorithm was applied to segment delivery points into a predefined number of clusters, offering a structured and computationally efficient approach to identifying delivery zones. Several clusters (k=20) were chosen since the original static zones contain 19 clusters (manually designed). The choice is picked by considering business and technical perspectives.

The company that possesses the dataset and utilizes the static zones has a small group of operating couriers (12 to 15). A slightly higher number of clusters brings flexibility and fair workload distribution among couriers.

Original static zones have 19 manually designed clusters. Using ensures that clustering methods align closely with the existing operational structure. It also provides a good balance between having too many small clusters (inefficient) and too few large clusters (long delivery times).

Dynamic distribution of delivery zones using the algorithm results in a form of zones designed in compact and evenly distributed across the map. Although there is an exception of empty areas that are not marked. This approach can help to evenly distribute work among couriers, since each courier will have approximately the same area in size. Compared to static areas on the left side of Figure 6, where one courier may have a significantly larger area, and another may have a relatively smaller one (Figure 6).

Static delivery zones **K-Means based zones**

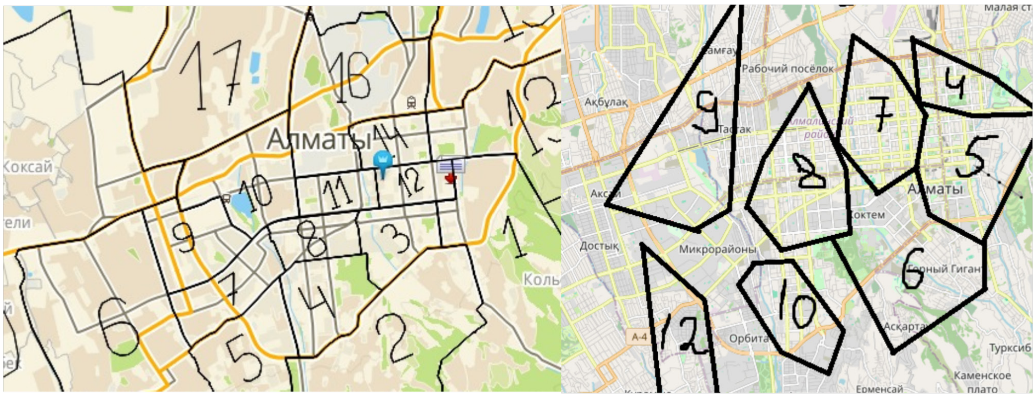

Figure 6 – Close look on static and algorithm generated zones

Future iterations could integrate dynamic adjustments to the value of k based on real-time data, such as traffic patterns or seasonal demand fluctuations, to make K-Means clustering more adaptable to the complexities of urban logistics.

Silhouette Score remains close to 1 when cluster , indicating that clusters are well-formed and clearly separated. This suggests that fewer, larger clusters maintain strong cohesion. There is a significant decline in the score after , suggesting that increasing the number of clusters beyond this point reduces cohesion. Thus, based on dataset size, clusters may lead to inefficient segmentation, with overlapping and less distinct clusters (Figure 7).

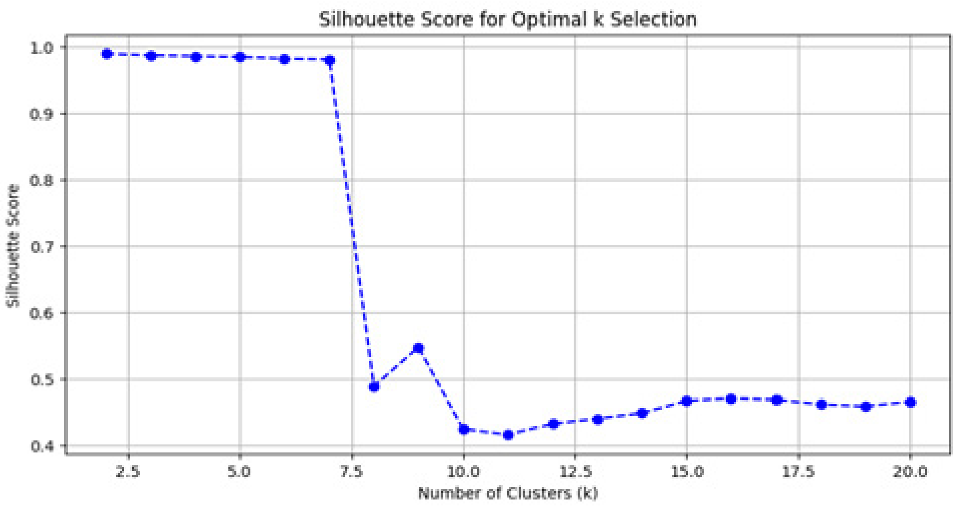

Figure 7 – Elbow method applied with Silhouette score

Unlike K-Means, DBSCAN does not require specific number of clusters. It makes it ideal for capturing anomalies and variations in delivery patterns. The output map showcases dynamic clusters generated by the algorithm, with each cluster represented by an unique color (Figure 8). The algorithm captures the natural structure of delivery zones, adapting to irregular geographical features and demand distributions. Basic DBSCAN results in a noise points percentage of 9.62%.

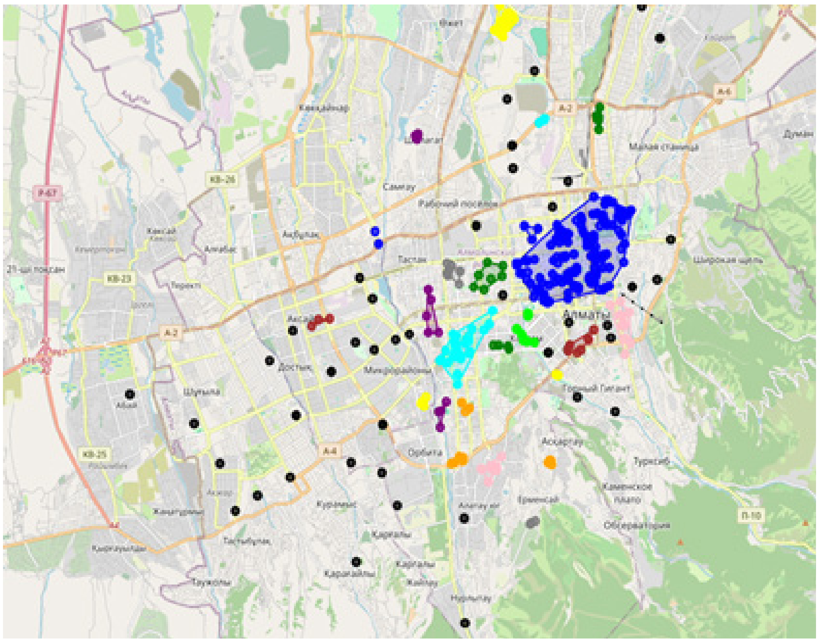

Figure 8 – DBSCAN density clustering with noise (grey dots)

Grey points that fall outside the defined density parameters are categorized as noise. These points typically correspond to isolated delivery locations that do not belong to any cluster.

Silhouette score evaluated 1-stage DBSCAN algorithm with moderate result output:

$$Silhouette\ Score\ for\ Single - Stage\ DBSCAN: 0.4229$$

Results are like the 3-stage DBSCAN approach and common interpretation of results are provided below.

To capture all grey points the algorithm was tuned to support a multi-stage clustering approach including carefully managed parameters to address the varying densities and noise levels (Figure 9). The primary clustering stage used small epsilon value $\varepsilon_{prim.} = 0.0035$ and minimum sample size $MinPts_{prim.} = 5$ to capture high-density delivery zones, such as those in the central urban areas. These strict parameters ensured compact clusters that accurately represented dense regions. In the secondary stage, noise points from the primary clustering were re-evaluated with lower epsilon $\varepsilon_{sec.} = 0.009$ and reduced minimum sample size $MinPts_{sec.} = 3$.

This allowed the algorithm to form clusters in medium-density areas while avoiding over merging. A tertiary stage was introduced with the largest epsilon $\varepsilon_{third.} = 0.035$ and keeping $MinPts_{third.} = 3$ to address remaining noise points in populated regions, ensuring comprehensive coverage of the dataset. The 3-Stage DBSCAN approach demonstrates a noise points percentage of 1.21%, which is almost 8 (7.9504) times lower than 1-Stage DBSCAN.

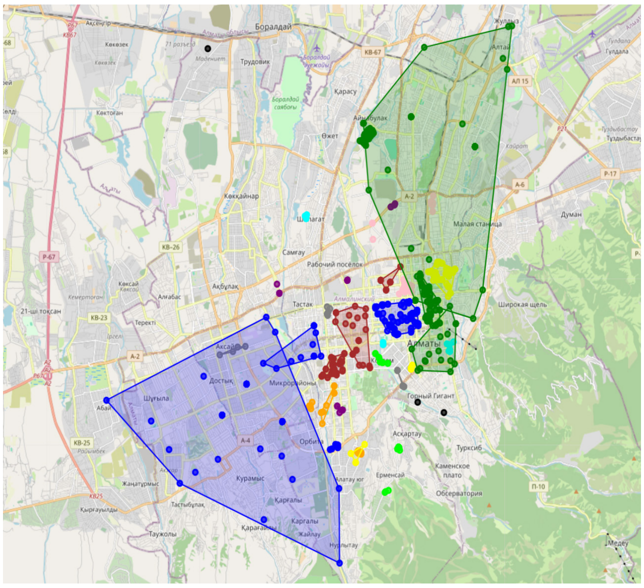

Figure 9 – DBSCAN 3-stage clusterization (primary, secondary and tertiary)

The multi-stage approach effectively allowed DBSCAN to adapt to urban, suburban, and peripheral delivery patterns. The approach ensured noise points were incorporated into clusters without precision degradation. The flexibility of DBSCAN in forming irregular cluster shapes captured real-world delivery zone layouts that static delivery zones could not provide. This method demonstrates how density-based clustering can optimize delivery zones, balancing the trade-off between precision in dense areas and inclusivity in regions. While the approach required careful settings of eps and min_samples parameters, its ability to handle noise and adapt to varying densities demonstrates its effectiveness in clustering geographical data points.

Silhouette score accessed 3-stage DBSCAN approach with moderate result:

$$Silhouette\ Score\ DBSCAN\ Clustering: 0.3613 \text{ (Moderate)}$$

The approach refines clusters in three steps: 1) identifies initial dense regions (clusters); 2) groups previously unclustered noise points; 3) attempts to cluster remaining noise. As a positive impact, more data points are assigned to clusters rather than being marked as noise (grey points). However, secondary and tertiary clusters may have less compact structures, leading to lower silhouette values (clusters overlapping).

Score 0.3613 is fine result for 3-stage DBSCAN but suggests that some clusters may be too loose or overlapping.

Performance comparison is based on Silhouette Score and Noise Points percentage (Table 2). K-Means has more effective clustering structure (optimal between $1 \leq k \leq 7$), whereas 1-Stage DBSCAN (0.4229) performs better than 3-Stage DBSCAN (0.3613), but still lower than K-Means, indicating that DBSCAN struggles with cluster separation.

Many clusters are algorithm-specific. K-Means is fixed to a specific number of clusters, providing a structured segmentation. DBSCAN has a variable number of clusters, as it determines clusters dynamically based on density.

K-Means has 0% noise level, it assigns every data point to cluster. 3-Stage DBSCAN algorithm is purposed to maximize the amount of data points binded to clusters, since grey points have low value in further application of data-driven approach. 1-Stage instance has a high noise rate (9.62%), meaning many points are left unclustered.

Table 2 – Performance metrics comparison

Category	K-Means	DBSCAN (1-Stage)	DBSCAN (3-Stage)
Silhouette score	$1 \leq k \leq 7$	0.4229	0.3613
Number of clusters	20	Varies	Varies
Noise points (%)	0%	9.62%	1.21%
Time elapsed (for 924 data points)	0.1384 seconds	0.0104 seconds	0.0375 seconds

In general, K-Means is the most structured approach, ensuring all points belong to clusters. DBSCAN (1-Stage) leaves too many points unclustered, making it less reliable. 3-Stage DBSCAN balances flexibility and data coverage, reducing noise while maintaining adaptive clustering.

K-Means take slightly longer time (~0.14 sec.) to capture all data points compared to DBSCAN instances (~0.01 sec. and 0.03 sec.). It is explained with the nature of K-Means, the algorithm covers all data points without leaving any unassigned points. DBSCAN, based on its parameters, may leave any points unclustered that do not satisfy the euclidean minimal distance (distance between points).

Conclusion

This research paper explored the potential of clustering algorithms to optimize delivery zones in Almaty, addressing the limitations of traditional static delivery zones. By leveraging geolocation data and applying clustering algorithms, the study demonstrates opportunities that dynamically managed delivery zones can provide in terms of alignment with real-world delivery demands in urban environments. The analysis revealed that static zones lack the flexibility to adapt to varying delivery demands. K-means clustering offered a structured approach to dividing delivery zones into uniform zones, highlighting its computational efficiency and suitability for evenly distributed data. High-demand areas were distributed into smaller, more manageable zones, while low-density regions were clustered into larger zones to minimize operational costs. The integration of heatmaps further revealed the insights by providing a visual representation of delivery density, identifying hotspots for decision-making.

Although the comparative results for K-Means and DBSCAN are promising, they rely on two-month temporal scope that provides seasonal fluctuations. To move beyond these constraints, future work should embed clustering within an end-to-end routing simulation that uses network-based travel metrics, consider wider range of temporal scope for dataset, explore spatial or streaming algorithms that adapt zones continuously, extend evaluation to multi-objective criteria covering cost and time-window compliance. Also, it should focus on integrating real-time data such as courier movement dynamics and order volumes to develop these dynamic zones further. The scalability of clustering algorithms like DBSCAN for larger datasets and real-time applications should be explored. This work lays the groundwork for adopting adaptive zoning models that respond to the evolving demands of urban logistics, contributing to more sustainable and efficient delivery systems in smart cities.

Acknowledgement

This research has been funded by the Committee of Science of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No.BR24992852 "Intelligent models and methods of Smart City digital ecosystem for sustainable development and the citizens' quality of life improvement").

REFERENCES

1 Kinelski G. Smart-city trends in the environment of sustainability as support for decarbonization processes // Polityka Energetyczna. – 2022. – Vol. 25. – No. 2.

2 Janjevic M., Winkenbach M. Characterizing urban last-mile distribution strategies in mature and emerging e-commerce markets // Transportation Research Part A: Policy and Practice. – 2020. – Vol. 133. – P. 164–196.

3 Russo F., Comi A. Urban courier delivery in a smart city: the user learning process of travel costs enhanced by emerging technologies // Sustainability. – 2023. – Vol. 15. – No. 23. – P. 16253.

4 Monferdini L., Bottani E. Examining the Response to COVID-19 in Logistics and Supply Chain Processes: Insights from a State-of-the-Art Literature Review and Case Study Analysis // Applied Sciences. – 2024. – Vol. 14. – No. 12. – P. 5317.

5 Li Z., Gu W., Meng Q. The impact of COVID-19 on logistics and coping strategies: a literature review // Regional Science Policy & Practice. – 2023. – Vol. 15. – No. 8. – P. 1768–1795.

6 Yang J. et al. Optimization of real-time traffic network assignment based on IoT data using DBN and clustering model in smart city // Future Generation Computer Systems. – 2020. – Vol. 108. – P. 976–986.

7 Rahman M. A. et al. Logistics Hub Location Optimization: A K-Means and P-Median Model Hybrid Approach Using Road Network Distances // arXiv preprint arXiv:2308.11038. – 2023.

8 Mandal M.P., Archetti C. A Decomposition Approach to Last Mile Delivery Using Public Transportation Systems // arXiv preprint arXiv:2306.04219. – 2023.

9 Barreto S. et al. Using clustering analysis in a capacitated location-routing problem //European journal of operational research. – 2007. – Vol. 179. – No. 3. – P. 968–977.

10 Mandal M.P., Archetti C. A Decomposition Approach to Last Mile Delivery Using Public Transportation Systems // arXiv preprint arXiv:2306.04219. – 2023.

11 Hess A., Spinler S., Winkenbach M. Real-time demand forecasting for an urban delivery platform // Transportation Research Part E: Logistics and Transportation Review. – 2021. – Vol. 145. – P. 102147.

12 Sawik B. Optimizing Last-Mile Delivery: A Multi-Criteria Approach with Automated Smart Lockers, Capillary Distribution and Crowdshipping // Logistics. – 2024. – Vol. 8. – No. 2. – P. 52.

13 Google Sites. K-means clustering algorithm. URL: https://sites.google.com/site/dataclusteringalgorithms/k-means-clustering-algorithm (дата обращения: 31.01.2025).

14 DataCamp. DBSCAN clustering algorithm URL: https://www.datacamp.com/tutorial/dbscan-clustering-algorithm (дата обращения: 31.01.2025).

15 KDnuggets. DBSCAN clustering algorithm in machine learningю – 2020, апрель. URL: https://www.kdnuggets.com/2020/04/dbscan-clustering-algorithm-machine-learning.html (дата обращения: 31.01.2025).

16 Shahapure K.R., Nicholas C. Cluster quality analysis using silhouette score // 2020 IEEE 7th international conference on data science and advanced analytics (DSAA). – IEEE, 2020. – C. 747–748.

17 Ogbuabor G., Ugwoke F. N. Clustering algorithm for a healthcare dataset using silhouette score value // Int. J. Comput. Sci. Inf. Technol. – 2018. – Vol. 10. – No. 2. – P. 27–37.

18 Analytics Vidhya. K-means: Getting the optimal number of clusters. – 2021, май. URL: https://www.analyticsvidhya.com/blog/2021/05/k-mean-getting-the-optimal-number-of-clusters/ (дата обращения: 31.01.2025).

19 GeeksforGeeks. Elbow method for optimal value of k in K-means. URL: https://www.geeksforgeeks.org/elbow-method-for-optimal-value-of-k-in-kmeans/ (дата обращения: 31.01.2025).

20 GeeksforGeeks. DBSCAN clustering in ML: Density-based clustering. URL: https://www.geeksforgeeks.org/dbscan-clustering-in-ml-density-based-clustering/ (дата обращения: 31.01.2025).

21 Srilekha S., Priyadharshini P., Adhilakshmi M. Comparative evaluation of K-Means, hierarchical clustering, and DBSCAN in blood donor segmentation // International Journal for Multidisciplinary Research. – 2024. – Vol. 6. – No. 4. – P. 1–5. https://doi.org/10.36948/ijfmr.2024.v06i04.26755.

REFERENCES

1 Kinelski G. Smart-city trends in the environment of sustainability as support for decarbonization processes. Polityka Energetyczna, 25 (2) (2022).

2 Janjevic M., & Winkenbach M. Characterizing urban last-mile distribution strategies in mature and emerging e-commerce markets. Transportation Research Part A: Policy and Practice, 133, 164–196 (2020).

3 Russo F., & Comi A. Urban courier delivery in a smart city: the user learning process of travel costs enhanced by emerging technologies. Sustainability, 15 (23), 16253 (2023).

4 Monferdini L., & Bottani E. Examining the Response to COVID-19 in Logistics and Supply Chain Processes: Insights from a State-of-the-Art Literature Review and Case Study Analysis. Applied Sciences, 14 (12), 5317 (2024).

5 Li Z., Gu W., & Meng Q. The impact of COVID-19 on logistics and coping strategies: a literature review. Regional Science Policy & Practice, 15 (8), 1768–1795 (2023).

6 Yang J., Han Y., Wang Y., Jiang B., Lv Z., & Song H. Optimization of real-time traffic network assignment based on IoT data using DBN and clustering model in smart city. Future Generation Computer Systems, 108, 976–986 (2020).

7 Rahman M.A., Basheer M.A., Khalid Z., Tahir M., & Uppal,M. Logistics Hub Location Optimization: A K-Means and P-Median Model Hybrid Approach Using Road Network Distances. arXiv preprint arXiv:2308.11038 (2023).

8 Mandal M.P., & Archetti C. A Decomposition Approach to Last Mile Delivery Using Public Transportation Systems. arXiv preprint arXiv:2306.04219 (2023).

9 Barreto S., Ferreira C., Paixao J., & Santos B.S. Using clustering analysis in a capacitated location-routing problem. European journal of operational research, 179 (3), 968–977 (2007).

10 Mandal M.P., & Archetti C. A Decomposition Approach to Last Mile Delivery Using Public Transportation Systems. arXiv preprint arXiv:2306.04219 (2023).

11 Hess A., Spinler S., & Winkenbach M. Real-time demand forecasting for an urban delivery platform. Transportation Research Part E: Logistics and Transportation Review, 145, 102147 (2021).

12 Sawik B. Optimizing Last-Mile Delivery: A Multi-Criteria Approach with Automated Smart Lockers, Capillary Distribution and Crowdshipping. Logistics, 8 (2), 52 (2024).

13 Google Sites. (n.d.). K-means clustering algorithm. Retrieved January 31, 2025. URL: https://sites.google.com/site/dataclusteringalgorithms/k-means-clustering-algorithm

14 DataCamp. (n.d.). DBSCAN clustering algorithm. Retrieved January 31, 2025. URL: https://www.datacamp.com/tutorial/dbscan-clustering-algorithm

15 KDnuggets. DBSCAN clustering algorithm in machine learning (2020, April). Retrieved January 31, 2025. URL: https://www.kdnuggets.com/2020/04/dbscan-clustering-algorithm-machine-learning.html

16 Shahapure K.R., & Nicholas C. Cluster quality analysis using silhouette score. In 2020 IEEE 7th international conference on data science and advanced analytics (DSAA, 2020, October).), pp. 747–748

17 Ogbuabor G., & Ugwoke F.N. Clustering algorithm for a healthcare dataset using silhouette score value. Int. J. Comput. Sci. Inf. Technol, 10 (2), 27–37 (2018).

18 Analytics Vidhya. K-means: Getting the optimal number of clusters Retrieved January 31, 2025. URL: https://www.analyticsvidhya.com/blog/2021/05/k-mean-getting-the-optimal-number-of-clusters/

19 GeeksforGeeks. (n.d.). Elbow method for optimal value of k in K-means. Retrieved January 31, 2025. URL: https://www.geeksforgeeks.org/elbow-method-for-optimal-value-of-k-in-kmeans/

20 GeeksforGeeks. (n.d.). DBSCAN clustering in ML: Density-based clustering. Retrieved January 31, 2025. URL: https://www.geeksforgeeks.org/dbscan-clustering-in-ml-density-based-clustering/

21 Srilekha S., Priyadharshini P., & Adhilakshmi M. Comparative evaluation of K-Means, hierarchical clustering, and DBSCAN in blood donor segmentation. International Journal for Multidisciplinary Research, 6 (4), 1–5 (2024). https://doi.org/10.36948/ijfmr.2024.v06i04.26755

[1]Талғатұлы Ж.,
ғылыми қызметкер, ORCID ID: 0009-0004-1537-7975,
e-mail: zforzesteam@gmail.com
[1]*Амиргалиев Б.Е.,
PhD, Профессор, ORCID ID: 0000-0003-0355-5856,
*e-mail: beibut.amirgaliyev@astanait.edu.kz
[1]Еділхан Д.,
PhD, Ассоциативті Профессор, ORCID ID: 0000-0002-6343-5277,
e-mail: d.yedilkhan@astanait.edu.kz
[1]Тургинбеков А.,
ғылыми қызметкер, ORCID ID: 0000-0002-2742-6779,
e-mail: turginbekovalmaz@gmail.com
[2]Гадаборшев Х.С.,
MBA, ORCID ID: 0009-0001-3868-3594,
e-mail: khavazh@rmggroup.kz

[1]Astana IT University, Астана қ., Қазақстан
[2]Almaty Management University, Алматы қ., Қазақстан

ҚАЛАЛЫҚ КУРЬЕРЛІК ЛОГИСТИКАДАҒЫ КЛАСТЕРЛЕУ ӘДІСТЕРІНЕ НЕГІЗДЕЛГЕН ДЕРЕКТЕРМЕН ОҢТАЙЛАНДЫРУ

Аңдатпа

Қалалар мен олардың инфрақұрылымының қарқынды дамуы жағдайында сапалы қалалық жеткізілімге деген сұраныс үздіксіз өсіп келеді. Бұл жұмыста курьерлік компания ұсынған деректер негізінде жеткізу аймақтарын үйлестіру мүмкіндіктері қарастырылады. Қолмен құрылған дәстүрлі аймақтар көбінесе қаладағы ауа райы, жол қозғалысы, көше желісі сияқты факторларды дұрыс ескере бермейді. Зерттеуде K-Means және DBSCAN кластерлеу алгоритмдерінің жеткізу аймақтарын динамикалық түрде қалыптастыруға қалай ықпал ететіні көрсетіледі. Салыстырмалы талдау Silhouette score және Big-O нотациясы тәрізді көрсеткіштерді пайдаланады. Нәтижелер K-Means алгоритмі біркелкі, құрылымды кластерлер құратынын, ал DBSCAN дерек тығыздығына бейімделіп, икемді шекаралар анықтайтынын дәлелдеді. Көп деңгейлі DBSCAN тәсілі шашыраңқы тапсырыстардың шоғырлануын азайтып, барлық жеткізу нүктелеріне қолжетімділікті арттырады. Осылайша, кластерлеу алгоритмдері курьерлер арасында тапсырыстарды әділ бөлуді жеңілдетіп, операциялық шығындарды қысқартуға мүмкіндік береді. Болашақта алгоритмдердің динамикалық ортада қалай жұмыс істейтінін бағалау үшін нақты уақыттағы деректерді енгізу қажет.

Тірек сөздер: логистика, кластерлеу алгоритмдері, курьерлік жеткізуді оңтайландыру, ақылды қалалар, жеткізу аймақтары, машиналық оқыту.

[1]Талғатұлы Ж.,
научный сотрудник, ORCID ID: 0009-0004-1537-7975,
e-mail: zforzesteam@gmail.com
[1]*Амиргалиев Б.Е.,
PhD, профессор, ORCID ID: 0000-0003-0355-5856,
*e-mail: beibut.amirgaliyev@astanait.edu.kz
[1]Еділхан Д.,
PhD, ассоц. профессор, ORCID ID: 0000-0002-6343-5277,
e-mail: d.yedilkhan@astanait.edu.kz
[1]Тургинбеков А.,
научный сотрудник, ORCID ID: 0000-0002-2742-6779,
e-mail: turginbekovalmaz@gmail.com
[2]Гадаборшев Х. С.,
MBA, ORCID ID: 0009-0001-3868-3594,
e-mail: khavazh@rmggroup.kz

[1]Astana IT University, Астана, Қазақстан
[2]Almaty Management University, Алматы, Қазақстан

МЕТОДЫ ОПТИМИЗАЦИИ, ОСНОВАННЫЕ НА КЛАСТЕРИЗАЦИИ НА ОСНОВЕ ДАННЫХ ГОРОДСКОЙ КУРЬЕРСКОЙ ЛОГИСТИКИ

Аннотация

В условиях стремительного развития городов и их инфраструктуры спрос к качественным городским доставкам возрастает с той же скоростью. Данная работа исследует возможности динамического распределения зон доставки для курьерских доставок на основе данных, предоставляемых курьерской компанией. Традиционные зоны доставки, создаваемые вручную, часто не обеспечивают релевантность картины по отношению к реальной ситуации в городе (погода, трафик, дороги и т.д.). В данном исследовании приводятся результаты того, как алгоритмы кластеризации K-Means и DBSCAN могут способствовать динамическому распределению зон доставок в виде кластеров. Сравнительный анализ включает в себя учет таких показателей, как значение Silhouette и вычислительная сложность Big-O Notation. Результаты показывают, что K-Means алгоритм создает структурированные и равномерные кластеры, в то время как DBSCAN показывает результаты в определении гибких кластеров с учетом плотности данных в регионе. Многоуровневый DBSCAN предоставляет возможность уменьшить концентрацию «шумов», тем самым увеличивая охват всех точек доставок. Полученные результаты отмечают преимущества использования алгоритмов кластеризации в создании динамических зон доставки для улучшения процессов распределения заказов между курьерами и уменьшением операционных расходов. В дальнейшие исследования следует включить получение данных в реальном времени для наблюдений за работой алгоритмов в динамической среде.

Ключевые слова: логистика, алгоритмы кластеризации, оптимизация курьерской доставки, умные города, зоны доставки, машинное обучение.

Article submission date: 27.02.2025