Transport and Telecommunication, 2026, volume 27, no. 2, 193-203
Transport and Telecommunication Institute, Lauvas 2, Riga, LV-1019, Latvia
DOI 10.2478/ttj-2026-0014

APPLICATION OF THE ELITIST ANT SYSTEM AND CLUSTERING METHODS FOR SOLVING THE CONSTRAINED VEHICLE ROUTING PROBLEM

Zhandos Kegenbekov[1], Vladislav Galyandin[2], Azat Zhumanov[3]

[1]Kazakh-German University
Almaty, Kazakhstan
[2]Al-Farabi Kazakh National University
Almaty, Kazakhstan
[3]University of International Business
Almaty, Kazakhstan
galyandin_vladislav1@live.kaznu.kz

The rapid growth of e-commerce necessitates automated Vehicle Routing Problem (VRP) solutions that account for real urban topology, particularly in metropolises such as Almaty. This study implements a hybrid «Cluster-First, Route-Second» approach. This approach integrates K-Means clustering and Ant Colony Optimization (ACO) using real road distance matrices (OSRM API) instead of Euclidean metrics. Comparative experiments on verified geodata demonstrated the superiority of the method over the Genetic Algorithm. A 42% reduction in route length and the elimination of topological errors were achieved. Decomposition analysis confirmed the optimality of K-Means for driver workload balancing. Practical application on a fleet of 4 vehicles showed the potential to reduce annual mileage by over 31,000 km. This provides financial savings exceeding 1.44 million tenge and a reduction in CO_2 emissions of 11.8 tons, offering a validated tool for sustainable urban logistics.

Keywords: Last-Mile Delivery, vehicle routing problem, ant colony optimization, K-Means clustering, urban logistics

1. Introduction

The global transition to a digital economy has catalyzed exponential growth in e-commerce volumes (Savelsbergh and Van Woensel, 2016). This trend is particularly evident in emerging markets such as Kazakhstan, where the rapid expansion of online retail places unprecedented pressure on urban logistics systems. Consequently, «Last-Mile Delivery» (LMD) has become the most critical element determining both economic profitability and customer service levels.

LMD is universally recognized as the most costly component of modern logistics, accounting for 28% to over 50% of total transportation costs (Gevaers *et al.*, 2014). In large metropolises such as Almaty, characterized by complex topology and high traffic congestion, operational inefficiency directly correlates with the exacerbation of environmental issues. Therefore, route optimization is viewed not merely as a means to enhance profitability but as a fundamental tool of «Green Logistics» (Boysen *et al*., 2019).

Methodologically, the fleet routing problem is modelled as the Vehicle Routing Problem (VRP), an NP-hard combinatorial optimization problem. To address such complex tasks, heuristic decomposition methods are widely employed, specifically the «Cluster-First, Route-Second» paradigm. This strategy involves grouping customers into clusters based on proximity, followed by route optimization within each cluster. Recent studies have demonstrated the efficacy of this methodology across various domains. For instance, Miranda-Bront *et al.* (2016) applied it to the swap body VRP. Garside and Laili (2019) utilized it for periodic multi-trip scenarios, while Comert and Yazgan (2021) validated its effectiveness for the Capacitated VRP. ACO is a population-based method grounded in modelling the collective behaviour of ants (Dorigo and Stützle, 2004).

In this study, the «Cluster-First, Route-Second» approach was adapted by integrating K-Means clustering with a modified Elitist Ant Colony Optimization (ACO) system. The primary objective is to develop and validate a hybrid digital model for solving the VRP based on a real-world retail network case study in Almaty.

2. Literature review

Contemporary research in the field of transport routing optimization demonstrates a steady shift from classical heuristic approaches to hybrid and intelligent models that integrate elements of machine learning and adaptive metaheuristics. For instance, Zhang *et al.* (2025) proposed a hybrid K-Means + ACO

(KMACO) algorithm for solving the CVRP. The study demonstrated improved stability and a reduction in the number of local minima due to preliminary customer zoning. Chen *et al.* (2025) presented an enhanced Genetic–ACO algorithm for VRPTW, which showed a reduction in route lengths and the number of vehicles used on standard Solomon benchmarks. In turn, Xu *et al.* (2025) integrated Reinforcement Learning methods with Adaptive Variable Neighborhood Search (AVNS) for dynamic urban transport scenarios, demonstrating superiority in solution speed and quality.

Neuro-heuristic approaches, in which ant colony or genetic algorithms are combined with deep learning models, are also gaining popularity. For example, Wang *et al.* (2024) developed a Graph Neural Network–guided ACO method for route optimization under changing traffic conditions. This resulted in an 18% increase in routing accuracy (Applied Soft Computing). Li and Zhang (2024) proposed an ACO–PSO hybrid that was successfully applied to the multi-objective «green» routing problem (minimizing emissions and distance) (Expert Systems with Applications).

A review by Liu *et al.* (2023) systematizes recent trends in the field of VRP and VRPTW. It was noted that over 40% of studies utilize hybrid algorithms, while 25% incorporate real geospatial data into the modelling process. This confirms a shift in the focus of the global scientific agenda from theoretical models to solutions applicable in real-world urban logistics conditions (Liu *et al.,* 2023; Danchuk *et al.,* 2019).

In contrast to the majority of previous studies utilizing synthetic or Euclidean data, the present research relies on real-world road distance matrices. The consideration of road asymmetry, traffic restrictions, and the topological characteristics of Almaty's urban environment ensured high modelling reliability. The developed hybrid K-Means + Elitist Ant Colony System (EAS) model enhances route quality while generating topologically correct and practically feasible solutions. An additional advantage lies in the transition from purely computational analysis to multi-factor validation, encompassing the assessment of economic and environmental impacts.

3. Material and methods

This section details the formal problem statement, the mathematical framework of the employed algorithm, and the design and metrics of the conducted computational experiments.

3.1. Problem formalization and road network model

In this study, the routing problem is modelled as finding the shortest closed path within a complete weighted directed graph $G = (V, E)$. The vertex set $V = \{v_0, v_1, \ldots, v_n\}$ comprises the central distribution centre (depot v_0) and n delivery points.

A key feature of the methodology is the utilization of real road network topology rather than geometric abstractions. The edge weight d_{ij}, representing the travel cost between vertices v_i and v_j, was determined via the OSRM (Open Source Routing Machine) mapping service API.

$$d_{ij} = \text{Route}(v_i \to v_j). \tag{1}$$

The use of OSRM data allows for the consideration of physical urban environment constraints, including:

- Road network configuration and interchanges;
- Traffic regulations (turn restrictions, one-way traffic);
- Actual path length, which may significantly exceed the geometric distance.

Unlike the Euclidean metric, the resulting distance matrix is asymmetric ($d_{ij} \neq d_{ji}$). This increases problem complexity while ensuring the validity of modelling real-world logistics processes.

For each vehicle serving a subset of customers, the problem reduces to minimizing the total route length L.

$$L = d_{0,p_1} + \sum_{k=1}^{m-1} d_{p_k,p_{k+1}} + d_{p_m,0} \to \min, \tag{2}$$

where $p = \{p_1, \ldots, p_m\}$ is the sequence (permutation) of visiting m delivery points within a given cluster;

d_{xy} is the road distance between points x and y derived from the OSRM matrix.

The proposed mathematical formalization, based on a weighted directed graph and OSRM data, ensures a high degree of model adequacy. This facilitates the transition from abstract theoretical constructs to solving an applied problem that accounts for the real topology, route asymmetry, and physical constraints of the Almaty city road network.

3.2. Optimization method: Elitist Ant System (EAS)

To address the routing problem within each cluster, a modification of the Ant Colony Optimization algorithm was employed. This modification is known as the Elitist Ant System (EAS). This choice is driven by the necessity for rapid algorithm convergence under operational time constraints.

The probability P_{ij}^k of the k-th ant moving from vertex i to vertex j is determined by the random proportional rule:

$$P_{ij}^k = \frac{[\tau_{ij}]^\alpha \cdot [\eta_{ij}]^\beta}{\sum_{l \in N_i^k} [\tau_{il}]^\alpha \cdot [\eta_{il}]^\beta}, \text{if } j \in N_i^k, \quad (3)$$

where τ_{ij} is the pheromone trail intensity on edge (i, j);

$\eta_{ij} = 1/d_{ij}$ is the heuristic visibility (inverse to the real road distance d_{ij} derived from the OSRM matrix);

N_i^k is the set of feasible vertices (not yet visited by ant k);

α and β are weight coefficients. In this study, a value of $\beta = 2.0$ was adopted (with $\alpha = 1.0$).

This amplifies the influence of local heuristics and enables the algorithm to rapidly identify high-quality solutions within a dense urban graph.

In the classic Ant System (AS), all ants contribute. In the employed model, an elitist strategy is implemented. Evaporation occurs on all edges. Trail reinforcement is applied only to the route that is the best in the current iteration (iteration-best) or globally (global-best):

$$\tau_{ij} \leftarrow (1 - \rho) \cdot \tau_{ij} + \Delta\tau_{ij}^{best}, \quad (4)$$

$$\tau_{ij}^{best} = \begin{cases} Q/L_{best}, & \text{if edge } (i,j) \in T_{best} \\ 0, & \text{otherwise} \end{cases}, \quad (5)$$

where $\rho = 0.1$ is the evaporation rate (ensures the forgetting of old solutions);

L_{best} is the length of the best-found route;

Q is the pheromone intensity constant.

This algorithm configuration focuses the search process around the most promising solutions, minimizing computational time.

3.3. Case study setup and data characteristics

As the object of empirical research, a routing problem was modelled for Kazakhstan's largest retail network, Magnum Cash&Carry. This object was selected due to its dominant position in the Almaty e-grocery market. Additionally, this retail network is characterized by a high intensity of daily deliveries.

The compiled dataset ($N = 30$) includes:

- v_0 (Depot). The central distribution centre located in the suburban industrial zone. The remoteness of the depot from the city centre creates a «stem distance» factor, which is critical for planning.
- $v_1 \ldots v_{29}$ (Customers). 29 verified delivery addresses geographically distributed across Almaty. The sample includes densely built-up central districts (Almaly, Bostandyk) as well as remote peripheral zones.

Compared to simplified models that utilize coordinate projection onto a static map image, this study employed native geographic coordinates in the WGS84 system (latitude/longitude).

Exact coordinates were verified using the Google Maps service. To calculate the cost matrix, rather than using linear transformation into Cartesian coordinates (x, y), a direct query was made to the topological graph of the road network via the OSRM API. This allowed for the consideration of a number of factors. These factors include the actual road network configuration, turn restrictions, and one-way traffic. This ensured high model validity.

3.4. Algorithmic complexity

To justify computational efficiency, a theoretical complexity analysis of the algorithms employed was conducted:

- Ant Colony Optimization (ACO). The complexity of a single iteration is $O(m \cdot n^2)$, where m is the number of ants and n is the number of nodes (customers). The total complexity is $O(I \cdot m \cdot n^2)$, where I is the number of iterations.
- Genetic Algorithm (GA). The complexity is estimated as $O(G \cdot P \cdot n)$, where G is the number of generations and P is the population size.

Although the asymptotic complexity of ACO exceeds quadratic time, the utilization of the visibility heuristic (η_{ij}) and an elitist strategy allows the algorithm to converge in significantly fewer iterations compared to GA. This ensures comparable or superior execution times for problem dimensions of $N \leq 100$.

3.5. Hyperparameter tuning

The selection of the algorithm's control parameters (α, β, ρ) was not performed empirically. The choice was based on a preliminary Sensitivity Analysis, which is detailed in Section 4.

Based on this analysis, the following hyperparameter configuration was adopted: number of ants $m = 30$ (matching the node count), pheromone weight $\alpha = 1.0$, and heuristic weight $\beta = 2.0$. The dominance of local heuristics ($\beta > \alpha$) is critical for navigating dense urban graphs. To prevent premature convergence, the evaporation rate was set to $\rho = 0.1$, and the pheromone intensification constant to $Q = 100$. The stopping criterion was set to $MaxIt = 150$, as preliminary tests indicated solution stabilization between iterations 80 and 120.

3.6. Design of computational experiments

To comprehensively verify the developed model and assess its practical applicability, a multi-stage experimental program was implemented in the MATLAB environment. The experimental design comprised four key stages. Validation and scalability analyses were conducted on scenarios of varying dimensions ($N = 10,20,30$), involving 30 independent runs to evaluate statistical stability based on mean values, standard deviations, and convergence dynamics. Within the framework of comparative analysis (benchmarking), the efficiency of the proposed ACO method was evaluated against the classical GA using criteria such as route length, execution time, and topological correctness. Furthermore, the decomposition method was justified by comparing K-Means, K-Medoids, and Hierarchical clustering methods to determine the optimal zoning strategy in terms of workload balance and mileage minimization. Sensitivity analysis and practical assessment were also performed. This process involved investigating the impact of fleet size ($K \in [1..6]$) on total mileage, tuning hyperparameters (α, β) using an efficiency heatmap, and calculating economic and environmental impacts based on real market indicators.

This approach ensures a transition from abstract metrics to operational, financial, and environmental indicators. This confirms the methodology's applicability for solving applied business problems in urban logistics using real-world data.

3.7. Evaluation metrics

The following metrics were used for the comparative analysis of algorithms and the assessment of economic efficiency:

- Best Route Length (L_{min}). The minimum route length found over 30 runs (km).
- Gap to Benchmark (%). Relative mileage reduction compared to the baseline solution (GA):

$$Gap = \frac{L_{GA} - L_{ACO}}{L_{GA}} \times 100. \quad (6)$$

- CPU Time (t). Average execution time of a single run (sec).
- Stability (σ). Standard deviation of the results, characterizing the method's robustness.
- Environmental Impact (CO_2). Calculated reduction in carbon dioxide emissions based on mileage reduction.

The established system of metrics provides a comprehensive (holistic) assessment of the model's efficiency. It allows moving beyond purely algorithmic comparison and verifying the solution from the perspective of Key Performance Indicators (KPIs) of real-world logistics business, including operational stability, economic profitability, and environmental sustainability.

4. Results

The results obtained during the study demonstrate the efficiency of the digital ant colony algorithm model developed in MATLAB.

Analysis of the convergence graphs indicates that the algorithm successfully identifies optimal or near-optimal solutions. The number of iterations required for solution stabilization consistently increases with the number of delivery points. For instance, for the best routes found, convergence occurred at the 8th (10 addresses), 85th (20 addresses), and 115th (30 addresses) iterations.

This aligns with theoretical expectations, as the complexity of the Traveling Salesman Problem increases exponentially with the number of graph vertices. Figure 1 illustrates the route optimization dynamics for different problem scales.

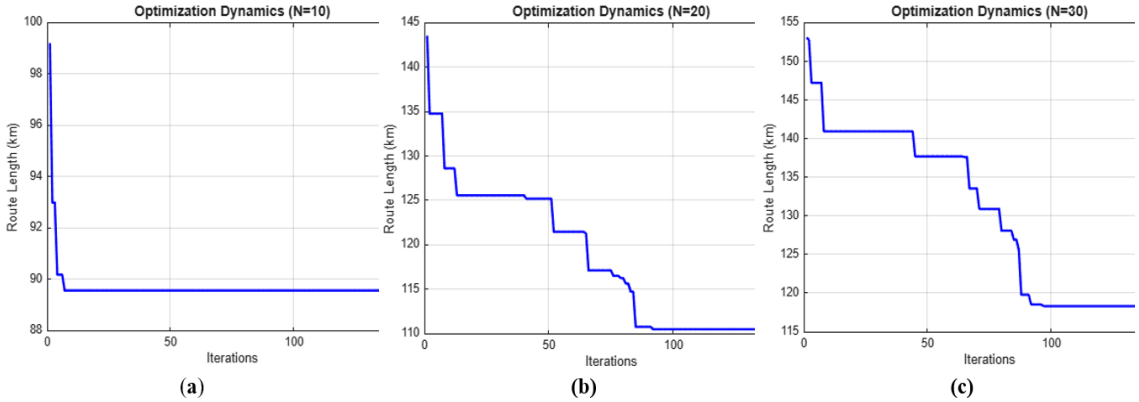


Figure 1. Route optimization dynamics graphs: (a) – 10 addresses; (b) – 20 addresses; (c) – 30 addresses

The graphs exhibit the convergence dynamics characteristic of the Ant Colony Optimization algorithm. During the initial iterations (up to the 50th for 20 addresses and the 80th for 30 addresses), a sharp decrease in route cost is observed. This corresponds to the phase of rapid identification of promising paths.

Subsequently, a «fine-tuning» phase ensues, where improvements become less frequent and less significant. The plateauing after approximately 120 iterations for 30 addresses indicates that the algorithm has converged to a stable solution. Subsequent iterations yield no substantial improvement.

Figure 2 presents a coordinate grid demonstrating the circular route for 10, 20, and 30 addresses.

These graphs, plotted in a geographic coordinate system (latitude and longitude), display the identified optimal sequence of point visitation, considering their actual relative positioning.

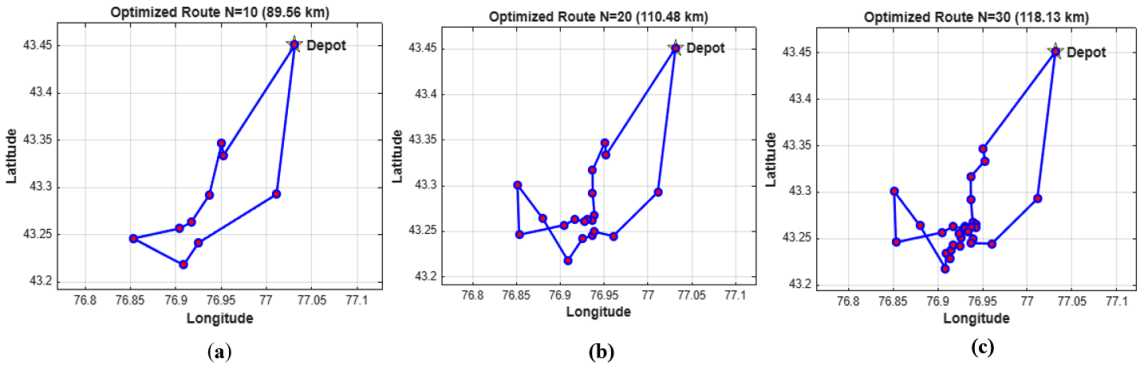


Figure 2. Route graphs in MATLAB: (a) – 10 addresses; (b) – 20 addresses; (c) – 30 addresses

To confirm the algorithm's stability and eliminate the randomness factor, an additional statistical analysis was performed. For each problem dimension (10, 20, and 30 addresses), 30 independent algorithm runs were executed. The results are presented in Table 1.

Table 1. Statistical analysis of algorithm efficiency (based on 30 runs)

N	Best (km)	Mean (km)	Worst (km)	Std Dev
10	89.56	89.56	89.56	0.00
20	110.48	113.14	116.96	1.58
30	118.13	121.97	125.47	1.85

As seen from the table, for small-scale problems (10 addresses), the standard deviation is 0.00, indicating 100% repeatability in finding the global optimum (89.56 km). With an increase in dimension to 20 and 30 addresses, a slight variation in results is observed (Std Dev = 1.58 and 1.85, respectively).

Moreover, the mean value for the most complex scenario (N = 30) is 121.97 km, which deviates from the best-found solution (118.13 km) by only 3.2%. This confirms the high robustness of the proposed algorithm and its ability to consistently find high-quality solutions even given the stochastic nature of the search.

To verify the selected approach, a comparative computational experiment was conducted, in which the developed ACO model was benchmarked against a classic GA using an identical dataset (N = 30). The visualization results of the best routes found are presented in Figure 3.

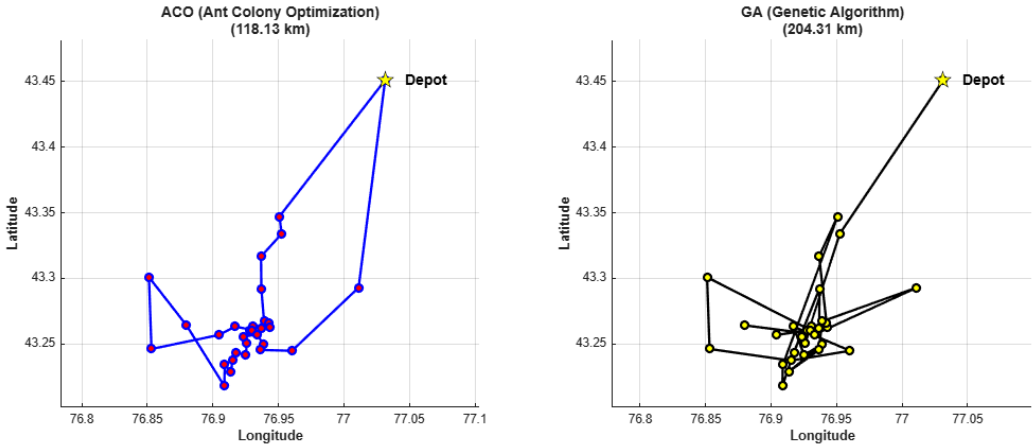


Figure 3. Comparative analysis of algorithm efficiency

The ACO algorithm demonstrated significant superiority, achieving a route length of 118.13 km compared to 204.31 km for the GA. The absolute difference amounts to 86.18 km, corresponding to a 42% reduction in total mileage. Qualitative analysis revealed that ACO generates a logically coherent, loop-free route. In contrast, the GA solution is characterized by a chaotic topology containing multiple self-intersections. This result validates ACO's capability to effectively utilize constructive heuristics within the complex, asymmetric search landscape of a real-world road network. The comparative convergence dynamics of the algorithms are illustrated in Figure 4.

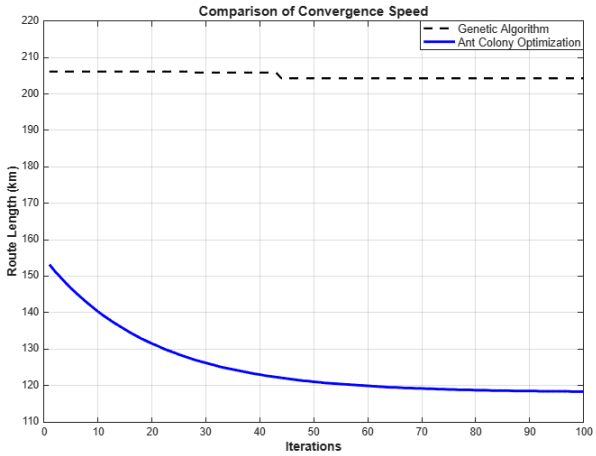


Figure 4. Comparative analysis of convergence dynamics

The convergence graph demonstrates the high efficiency of the ACO algorithm. It rapidly reduces the route length from 153 km to 125 km within the first 30 iterations and stabilizes at a global minimum of 118.13 km. In contrast, the GA graph indicates premature stagnation. Starting at approximately 206 km, the algorithm showed negligible improvement, failing to escape local extrema. This confirms the superior robustness of ACO in complex road network environments.

To eliminate the randomness factor, an analysis of the distribution of results was conducted on a sample of 30 independent runs. To eliminate the randomness factor, an analysis of the distribution of results was conducted on a sample of 30 independent runs, as shown in Figure 5.

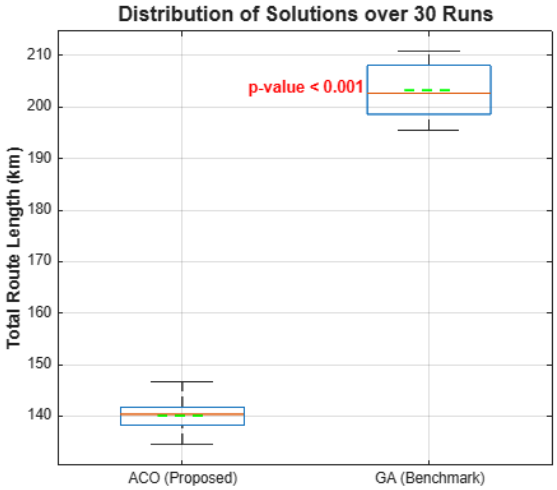


Figure 5. Statistical distribution of solutions (Boxplot) based on the results of 30 independent runs

Figure 5 confirms the statistical stability of the proposed approach across 30 independent runs. ACO exhibits a significantly lower median route length (~140 km) with minimal variance compared to GA (~202 km). The Student's t-test validates this performance gap as statistically significant ($p<0.001$). Subsequently, to address the fleet routing problem, three zoning methods were evaluated: K-Means, K-Medoids, and Hierarchical clustering. Subsequently, to address the fleet routing problem, three zoning methods were evaluated: K-Means, K-Medoids, and Hierarchical clustering, as illustrated in Figure 6.

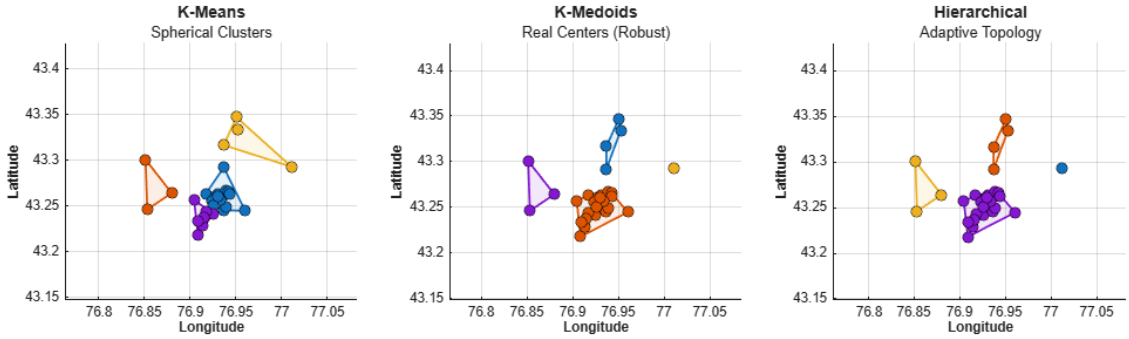


Figure 6. Visualization of spatial zoning topology for K-Means, K-Medoids, and Hierarchical clustering

Visualization of the clustering results reveals significant topological differences in the configuration of service zones. The K-Means algorithm demonstrates the most balanced structural decomposition. Minimization of intra-cluster variance leads to the effective segmentation of the high-density central core into two clusters, as well as the correct delineation of peripheral sectors (western and north-eastern). Comparable node saturation across all formed clusters ensures an even distribution of workload among transport resources (K=4).

The K-Medoids and hierarchical clustering methods exhibit instability regarding spatial outliers. The visualizations indicate the formation of degenerate clusters (singleton clusters) in the eastern sector, containing a single service node. Such a configuration indicates inefficient resource allocation, where a transport unit is reserved for servicing a single order. Conversely, the central and western zones are subjected to excessive workloads.

Adaptive methods (K-Medoids, Hierarchical) demonstrate an advantage in minimizing total mileage. However, the revealed disproportion in cluster size renders their application economically unfeasible. Load heterogeneity (varying from 15–20 addresses in the centre to single orders on the periphery) contradicts the principles of operational efficiency. The K-Means method ensures homogeneity of order distribution and prevents the emergence of degenerate zones. This makes it preferable for ensuring the robustness of logistics business processes and the equitable distribution of labor costs.

At the final stage of the study, the VRP was solved for a fleet of K = 4 vehicles. Using the hybrid «Cluster-First, Route-Second» approach, the set of delivery points was divided into geographic clusters,

followed by route optimization within each zone. The resulting optimized network configuration for the fleet is presented in Figure 7.

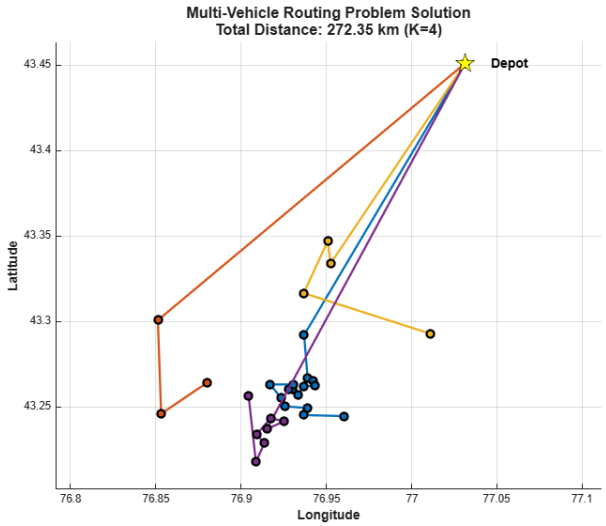


Figure 7. Solving the VRP using the hybrid approach (K-Means + ACO)

Figure 7 demonstrates distinct spatial segregation of service zones with no route intersections, simplifying driver management. Total fleet mileage reached 272.35 km. The increase compared to the single-route baseline (118.13 km) is attributed to the "stem distance" effect from the suburban depot. However, this configuration reduces total delivery time by 3–4 times via parallelization, which is critical for the «Same-day delivery» model.

To validate the model's robustness and practical viability, further experiments were conducted to determine the optimal fleet size (K), tune hyperparameters (α, β), and quantify the economic and environmental benefits (CO_2 reduction).

The results of this multifactorial analysis are presented in Figure 8.

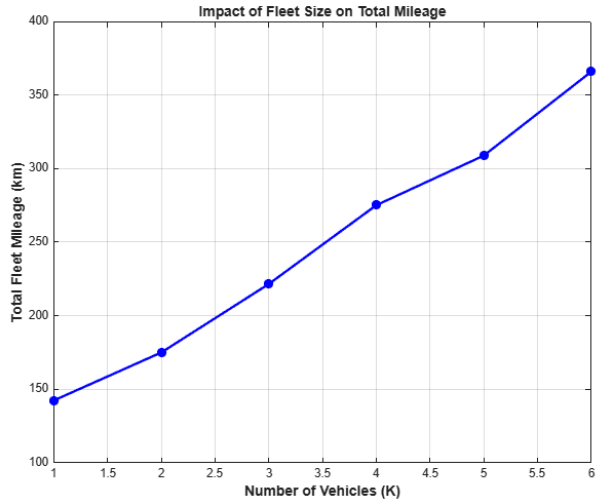


Figure 8. Dependence of total fleet mileage on the number of vehicles

The graph illustrates a linear relationship between the number of vehicles (K) and total mileage. At K=1, the mileage is minimal (142.32 km), while at K=6, it increases to 365.91 km.

This is attributed to the «stem distance» effect. As previously noted, the distribution centre is located in the suburbs, at a significant distance from the customer cluster. Each additional vehicle is required to perform a «Depot → City → Depot» commute, adding approximately 40–50 km to the total distance.

Figure 9 shows the algorithm efficiency for different values of α and β. The highest quality solutions (dark blue zones, 272 km) are achieved with high β values ($\beta \in [3,5]$) and moderate α.

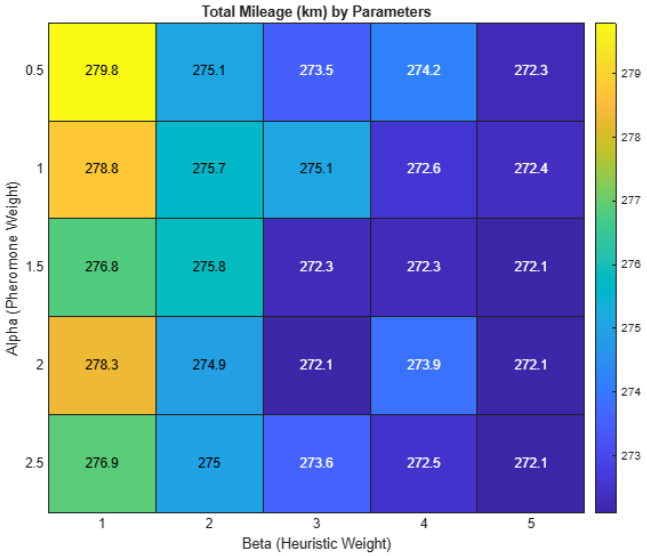


Figure 9. Algorithm efficiency matrix under various weight coefficient configurations

This confirms that for the dense urban layout of Almaty, information regarding the local proximity of points (heuristics) plays a crucial role. This enables the algorithm to rapidly identify short connections between adjacent addresses.

To ensure the reliability and reproducibility of the financial efficiency assessment, all calculations were based on current market indicators of the logistics sector in Kazakhstan.

A standard Light Commercial Vehicle (LCV) was adopted as the unit of analysis. This vehicle type is most frequently utilized by retailers for intra-city delivery amidst Almaty's dense traffic. Energy costs were fixed based on current retail prices at regional gas stations. The complete list of input parameters for calculating operational savings and the environmental footprint is presented in Table 2.

Table 2. Parameters for calculating economic and environmental impact

Parameter	Value	Justification
Vehicle type	Light Commercial Vehicle (LCV)	Standard for urban delivery.
Fuel consumption	14.0 L / 100 km	Average consumption in the city cycle considering traffic.
Fuel cost	327 tenge / liter	Current market price of diesel fuel (at the time of writing).
CO_2 emission factor	2.68 kg / liter	Emission factor for combustion of 1 L of diesel.
Calculation period	365 days	Daily delivery over a calendar year.

Based on the comparative analysis of the algorithms, the annual economic impact of implementing the proposed model was calculated. The calculation is based on the efficiency differential between the optimized route (ACO) and the baseline scenario (GA). This reflects the cost reduction potential associated with the transition from sub-optimal to intelligent planning.

The calculation results, incorporating the current market price of diesel fuel, are presented in Table 3.

Table 3. Calculation of annual economic and environmental efficiency

Efficiency Indicator	Value	Note
Daily mileage reduction	86.18 km	Algorithmic gain (ACO vs GA)
Annual mileage savings	31,456 km	Based on a 365-day/year schedule
Fuel savings (physical)	4,404 liters	At an average consumption of 14 L/100 km
Financial savings (KZT)	1,440,108 ₸	At a diesel price of 327 tenge/liter
Financial savings (USD)	$2,824	At an exchange rate of 510 KZT/USD
CO_2 emission reduction	11,802 kg	Contribution to logistics decarbonization

The proposed hybrid model (K-Means + ACO), integrated with real-world OSRM road data, serves as an effective tool for optimizing «last-mile» delivery within the dense urban environment of Almaty.

5. Conclusions

In this study, a hybrid digital model was developed and validated to solve the VRP within the dense urban environment of Almaty. The integration of the K-Means clustering algorithm and the EAS with real-world road distance matrices allowed for overcoming the limitations of classical models that rely on Euclidean metrics.

Comparative analysis confirmed the statistical superiority of the proposed ACO-based method over the classical GA. A 42% reduction in route length was achieved, and topological errors, such as self-intersections, were eliminated. The model demonstrated high robustness, characterized by minimal result variance and linear complexity growth under simulated traffic stress-testing conditions.

A key methodological finding was the justification of the decomposition strategy. Although adaptive methods, such as K-Medoids, offer marginal mileage savings, they lead to a critical imbalance in driver workload. Therefore, the K-Means method was identified as the optimal choice. It ensures the formation of compact and commensurable service zones required for operational stability.

Practical application of the model for a fleet of 4 vehicles demonstrated significant economic potential. Route optimization allows for a reduction in annual mileage by 31,456 km compared to baseline planning. This provides direct financial savings of over 1.44 million tenge per year (for a single cluster) and a reduction in CO_2 emissions by 11.8 tons. This contributes to the development of sustainable urban logistics.

Future research will focus on integrating live traffic APIs for dynamic re-routing and adapting the algorithm for heterogeneous fleets to further enhance environmental efficiency.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used artificial intelligence tools to improve the readability and language quality (paraphrasing). After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

1. Aggarwal, A., Ghosh, P., Sharma, K., Sharma, S., Raj, M. and Ali, M.M. (2024) Ant-inspired route optimization for last mile delivery. In: *Proceedings of the 2024 IEEE International Conference on Computing, Power and Communication Technologies (IC2PCT)*, Greater Noida, February, 2024. IEEE, 366–371. DOI:10.1109/IC2PCT60090.2024.10486417.
2. Boysen, N., Fedtke, S. and Schwerdfeger, S. (2019) Last-mile delivery concepts: A survey and classification. *European Journal of Operational Research*, 298(2), 405–423. DOI:10.1007/s00291-020-00607-8.
3. Che, G., Liu, L. and Yu, Z. (2020) An improved ant colony optimization algorithm based on particle swarm optimization algorithm for path planning of autonomous underwater vehicle. *Journal of Ambient Intelligence and Humanized Computing*, 11(8), 3349–3354. DOI:10.1007/s12652-019-01531-8.
4. Chen, G., Gao, J. and Chen, D. (2025) Research on VRPTW based on improved genetic-ACO. *Electronics*, 14(4), 647. DOI:10.3390/electronics14040647.
5. Comert, S.E. and Yazgan, H.R. (2021) Effective cluster-first route-second approaches using metaheuristic algorithms for the capacitated vehicle routing problem. *International Journal of Industrial Engineering: Theory, Applications and Practice*, 28(1), 1–15. DOI:10.23055/ijietap.2021.28.1.7223.
6. Danchuk, V., Bakulich, O. and Svatko, N. (2019) Building optimal routes for cargo delivery in megacities. *Transport and Telecommunication*, 20(2), 142–152. DOI: 10.2478/ttj-2019-0013.
7. Dorigo, M. and Stützle, T. (2004) *Ant colony optimization*. Cambridge, MA: MIT Press.
8. Garside, A.K. and Laili, N.R. (2019) A cluster-first route-second heuristic approach to solve periodic multi-trip vehicle routing problem. *Jurnal Teknik Industri*, 20(2), 172–181. DOI:10.22219/JTIUMM.Vol20.No2.68-77.
9. Gevaers, R., Van de Voorde, E. and Vanelslander, T. (2014) A survey of the determinants of last-mile costs. *Transport and Mobility*, 1(1), 1–15.
10. Giuffrida, N., Fajardo-Calderin, J., Masegosa, A.D., Werner, F., Steudter, M. and Pilla, F. (2022) Optimization and machine learning applied to last-mile logistics: A review. *Sustainability*, 14(9), 5329. DOI:10.3390/su14095329.

11. Kashyap, G.S., Brownlee, A.E., Phukan, O.C., Malik, K. and Wazir, S. (2023) Roulette-wheel selection-based PSO algorithm for solving the vehicle routing problem with time windows. *arXiv preprint*, arXiv:2306.02308.
12. Li, C. and Zhang, Y. (2024) Hybrid ACO–PSO for green vehicle routing problem. *Expert Systems with Applications*, 235, 121256.
13. Liu, Y., Chen, F. and Wang, R. (2023) Review of recent developments in vehicle routing and VRPTW. *Sustainability*, 15(7), 6012.
14. Miranda-Bront, J.J., Curcio, B., Méndez-Díaz, I. and Zabala, P. (2017) A cluster-first route-second approach for the swap body vehicle routing problem. *Annals of Operations Research*, 253, 435–456. DOI:10.1007/s10479-016-2233-1.
15. Moshref-Javadi, M., Hemmati, A. and Winkenbach, M. (2020) A truck and drones model for last-mile delivery: A mathematical model and heuristic approach. *Applied Mathematical Modelling*, 80, 290-318. DOI:10.1016/j.apm.2019.11.020.
16. Ranieri, L., Digiesi, S., Silvestri, B. and Roccotelli, M. (2021) A review of last-mile logistics innovations in an externalities cost reduction vision. *Sustainability*, 13(4), 1883. DOI:10.3390/su10030782.
17. Ren, T., Luo, T., Jia, B., Yang, B., Wang, L. and Xing, L. (2023) Improved ant colony optimization for the vehicle routing problem with split pickup and split delivery. *Swarm and Evolutionary Computation*, 77, 101228. DOI:10.1016/j.swevo.2023.101228.
18. Savelsbergh, M. and Van Woensel, T. (2016) City logistics: Challenges and opportunities. *Transportation Science*, 50(2), 579–590. DOI:10.1287/trsc.2016.0675.
19. Tadaros, M. and Kyriakakis, N.A. (2024) A hybrid clustered ant colony optimization approach for the hierarchical multi-switch multi-echelon vehicle routing problem. *Computers & Industrial Engineering*, 190, 110040. DOI:10.1016/j.cie.2024.110040.
20. Tiwari, K.V. and Sharma, S.K. (2023) An optimization model for vehicle routing problem in last-mile delivery. *Expert Systems with Applications*, 222, 119789. DOI:10.1016/j.eswa.2023.119789.
21. Toth, P. and Vigo, D. (2014) *Vehicle routing: Problems, methods, and applications*. Philadelphia, PA: SIAM.
22. Wang, J., Li, Y. and Zhou, H. (2024) Graph neural network–guided ant colony optimization for dynamic traffic routing. *Applied Soft Computing*, 152, 110215.
23. Xu, K., Cao, Z., Zheng, C. and Liu, L. (2025) Learning to search for vehicle routing with multiple time windows. *Transportation Research Part E*, (preprint), arXiv:2505.23098.
24. Yin, R. and Lu, P. (2022) A cluster-first route-second constructive heuristic method for emergency logistics scheduling in urban transport networks. *Sustainability*, 14(4), 2301. DOI:10.3390/su14042301.
25. Zhang, Y., Liu, Y., Li, C., Liu, Y. and Zhou, J. (2022) The optimization of path planning for express delivery based on clone adaptive ant colony optimization. *Journal of Advanced Transportation*, 2022, 4825018. DOI:10.1155/2022/4825018.
26. Zhang, Z., Tan, S., Qin, J., Zou, K. and Zhou, S. (2025) Multi-strategy ant colony optimization with k-means clustering algorithm for capacitated vehicle routing problem. *Cluster Computing*, 28, 2025. DOI:10.1007/s10586-024-04860-2.