

Article

A Spatially Constrained PCA–MST-Based Clustering Model for Railway Freight Management in Kazakhstan

Aizhan Mukhametzhanova ¹, Marat Baiseitov ¹, Aliya Izbairova ^{2,*}, Dariga Kushtayeva ² and Gabit Bakyt ^{3,*}

¹ Department of Organization of transport, traffic and transport operation, L. N. Gumilyov Eurasian National University, Astana 010000, Kazakhstan; mukhametzhanova_av@enu.kz (A.M.); marat2240@gmail.com (M.B.)

² School of Transport Engineering and Logistics Named After M. Tynyshpayev, Satbayev University, Almaty 050013, Kazakhstan; d.kushtaeva@satbayev.university

³ Department of Rolling stock, Mukhametzhana Tynyshbayev ALT University, Almaty 050012, Kazakhstan

* Correspondence: a.izbairova@satbayev.university (A.I.); g.bakyt@alt.edu.kz (G.B.)

Abstract

Background: Kazakhstan's railway network exhibits substantial spatial heterogeneity, limiting the effectiveness of uniform freight management strategies and necessitating differentiated analytical approaches to regional transport planning. This study aims to develop a spatially constrained clustering model for railway freight management based on the economic, infrastructural, and operational characteristics of the regions served by KTZh—Freight Transportation LLP. **Methods:** The proposed methodology integrates principal component analysis (PCA) with a minimum spanning tree (MST) algorithm under railway connectivity constraints. A dataset comprising 20 standardized indicators for 17 regions of Kazakhstan was analyzed. **Results:** PCA reduced the original variable space to five principal components, explaining 77.9% of the cumulative variance. Cluster validity was confirmed using the Elbow, Silhouette, and Calinski–Harabasz indices, resulting in the identification of six spatially connected transport clusters with distinct functional profiles. The clusters revealed significant regional differences in freight generation, logistics infrastructure, transit potential, and investment characteristics. **Conclusions:** The proposed framework provides an evidence-based analytical tool for railway freight management, infrastructure planning, and the prioritization of regional development strategies while accounting for spatial connectivity constraints.

Keywords: transportation process; rail transport; transport clusters; PCA; MST; spatial clustering

Academic Editor: Robert Handfield

Received: 7 July 2026

Revised: 27 July 2026

Accepted: 4 August 2026

Published: 5 August 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

The territorial length and export and raw material economic profile of the Republic of Kazakhstan determine the strategic importance of the railway network. The efficiency of the transportation process significantly affects the stability of the national economy, since the network acts as the basis both for maintaining domestic supply chains and for integrating the domestic market into world trade. At the same time, the operational management of the transport system operates within the strict boundaries of industry legislation and rules for transportation and technical operation [1–3]. Such a framework restrains extensive network expansion, which requires searching for internal infrastructure reserves. In this context, the development of the railway sector can use political, business

and operational factors to improve efficiency [4]. However, the implementation of such strategies is complicated by pronounced regional and infrastructural imbalances, which lead to uneven loading of transport hubs and act as a barrier to inclusive economic growth. In the context of such spatial heterogeneity of cargo flows, traditional planning methods are not effective enough, which necessitates the use of spatial modeling methods. In this regard, the coordination of transport accessibility with the socio-economic indicators of the regions is becoming an important condition for minimizing logistics costs and increasing the overall technical efficiency of railway operators.

Within the framework of this methodological approach, the spatial organization of transport systems is considered as a complex of interconnected nodes [5], the analysis of which through clustering acts as an effective tool for identifying infrastructure imbalances [6]. The observed tendency towards power concentration at strategically important points determines the feasibility of using network analysis to identify hidden topological characteristics of the system [7,8]. Moreover, the definition of efficiency drivers within business clusters forms the basis for the integration of macroeconomic indicators in the transportation management model [9]. This approach is organically complemented by clustering methods, which provide deep segmentation of heterogeneous cargo flows and make it possible to objectively assess the significance of specific network nodes [10]. In summary, spatial analysis provides a solid foundation for transitioning from descriptive logistics models to proactive planning tools.

The development of spatial modeling methods is naturally based on graph theory, mathematical modeling and software, which makes it possible to maintain the stability of the transport network under dynamic changes in demand [11]. Such approaches provide the necessary routing flexibility [12] aimed at minimizing total logistics costs [13]. Among these mathematical tools, minimum spanning tree (MST) search algorithms are of particular practical value for optimizing the topology of the transport network. They allow you to design the configuration of the links so that the total network resistance, expressed in terms of the weights of the edges of the graph, is minimal. The validity of the application of the MST algorithm is supported by its ability to maintain stability of solutions in conditions of uncertainty, and makes it possible to form a reliable architecture of internal logistics links [14,15].

Unlike conventional k-means and hierarchical clustering methods, the proposed framework incorporates the physical railway connectivity constraint through the minimum spanning tree (MST), preventing the formation of geographically disconnected clusters. In contrast to existing railway corridor classification approaches, the proposed model simultaneously integrates economic, infrastructural, and transport-operational indicators within a unified PCA-based dimensionality reduction framework [16,17]. The proposed methodology therefore provides an analytically consistent basis for identifying functionally specialized regional railway freight clusters that can support differentiated railway freight management in Kazakhstan. An additional barrier is the lack of universal mechanisms for integrating economic, infrastructure and operational metrics into the clustering process. Despite the reasonable effectiveness of geospatial multicomponent systems [18], their introduction into the production activities of KTZh—Freight Transport LLP necessitates the formation of an adapted methodological approach.

However, despite these advances, an important methodological problem remains unresolved. Existing studies do not provide a unified framework that simultaneously integrates economic, infrastructural, and operational characteristics of railway regions while preserving the spatial connectivity of the railway network. As a result, regional freight management often relies on administrative boundaries or isolated performance indicators, which may not adequately reflect the functional structure of the railway system. Consequently, there is a need for a spatially constrained clustering model capable of

identifying functionally homogeneous railway regions and supporting evidence-based freight management decisions under real network connectivity constraints.

Despite the growing application of PCA, graph-based methods, and clustering techniques in transport research, existing studies generally analyze economic, infrastructural, or operational characteristics separately or ignore the physical topology of railway networks. Consequently, there is still no methodological framework that integrates multivariate regional indicators with railway connectivity constraints to identify functionally homogeneous freight regions in Kazakhstan. This research gap limits the applicability of existing clustering approaches for evidence-based railway freight management.

Based on the identified research problem, the objective of this study is to develop a spatially constrained PCA–MST-based clustering model for railway freight management in Kazakhstan. The proposed framework integrates economic, infrastructural, and operational indicators with railway network connectivity to identify functionally homogeneous transport regions that can support differentiated planning and resource allocation. The scientific novelty of the study lies in combining principal component analysis with a railway connectivity-constrained minimum spanning tree to construct spatially coherent transport clusters for decision support [19,20].

The scientific novelty of this study lies in the following contributions:

- Integration of PCA and MST within a spatially constrained clustering framework;
- Incorporation of railway network connectivity into the clustering procedure;
- Simultaneous analysis of economic, infrastructural, and operational regional indicators;
- Development of a methodological framework for regional railway freight management in Kazakhstan.

2. Materials and Methods

The transition to spatial management necessitates the formalization of the parameters of the transport system. In this study, the railway network of the Republic of Kazakhstan is conceptualized as a complex spatial structure, mathematically represented as a weighted undirected graph $G = (V, E)$ [5,21]. This representation enables the evaluation of railway network connectivity rather than isolated station characteristics. The set of vertices (V) represents cargo-forming regions, and the set of edges (E) represents railroad tracks. For the purposes of spatial clustering, the railway network is represented as a connected weighted undirected graph. Although railway operations involve directional train movements, the graph used in this study represents only the physical connectivity between neighboring regions. Therefore, the existence of a railway connection between two adjacent regions is treated as a symmetric relation, which satisfies the assumptions of the minimum spanning tree (MST) algorithm.

The regional railway network was represented as a connected weighted undirected graph, where an edge was established between two regions if at least one direct railway connection existed between them. When multiple railway lines connected the same pair of regions, they were represented by a single edge in the adjacency matrix, thereby preserving only the existence of physical connectivity rather than the number of parallel routes.

Vertices correspond to the 17 administrative regions included in the analysis, while an edge exists only if two regions are directly connected by the railway network. The weight of each admissible edge is defined as the Euclidean distance between the corresponding regional observations in the reduced PCA feature space. The railway adjacency matrix is then used to constrain the set of admissible edges during MST construction, ensuring that the minimum spanning tree is built only over physically existing railway connections.

Although railway operations involve directional train movements, the graph adopted in this study represents only the physical connectivity of the railway network. Therefore, railway connections are treated as symmetric, satisfying the assumptions of the minimum spanning tree algorithm. Since the network topology is the primary factor determining the route configuration and logistics costs, the proposed graph model takes into account the operational limitations of the infrastructure of KTZh—Freight Transportation LLP.

The empirical base of the study was an array of data for the period 2022–2024, formed on the basis of the registers of the Bureau of National Statistics, the Ministry of Transport of the Republic of Kazakhstan and corporate reporting of the national operator [22–24]. Data were grouped into three interrelated analytical blocks.

The dataset covers the period 2022–2024; however, each region is represented by a single observation composed of the most recent available value or an aggregated indicator (e.g., average annual growth rates) depending on the nature of the variable. This approach ensured temporal comparability while avoiding duplication of regional observations in the PCA dataset.

The first, the economic block, determines the macroeconomic potential of cargo formation in each region (Table 1). The spatial units of analysis correspond to the 17 administrative regions of Kazakhstan, representing the official territorial level for which consistent economic, infrastructural, and transport statistics are available. Major urban freight nodes, including Astana, Almaty, and Shymkent, were not treated as separate observations. Instead, their economic and transport characteristics are incorporated into the corresponding regional statistics published by the official national statistical sources, ensuring consistency and comparability across all spatial units. This aggregation approach avoids inconsistencies associated with combining regional and city-level observations within the same multivariate dataset.

Table 1. Economic block of the database.

Region	Gross Regional Product (GRP), Million Tenge	Industry Share, %	Share of Agriculture, Forestry and Fisheries, %	Share of Construction, %	Share of Transport and Warehousing, %	Share of Transport and Warehousing in Fixed Capital Investments, %	Average Annual Growth rate of Investments in Fixed Assets of Transport and Warehousing, %
Abay Region	5,087,752.9	43.5	10.3	7.9	4.0	20.6	96.8
Akmola region	22,705,137.7	19.8	4.6	8.4	12.7	23.6	44.8
Aktobe region	7,548,720.3	42.1	6.1	6.9	7.4	14.0	42.6
Almaty region	38,647,788.6	11.6	2.2	4.5	11.0	16.0	23.3
Atyrau region	25,487,683.6	47.4	0.6	6.8	3.5	9.6	3.0
East Kazakhstan region	7,463,228.5	46.8	7.9	7.1	2.6	11.7	47.2
Zhambyl region	4,950,019.6	22.0	12.4	9.5	13.2	10.6	50.1
Zhetysay Region	2,987,727.0	12.0	17.4	16.0	7.2	40.3	111.2
Karaganda region	14,158,822.4	53.0	3.8	6.0	3.4	16.6	144.1
Kostanay region	7,437,694.9	39.5	12.3	6.2	5.5	9.9	36.6
Kyzylorda region	4,137,287.1	27.8	6.3	8.4	9.5	16.4	25.9
Mangystau region	6,916,229.9	49.3	0.7	7.8	8.5	19.6	35.3
North Kazakhstan region	4,055,197.1	19.0	26.6	6.8	7.4	25.9	87.5
Pavlodar region	8,248,173.8	49.2	6.4	6.4	4.7	6.4	56.1
Turkestan region	12,811,701.2	19.7	9.7	8.8	8.3	10.0	59.4

Ulytau Region	3,887,335.3	73.7	2.4	5.2	4.0	8.8	−6.4
West Kazakhstan region	7,350,415.9	53.2	4.6	6.5	3.6	13.4	5.0

The second, the infrastructure block, reveals the physical capabilities of the railway network to meet logistical demand (Table 2).

Table 2. Database infrastructure unit.

Region	Track Operating Length, km	Track Density, km/1000 km ²	Border Stations	Transport and Logistics Centers (TLC)	Engel Coefficient	Temporary Storage Ware- houses (TSW), m ²	Grain Warehouses, Places	Petroleum Product Depots, Locations	Warehouses of Other Goods, Places	Warehouses of Vegetables and Fruits, Places	Warehouses Food, Places
Abay Region	802.8	0.004	1	1	0.0024	0	9	81	27	7	7
Akmola region	1565.8	0.011	0	2	0.0028	76,839	543	264	98	17	12
Aktobe region	1817.3	0.006	2	1	0.0034	41,031	25	44	30	4	1
Almaty region	510.0	0.005	0	1	0.0008	489,427	109	148	255	47	26
Atyrau region	742.3	0.006	1	1	0.0026	36,560	0	80	54	7	1
East Kazakhstan region	424.9	0.004	0	1	0.0016	12,311	130	17	119	133	9
Zhambyl region	1029.1	0.007	1	0	0.0025	8825	65	27	15	0	22
Zhetysu Region	872.6	0.007	2	1	0.0030	2520	20	58	25	0	6
Karaganda region	1677.0	0.007	0	0	0.0032	7035	23	178	35	212	11
Kostanay region	1272.2	0.006	3	1	0.0032	36,370	177	47	9	16	1
Kyzylorda region	870.9	0.004	0	1	0.0020	33,643	16	32	22	0	2
Mangystau region	1096.6	0.007	4	1	0.0031	13,728	18	83	45	2	0
North Kazakhstan re- gion	618.7	0.006	1	0	0.0027	0	85	14	9	0	6
Pavlodar region	766.0	0.006	1	1	0.0025	18,020	86	133	36	4	1
Turkestan region	551.6	0.024	2	2	0.0009	48,825	0	87	41	17	0
Ulytau Region	795.9	0.041	0	0	0.0039	0	0	0	2	0	0
West Kazakhstan region	424.9	0.027	1	1	0.0013	9028	1	96	61	0	10

The third, the transport and operational unit, reflects the actual performance of the railway network (Table 3).

Table 3. Transport and operational block of the database.

Region	Freight Turnover, mln t-km	Average Annual Growth Rate of Freight Turnover, %	Average Transporta- tion Dis- tance, km	Freight Traffic Intensity, mln t-km/km	Foreign Trade Bal- ance Ratio in Physi- cal Terms
Abay Region	12,919.32	11.32	825.35	16.09	0.629
Akmola region	62,907.05	12.55	827.00	40.18	0.470
Aktobe region	31,599.89	0.89	825.64	17.39	0.338
Almaty region	1,792,068	−10.36	827.65	35.14	−0.055
Atyrau region	15,883.02	7.39	769.77	21.40	0.981
East Kazakhstan region	5860.61	−13.56	825.94	13.79	−0.484
Zhambyl region	23,771.41	−4.64	827.24	23.10	0.463
Zhetysu Region	7162.22	9.42	822.52	8.21	−0.827

Karaganda region	29,832.20	7.75	799.63	17.79	0.756
Kostanay region	23,456.58	1.12	848.38	18.44	0.713
Kyzylorda region	19,187.04	3.64	826.29	22.03	0.823
Mangystau region	14,991.24	−3.20	764.47	13.67	0.869
North Kazakhstan re- gion	6649.12	−11.51	824.96	10.75	0.215
Pavlodar region	24,041.16	16.64	515.57	31.39	0.891
Turkestan region	14,186.02	7.37	644.03	25.72	−0.083
Ulytau Region	8175.75	−13.91	824.64	10.27	0.840
West Kazakhstan re- gion	4244.24	−5.74	825.42	9.99	0.828

During the preparation of the array, the initial set of 23 variables was optimized to 20 key figures by aggregating warehouse capacity data. It can be assumed that this aggregation reduces information redundancy and facilitates subsequent multivariate analysis while preserving the main characteristics of the original dataset.

Given the sensitivity of clustering algorithms to the scale of measurements, the direct use of different-sized indicators is recognized as methodically incorrect. To prevent distortion of the results, the initial array was standardized by the z-score method (Standard Score); as a result, all metrics are reduced to a single scale with zero mathematical expectation and a single standard deviation [25].

The formed matrix of variables is characterized by a high degree of multicollinearity. The use of arrays with a similar structure in clustering algorithms probably leads to distortion of the results [25].

To eliminate this problem, the Principal Component Analysis (PCA) method is used, which converts a set of correlated features into an abbreviated system of orthogonal, independent factors [26,27]. The method involves calculating the covariance matrix with the calculation of its eigenvectors and numbers. Vectors determine the direction of new axes in a multidimensional space with maximum variance, and eigenvalues reflect the amount of information explained [28]. The identified principal components (PCs) become new interpretable metrics for subsequent clustering.

During PCA, the optimal number of PCs is determined, since excessive reduction in dimension can lead to loss of significant data. For verification, the Kaiser criterion is used, which assumes the retention of factors with eigenvalues $\lambda > 1$ [27], and the analysis of the scree plot graph, which makes it possible to identify the point of structural inflection after which the contribution of the remaining components is recognized as information noise [27]. The threshold for the cumulative explained variance is set at least 70%, which is sufficient for social and economic research [27].

Since the initial extraction of the main components often forms a complex matrix of factor loads that makes it difficult to interpret latent variables, the method of orthogonal rotation of the Varimax axes was used [27,29]. This method maximizes the variance of squared factor loadings within each component, approximating the values of correlations of initial variables with factors to extremes [28]. Varimax rotation was applied to improve the interpretability of the principal components by achieving a simpler loading structure.

The next step is to cluster the vertices (regions) based on the resulting principal components. To prevent the formation of geographically disparate clusters, a binary matrix of railway connectivity is integrated into the model. It establishes a limitation: the unification of regions into a single cluster is allowed if there is a direct railway connection. This approach allows you to analyze real material flows, and not abstract Euclidean distances between administrative units.

The central segmentation tool is the Minimum Spanning Tree (MST) algorithm. Studies confirm that MST-based algorithms provide high computational efficiency and natural representation of data when clustering spatial networks [30]. The integration of backbone weight constraints into the spanning tree structure contributes to the accurate identification of stable transport systems under uncertainty [30,31].

The operation of the algorithm is an iterative dividing process. First, the weights of all possible edges of the graph are calculated, where the Euclidean metric in the space of previously allocated main components acts as a measure of distance. Then, the algorithm forms a single connected network with a minimum total mass of edges [32]. The final transport clusters are distinguished by removing a predetermined number of the heaviest edges of the spanning tree [32].

Algorithms based on the minimum spanning tree (MST) have high sensitivity to starting parameters, which necessitates objective mathematical verification of the optimal number of clusters (k). Exclusively heuristic determination of this parameter with a high probability leads to excessive generalization or unreasonable fragmentation of the transport network.

To support the selection of an appropriate clustering solution, three complementary cluster validation criteria were applied: the Elbow method, the Silhouette coefficient, and the Calinski–Harabasz index. These criteria were used to evaluate the relative quality of alternative clustering solutions within the analyzed dataset [33]:

- Elbow Method—used to determine the saturation point of the model based on the dynamics of intra-cluster distances
- Silhouette Coefficient—used to assess the clarity of the boundaries and the degree of isolation of clusters
- Calinski–Harabasz Index—used to verify the optimal balance between compactness and separability of groups

Software implementation of graph modeling, as well as calculations and route construction, is carried out using the Python 3.13 language in the Google Colab environment. The processing of graph structures taking into account topological constraints is carried out using the NetworkX library, and the integration of the connectivity matrix is carried out using the GeoPandas package. This software solution provides a high reproducibility of calculations and allows you to correctly synthesize a multidimensional feature space with a physical map of the railway network.

The analysis was implemented using Python, including the NetworkX and GeoPandas libraries. All preprocessing procedures, PCA, cluster analysis, MST construction, and spatial visualization were performed using fixed parameter settings described in the manuscript.

3. Results

Standardization of 20 heterogeneous variables using the Z-score standardization method made it possible to neutralize analytical distortions caused by different dimensions of indicators. The assessment of standardized metrics revealed significant spatial heterogeneity: exceeding the standard deviation of one in some parameters in a number of regions indicates local concentrations of transport and economic potential (Figure 1).

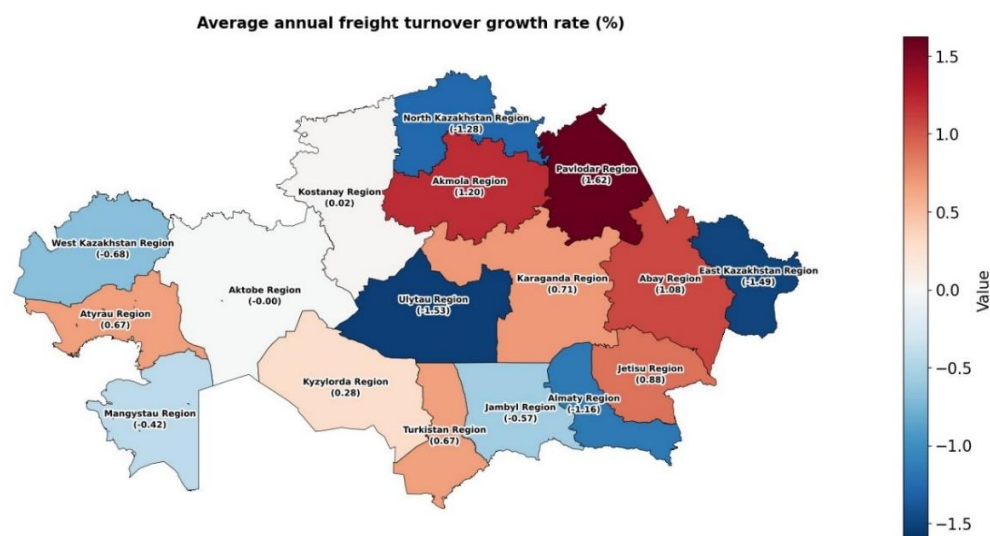


Figure 1. Data standardization results—choropleth map of average annual cargo turnover growth (%).

In addition, the correlation matrix recorded a high degree of multicollinearity of the features, confirming the information redundancy of the raw data for clustering (Figure 2).

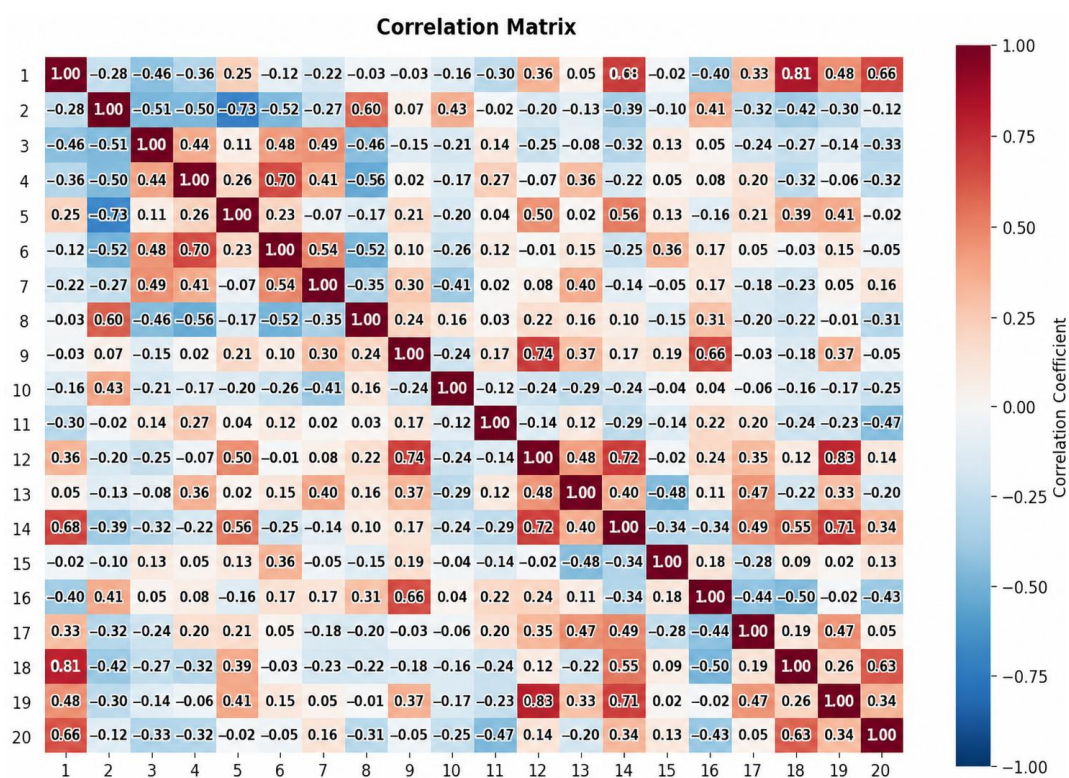


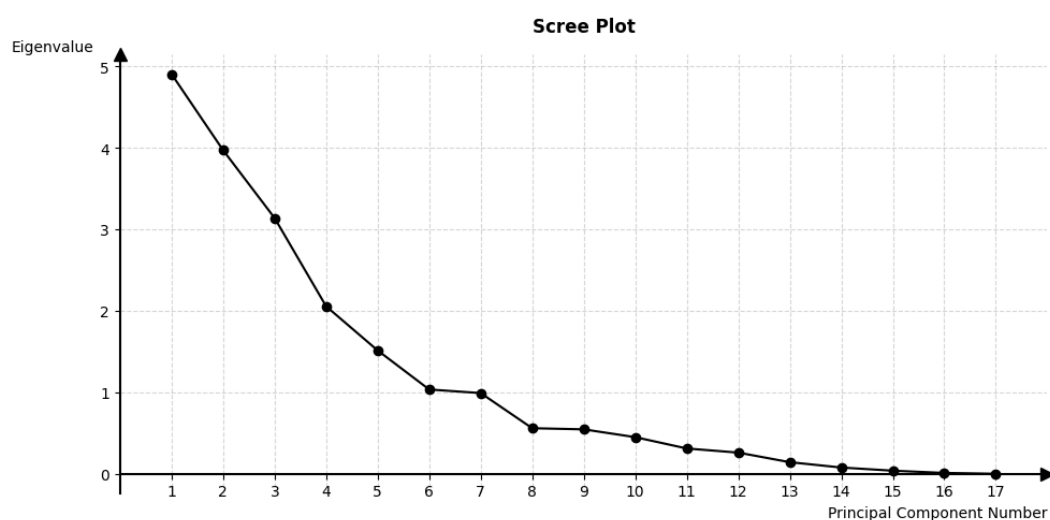
Figure 2. Data standardization results—correlation matrix.

The breakdown of indicators on the correlation matrix is presented in Table 4.

Table 4. List and coding of indicators.

Nº	Indicator Name	Units of Measure	Analysis Unit
1	Gross Regional Product (GRP)	million tenge	Economic
2	Share of industry in GRP structure	%	Economic
3	Share of agriculture, forestry and fisheries in GRP	%	Economic
4	Share of construction in GRP structure	%	Economic
5	Share of transport and warehousing in GRP structure	%	Economic
6	Share of transport and warehousing in investments	%	Economic
7	Average annual growth rate of investments in fixed assets of transport	%	Economic
8	Foreign Trade Balance Ratio	units	Economic
9	Operational length of railway tracks	km	Infrastructural
10	Density of the railway network	km/1000 km ²	Infrastructural
11	Number of border railway stations	units	Infrastructural
12	Rail freight turnover	mln t-km	Transport and operational
13	Average annual growth rate of freight turnover	%	Transport and operational
14	Traffic intensity on the network	mln t-km/km	Transport and operational
15	Average cargo transportation range	km	Transport and operational
16	Engel coefficient (transport connectivity)	units	Infrastructural
17	Availability of transport and logistics centers (TLC)	units	Infrastructural
18	Total area of temporary storage facilities	m ²	Infrastructural
19	Raw material warehouses (grain and oil products)	units	Infrastructural
20	“Last mile” warehouses (food, other goods)	units	Infrastructural

The elimination of the multicollinearity problem by the principal component method (PCA) provided a mathematically justified reduction in the dimensionality of factor space. Based on the Kaiser test ($\lambda > 1$) and analysis of the scree plot, five main components were retained for further modeling (Figure 3).

**Figure 3.** Scree graph (Scree plot).

The total explained variance reached 77.9%, while the remaining 22.1% may reflect local factors, random variation, or other sources of unexplained heterogeneity that were not captured by the retained components [34]. Varimax orthogonal rotation was subsequently applied to improve the interpretability of the factor structure by simplifying the distribution of variable loadings across components (Figure 4).

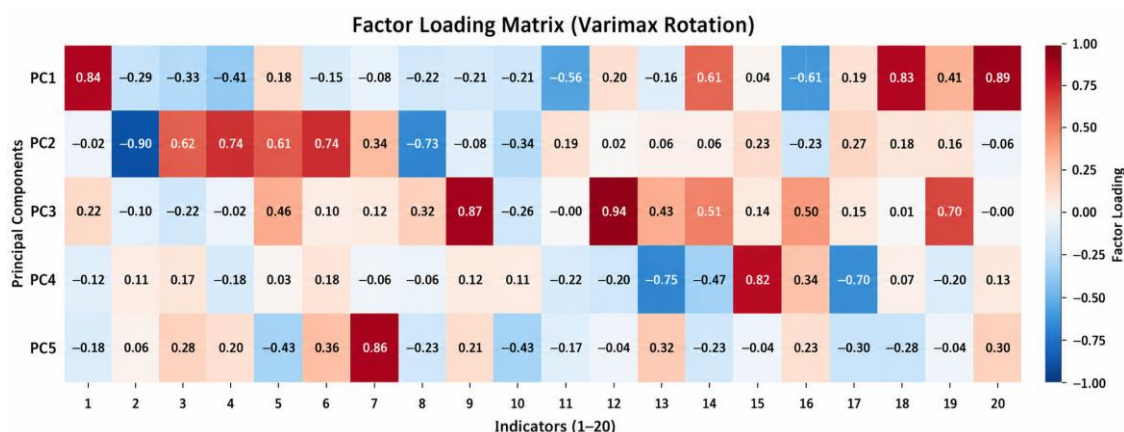


Figure 4. Matrix of factor loads.

Reducing the dimension of the factor space made it possible to move from the analysis of the initial correlated indicators to the study of stable interregional relationships based on metrically agreed components. As a result of the redistribution of information weight, it became possible to interpret each of the five latent variables. The orthogonal space formed provided a statistically robust basis for subsequent MST construction and identification of transport clusters.

The determination of the optimal number of clusters (k) was based on cross-analysis of validity indices (Figure 5, Table 5).

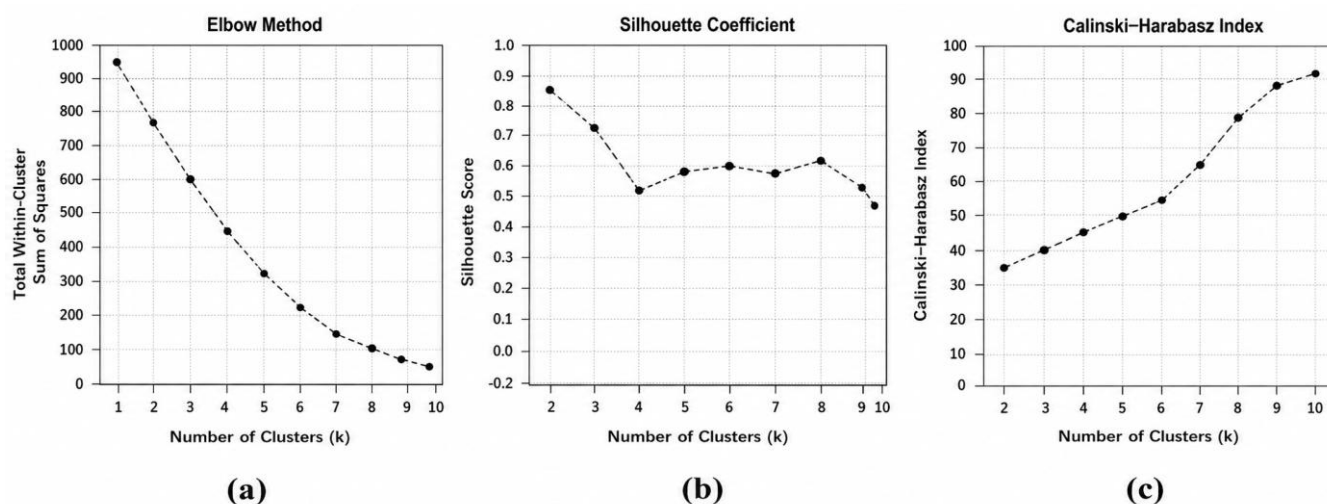


Figure 5. The result of determining the number (k): (a) the elbow method; (b) silhouette coefficient; (c) Calinski–Harabasz index.

Table 5. Internal validation indices for different numbers of clusters.

k	Silhouette Coefficient	Inertia (Elbow)	Calinski–Harabasz
3	0.32	160	3.6
4	0.27	125	4.2
5	0.16	100	4.6
6	0.21	70	5.8
7	0.21	52	6.5
8	0.20	40	7.1
9	0.22	30	7.6
10	0.16	22	8.0

As shown in Table 5, the three validation indices exhibit different trends across the tested values of k . Therefore, the final selection of six clusters was based on a combination of internal validation results, preservation of railway connectivity, and practical interpretability of the resulting regional structure [35,36].

The elbow method recorded the point of structural inflection of the intracluster dispersion at $k = 4$, suggesting a relatively compact clustering solution (Figure 5a). However, the Kalinsky–Kharabash index increased substantially from 4.56 to 5.73 when moving to $k = 6$ (Figure 5c), while the silhouette coefficient showed a partial recovery to 0.207 after the decline observed for the previous solution (Figure 5b). Taken together, these indices did not indicate a single universally optimal number of clusters but rather suggested several plausible partitioning alternatives.

Solutions with fewer clusters (e.g., $k = 4$ and $k = 5$) produced more aggregated regional groups, combining regions with distinct economic, infrastructural, and operational characteristics. Although these solutions were more compact according to some validation criteria, they provided less differentiation for practical railway freight management. In contrast, the six-cluster solution better preserved regional heterogeneity and enabled a clearer functional interpretation of the identified transport clusters. Therefore, the selected partition was considered more suitable for the intended analytical application.

Comparative analysis of the candidate clustering solutions showed that the configuration with $k = 4$ resulted in a relatively aggregated partition in which several regions with different transport characteristics were grouped together, reducing the interpretability of the resulting clusters. The six-cluster configuration provided a more differentiated regional structure while preserving railway connectivity constraints and facilitating the functional interpretation of the identified transport clusters. Therefore, the selection of $k = 6$ was based not only on the internal validation indices but also on considerations of regional interpretability and the intended application of the clustering results for railway freight management (Figure 6). Accordingly, the selected solution should be regarded as a practically justified partition rather than the unique statistically optimal configuration.

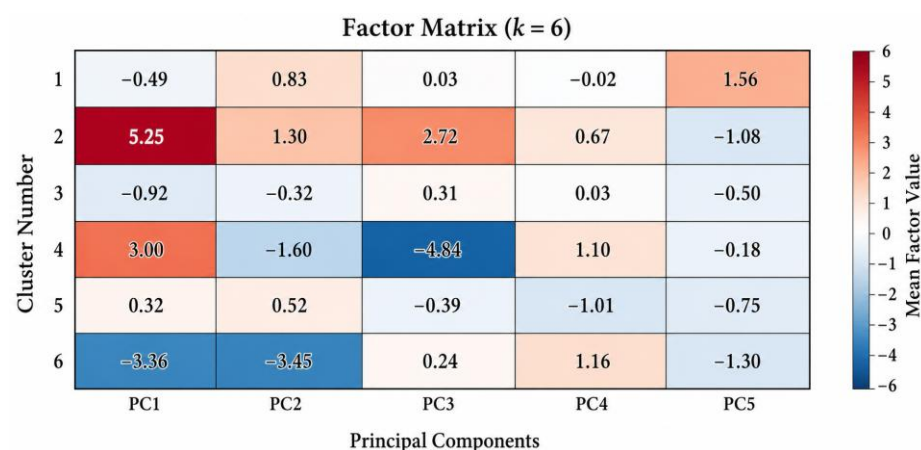
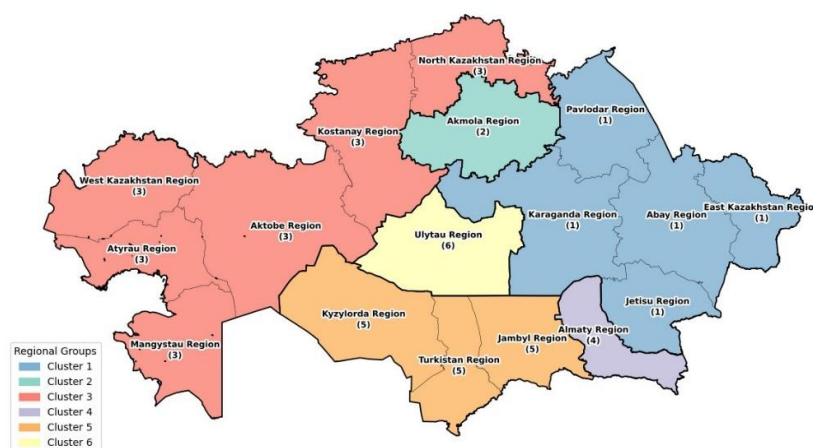


Figure 6. Cluster centroids in the PCA component space.

The spatial distribution of the formed regional clusters is shown in Figure 7.

Results of Regional Spatial Clustering of the Republic of Kazakhstan (K=6)

**Figure 7.** Map of received clusters.

The analysis of the formed structure confirms the presence of pronounced spatial heterogeneity of the transportation process and differences in the level of transport and infrastructure development of the regions. The use of PCA ensured the elimination of information redundancy of the initial data, while the use of MST made it possible to identify stable connections between regions and form transport clusters that reflect the structure of logistics interactions in the railway network of KTZh—Freight Transportation LLP. The obtained results create an analytical basis for the development of differentiated solutions focused on improving the efficiency of the transportation process.

4. Discussion

The transition from analysis to interpretation of the obtained data requires an understanding of the physical nature of the identified main components. Statistical processing of an array of twenty indicators revealed five latent factors that together explain the bulk of the variance of the system under study. The use of PCA in this case is due to the need to decompose the complex structure of the transportation process into independent measurements. This avoids redundancy in the description of regions and focuses on fundamental drivers that determine the typology of regional clusters.

The highlighted main components (PC) are synthetic indicators, behind which are specific technological and economic aspects of the activities of KTZh—Freight Transportation LLP. Since each component is a linear combination of initial features, the analysis of their factor loads makes it possible to determine the dominant specialization of the regions. This approach makes it possible to classify network nodes not according to formal administrative criteria, but according to the actual role in ensuring cargo flows, be it mass cargo generation, transit potential or infrastructure equipment.

A detailed consideration of the composition of each of the five components allows them to be given the following meaningful interpretation. Cluster specialization was assigned based on the dominant principal component scores together with the corresponding economic, infrastructural, and transport-operational characteristics of the member regions.

The first main component (PC1), with the largest contribution to variance (24.5%), is interpreted as a factor of concentration of consumption and service logistics. The dominant factor loads are in the warehouse infrastructure of the last mile (0.89), GRP (0.84) and TSF area (0.83). In combination with high traffic intensity (0.61) and negative values for the number of border stations (−0.56) and Engel coefficient (−0.61), the PC1 describes economically developed internal distribution centers. In these nodes, transport operations are

closely integrated with the terminal functions of consolidation and customs processing, while the accumulative functions of raw materials warehouses (0.41) play only a secondary role. Consequently, PC1 reflects the concentration of economic activity, warehouse infrastructure and service and distribution functions in regions with high domestic demand.

The second main component (PC2) has a slightly lower contribution (19.9%), which highlights the factor of infrastructure development with a mixed sectoral structure. The inverse relationship with the share of industry in GRP (−0.90) and the trade balance ratio (−0.73) is compensated by the significant contribution of the share of construction (0.74) and agriculture (0.62), as well as the share of transport in investments (0.74) and GRP (0.61). This allows us to conclude that high PC2 values are characteristic of regions in the economic structure of which industrial specialization plays a relatively smaller role. Presumably, such regions are more characterized by internal, infrastructure and intersectoral transport links than export and raw material specialization.

The third main component (PC3) with a contribution of 15.7% is an indicator of the trunk potential and mass cargo flows. The maximum loads in terms of freight turnover (0.94) and operational length of tracks (0.87) in combination with the capacity of raw material warehouses (0.70), freight traffic intensity (0.51) and Engel coefficient (0.50) describe the territories of transit and raw material specialization. PC3 records the ability of the regional system to implement large-scale transportation of mass goods over long distances with a high territorial network provision.

The fourth main component (PC4) with a contribution of 10.3% reflects the specifics of the range of transportation and logistics. Direct correlation with the average cargo transportation range (0.82) is accompanied by pronounced negative loads on the number of TLC (−0.70) and the average annual growth of cargo turnover (−0.75). This indicates the presence of extensive transport links: the transportation process relies on long routes, but its effectiveness is limited by the stagnation of the cargo base and the shortage of terminal infrastructure.

The fifth main component (PC5) has the smallest contribution (7.6%) and determines the investment dynamics of regional subsystems. The determining indicator is the average annual growth of capital investments (0.86), partially associated with the share of transport in investments (0.36) and an increase in cargo turnover (0.32). At the same time, negative loads in terms of the share of transport in GRP (−0.45) and the density of tracks (−0.43) indicate that the maximum investment activity is typical not for historically established industrial centers with a high network density, but for territories that are in the phase of active infrastructure renewal and expansion of logistics potential.

Based on the obtained five main components and their interpretation, it is possible to determine the specialization of the allocated transport clusters.

Cluster No. 1 (Aby, East Kazakhstan, Zhetysay, Karaganda and Pavlodar regions) acts as a zone of a mixed cargo base, where the development vector is determined by servicing diversified logistics flows. This is confirmed by positive PC2 values (0.83).

In contrast, Cluster No. 2 (Akmola region) functions as a central multimodal hub. The extreme values of PC1 (5.25) and PC3 (2.72) indicate a powerful synergistic effect of the concentration of warehouse infrastructure and mass cargo work.

Cluster No. 3 (Aktobe, Atyrau, Kostanay, Mangistau, North Kazakhstan, and West Kazakhstan regions) is a transit and export zone. The moderate influence of PC3 (0.31) against the background of negative values of other metrics characterizes the territory as focused on transit passage and mass export of raw materials.

A specific profile is demonstrated by Cluster No. 4 (Almaty region), which forms a terminal distribution cluster. The high concentration of PC1 (5.09) with a sharp negative

PC3 (−3.84) indicates a narrow specialization: acceptance, terminal processing and final distribution of imported goods.

The southern corridor of the domestic market is allocated to Cluster No. 5 (Zhambyl, Kyzylorda, and Turkestan regions), where the negative PC4 value (−1.10) reflects the predominance of short-haul traffic to meet local agro-industrial demand.

Finally, Cluster No. 6 (Ulytau region) is isolated in the long-distance transportation zone. The high transportation range (PC4 = 1.16) contrasts here with low consumption (PC1 = −3.36) and a pronounced investment deficit (PC5 = −1.3), identifying the most vulnerable segment of the railway network.

The practical significance of the results obtained lies in the possibility of transition from unified network management to differentiated regulation of transport clusters, taking into account their functional specialization. The resulting cluster structure confirms the low efficiency of using average general network approaches to organizing the transportation process in conditions of high spatial heterogeneity of the railway network.

The main tool for improving efficiency within the identified transport clusters is the introduction of intra-cluster routing. This solution provides for the concentration of wagon flows at the most equipped reference stations within each cluster to form unit train services.

The implementation of the proposed approach is based on the use of the identified specialization of the regions. Consolidation of cargo flows at the intra-cluster level contributes to the emergence of economies of scale in the formation of trains of increased weight and length. The technological advantage of this solution is to minimize the volume of shunting work at associated technical stations and accelerate the turnover of the car. Therefore, network segmentation based on a hybrid clustering algorithm is a promising tool for optimizing the distribution of traction resources and increasing the productivity of the backbone infrastructure in conditions of high heterogeneity of the transportation process.

Limitations

This study has several limitations that should be acknowledged. First, the analysis was conducted using aggregated regional data for 17 administrative units of Kazakhstan, which limits the statistical power of multivariate analysis and the generalizability of the results. Second, the proposed clustering framework was evaluated using internal clustering validation criteria only. Sensitivity analyses, such as bootstrap resampling or leave-one-region-out validation, were beyond the scope of the present study and therefore the robustness of the identified cluster structure was not formally assessed. Third, the proposed framework provides analytical support for railway freight management rather than direct evidence of operational efficiency improvement. In addition, the dataset combines variables obtained from official statistical sources that may refer to slightly different reporting periods, while commodity-specific railway freight flows were not available for inclusion in the analysis. These factors may influence the detailed interpretation of regional transport specialization. Future studies should validate the stability of the identified clusters using resampling techniques, incorporate temporal datasets, and compare the proposed clustering framework with operational railway performance indicators.

Although z-score standardization reduces differences in variable scales, extreme regional values may still influence the PCA results. Alternative preprocessing approaches, such as logarithmic transformation, robust scaling, or winsorization, were beyond the scope of this study and should be investigated in future research.

5. Conclusions

The study made it possible to develop a topologically correct model of transport clusters, which provides an analytical basis for increasing the efficiency of the transportation process of KTZh—Freight Transportation LLP. Integration of the hybrid approach using graph modeling and the minimum spanning tree (MST) algorithm with the separation of synthesized metrics from a set of economic, infrastructure and operational indicators in the regional context of the country has proven its effectiveness in spatial segmentation of the railway network. As a result of this hybrid approach, the multidimensional data array was successfully transformed into six specialized transport clusters, taking into account the physical connectivity of the infrastructure.

Based on the results of structural modeling, the need to switch from averaged general network metrics of operational planning to differentiated control, taking into account the technological specifics of each allocated cluster, is justified. A practical solution is the introduction of intra-cluster routing, which involves the consolidation of wagon flows at the most equipped reference stations within the operational clusters to form unit train services. It has been established that taking into account the identified regional specialization creates a technological basis for minimizing the volume of technological operations and accelerating the wagon turnaround due to the formation of trains of increased mass and length.

Summarizing, the proposed methodology of spatial clustering and the technological solutions arising from it act as an objective tool for the rational distribution of traction resources and increasing the productivity of the railway network in conditions of its pronounced heterogeneity.

Future research should include sensitivity analyses based on bootstrap resampling and leave-one-region-out validation to assess the robustness of the identified cluster structure. In addition, integrating temporal freight data, operational railway performance indicators, and simulation-based evaluation would further strengthen the practical applicability of the proposed framework.

Author Contributions: Conceptualization, A.M. and A.I.; methodology, M.B.; software, G.B.; validation, A.M., A.I. and G.B.; formal analysis, D.K.; investigation, G.B.; resources, M.B.; data curation, A.I.; writing—original draft preparation, A.M.; writing—review and editing, G.B.; visualization, M.B.; supervision, D.K.; project administration, A.M.; funding acquisition, A.I. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable

Informed Consent Statement: Not applicable

Data Availability Statement: The data used in this study were compiled from publicly available statistical sources, including the Bureau of National Statistics of the Republic of Kazakhstan, JSC National Company Kazakhstan Temir Zholy, and the Ministry of Transport of the Republic of Kazakhstan. The processed dataset, together with variable definitions and methodological details required to reproduce the analysis, is available from the corresponding author upon reasonable request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Law of the Republic of Kazakhstan on Railway Transport. Available online: https://adilet.zan.kz/eng/docs/Z010000266_ (accessed on 12 May 2026).
2. Order of the Minister of Industry and Infrastructure Development of the Republic of Kazakhstan No. 612 of 2 August 2019 on Approval of the Rules for the Transportation of Goods by Rail. Available online: <https://adilet.zan.kz/eng/docs/V1900019188> (accessed on 12 May 2026).
3. Order of the Minister for Investment and Development of the Republic of Kazakhstan No. 544 of 30 April 2015 on Approval of the Rules for the Technical Operation of Railway Transport. Available online: <https://adilet.zan.kz/rus/docs/V1500011897> (accessed on 12 May 2026).
4. Warnapura, L.L.; Kumara, A.S. A Model for Assessing Rail Freight Performance Using Global Empirical Analysis to Guide Railway Development. *Future Transp.* **2026**, *6*, 33. <https://doi.org/10.3390/futuretransp6010033>.
5. Rodrigue, J.-P. *The Geography of Transport Systems*, 4th ed.; Routledge: London, UK, 2016. <https://doi.org/10.4324/9781315618159>.
6. Shafiq, M.; Rocha, H.; Couto, A.; Ferreira, S. A Clustering Approach for Analyzing Access to Public Transportation and Destinations. *Sustainability* **2024**, *16*, 6944. <https://doi.org/10.3390/su16166944>.
7. Zhao, X.; Miao, C. Research on the Spatial Pattern of the Logistics Industry Based on POI Data: A Case Study of Zhengzhou City. *Sustainability* **2023**, *15*, 15574. <https://doi.org/10.3390/su152115574>.
8. Lu, M.; Jimenez Perez, E.; Mason, K. Complex Network Analysis of China's Integrated Air–High-Speed Rail Network: Topological Characteristics, Centrality Measures, and Cluster Analysis. *J. Transp. Geogr.* **2025**, *127*, 104292. <https://doi.org/10.1016/j.jtrangeo.2025.104292>.
9. Zizka, M.; Rydvalova, P. *Innovation and Performance Drivers of Business Clusters: An Empirical Study*; Springer: Cham, Switzerland, 2021. <https://doi.org/10.1007/978-3-030-79907-6>.
10. Pan, R.; Yuan, Q.; Liu, C.; Cao, J.; Liang, X. Intelligent Functional Clustering and Spatial Interactions of Urban Freight System: A Data-Driven Framework for Decoding Heavy-Duty Truck Behavioral Heterogeneity. *Appl. Sci.* **2025**, *15*, 8337. <https://doi.org/10.3390/app15158337>.
11. Strzelczyk, A.; Guze, S. An Approach to the Analysis of Critical Elements of Transport and Logistics Networks Using Graph Theory. *TransNav Int. J. Mar. Navig. Saf. Sea Transp.* **2024**, *18*, 535–544. <https://doi.org/10.12716/1001.18.03.07>.
12. Zhang, X.; Adamatzky, A.; Chan, F.; Deng, Y.; Yang, H.; Yang, X.S.; Tsompanas, M.A.I.; Sirakoulis, G.C.; Mahadevan, S. A Biologically Inspired Network Design Model. *Sci. Rep.* **2015**, *5*, 10794. <https://doi.org/10.1038/srep10794>.
13. Crainic, T.G.; Gendreau, M.; Gendron, B. *Network Design with Applications to Transportation and Logistics*; Springer: Cham, Switzerland, 2021. <https://doi.org/10.1007/978-3-030-64018-7>.
14. Hosseini, A. Uncertainty-Driven Stability Analysis of Minimum Spanning Tree under Multiple Risk Variations. *Mathematics* **2025**, *13*, 3100. <https://doi.org/10.3390/math13193100>.
15. Suguira, S.; Cheung, K.-F.; Bell, M.; Nakanishi, H.; Kurauchi, F.; Perera, S.; Vidyattama, Y. Urban Transit Network Design Using Spanning Tree: A Case Study of Canberra Transit Network. *arXiv* **2026**, arXiv:2604.15754. <https://doi.org/10.48550/arXiv.2604.15754>.
16. Chiang, T.-A.; Che, Z.-H.; Hung, C.-W. A K-Means Clustering and the Prim's Minimum Spanning Tree-Based Optimal Picking-List Consolidation and Assignment Methodology for Achieving Sustainable Warehouse Operations. *Sustainability* **2023**, *15*, 3544. <https://doi.org/10.3390/su15043544>.
17. Abdullayev, S.; Bakyt, G.; Kamzina, A.; Suleyeva, N.; Tokmurzina-Kobernyak, N. Dynamic Interaction of the TE-33A Diesel Locomotive and the Track in a Curve with a Radius of 600 Meters. *Int. J. Mech. Eng. Robot. Res.* **2024**, *13*, 205–212. <https://doi.org/10.18178/ijmerr.13.2.205-212>.
18. Boyko, N.I.; Salanchii, T.O. Development of Innovative Approaches for Network Optimization Using Geospatial Multi-Component Systems. *Radio Electron. Comput. Sci. Control* **2025**, *2*, 182–195. <https://doi.org/10.15588/1607-3274-2025-2-16>.
19. Liu, S.; He, N.; Cao, X.; Li, G.; Jian, M. Logistics Cluster and Its Future Development: A Comprehensive Research Review. *Transp. Res. Part E Logist. Transp. Rev.* **2022**, *168*, 102974. <https://doi.org/10.1016/j.tre.2022.102974>.
20. Regal, A.; Gonzalez-Feliu, J.; Rodriguez, M. A Spatio-Functional Logistics Profile Clustering Analysis Method for Metropolitan Areas. *Transp. Res. Part E Logist. Transp. Rev.* **2023**, *179*, 103312. <https://doi.org/10.1016/j.tre.2023.103312>.
21. Barthélemy, M. Spatial Networks. *Phys. Rep.* **2011**, *499*, 1–101. <https://doi.org/10.1016/j.physrep.2010.11.002>.
22. Bureau of National Statistics of the Agency for Strategic Planning and Reforms of the Republic of Kazakhstan. Transport Statistics. Available online: <https://stat.gov.kz/ru/industries/business-statistics/stat-transport/> (accessed on 12 May 2026).

23. JSC National Company Kazakhstan Temir Zholy. Integrated Annual Report 2024. Available online: https://railways.kz/media/filer_public/ab/5f/ab5fc04e-361f-4069-ab50-ec861622a8b8/godovoi_otchet_zh_2024_god.pdf (accessed on 12 May 2026).
24. Ministry of Transport of the Republic of Kazakhstan. Available online: <https://www.gov.kz/memleket/entities/transport?lang=ru> (accessed on 12 May 2026).
25. Sato-Ilic, M. Cluster-Scaled Principal Component Analysis. *WIREs Comput. Stat.* **2022**, *14*, e1572. <https://doi.org/10.1002/wics.1572>.
26. Matsunaga, M. How to Factor-Analyze Your Data Right: Do's, Don'ts, and How-To's. *Int. J. Psychol. Res.* **2010**, *3*, 97–110.
27. Jolliffe, I.T.; Cadima, J. Principal Component Analysis: A Review and Recent Developments. *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.* **2016**, *374*, 20150202. <https://doi.org/10.1098/rsta.2015.0202>.
28. Lever, J.; Krzywinski, M.; Altman, N. Principal Component Analysis. *Nat. Methods* **2017**, *14*, 641–642. <https://doi.org/10.1038/nmeth.4346>.
29. Bakyt, G.; Jailaubekov, Y.; Abdullayev, S.; Ashirbayev, G.; Ashirbayeva, I. Assessment of Carbon Dioxide Emissions in Road Transport Using Exhaust Gas Cleaning Technology in the Republic of Kazakhstan. *Vibroeng. Procedia* **2023**, *48*, 87–92. <https://doi.org/10.21595/vp.2023.23163>.
30. Rohe, K.; Zeng, M. Vintage Factor Analysis with Varimax Performs Statistical Inference. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **2023**, *85*, 1037–1060. <https://doi.org/10.1093/jrsssb/qkad029>.
31. Gagolewski, M.; Cena, A.; Bartoszek, M.; Brzozowski, Ł. Clustering with Minimum Spanning Trees: How Good Can It Be? *J. Classif.* **2025**, *42*, 90–112. <https://doi.org/10.1007/s00357-024-09483-1>.
32. Hosseini, A. Uncertainty Modeling and Stability Assessment of Minimum Spanning Trees in Network Design. *Mathematics* **2024**, *12*, 3812. <https://doi.org/10.3390/math12233812>.
33. Vathy-Fogarassy, Á.; Abonyi, J. *Graph-Based Clustering and Data Visualization Algorithms*; Springer: London, UK, 2013. <https://doi.org/10.1007/978-1-4471-5158-6>.
34. Todeschini, R.; Ballabio, D.; Termopoli, V.; Consonni, V. Extended Multivariate Comparison of 68 Cluster Validity Indices: A Review. *Chemom. Intell. Lab. Syst.* **2024**, *251*, 105117. <https://doi.org/10.1016/j.chemolab.2024.105117>.
35. Barkund, S.H.; Sharma, A.; Bhapkar, H.R. Survey of Shortest Path Algorithms. *Int. J. Renew. Energy Exch.* **2022**, *11*, 46–57.
36. Qiu, Y.X.; Wen, D.; Qin, L.; Li, W.; Li, R.H. Efficient Shortest Path Counting on Large Road Networks. *Proc. VLDB Endow.* **2022**, *15*, 2098–2110. <https://doi.org/10.14778/3547305.3547315>.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.