Optimization of production and transport infrastructure based on cluster analysis methods

Oleg Moskvichev[1,*], *Sergei* Nikishchenkov[1], and *Elena* Moskvicheva[1]

[1] Samara State University of Transport, 443066, Svobody Street, 2V, Samara, Russian Federation

Abstract. In order to solve the problems of optimizing production and transportation systems, a clustering procedure for objects is suggested. The procedure is a universal methodology for dividing a set of objects into subsets with their centers possessing optimal properties. At the same time, the use of point proximity metrics used in cluster analysis models the minimization of distances during transportation. If the volume of produced/extracted containerisable products of a production point is considered as the "weight" of each point, than the problem of minimizing transportation costs can be solved as a problem of optimizing clusters and their centers. A set of analytical models has been developed to describe and optimize the choice of location and number of container terminals (CT) at the first level and container storage and distribution centers (CSDC) at the second level of a two-level terminal model and logistics infrastructure of the container transport system (CTS). New clustering algorithms are suggested to determine the locations of CT and CSDC based on the condition of minimizing transportation costs and creating a terminal and logistics infrastructure, taking into account given or random number of clusters.

1 Problem statement

The research concepts of transport systems have not undergone significant changes up to date and are based on the approach to transport as a supporting system of the country's economy.

Historically, the leading link in the Russian transport system is railway transport. Its effective functioning plays an exceptional role in creating the conditions for modernization, transition to an innovative development path and sustainable growth of the national economy. The transport component in the final price of the product is an important factor when selecting the carrier and transport type by the cargo owner and ultimately depends on the condition and development of the transport system, the availability of infrastructure, changes in tariffs, and the offered logistics service.

* Corresponding author: moskvichev063@yandex.ru

In order to identify the development reserves of the transport services sector, to improve the transport work quality and forecast development, the models of transport systems are created [1-10].

A number of scientists have proposed a functional classification of applied economic and analytical models for the analysis of transport systems [2]:

- models that determine transport and economic relations [4-6,11];
- organization models of the transportation process [8,10,12-14];
- functioning and interaction models of the elements of transport systems [1,3,7,9,15-19].

In this paper, in order to remove the restrictions associated with insufficient development of the transport infrastructure, as well as for the goals of further country's economy growth, to identify development reserves, a model of a container transport system that defines transport and economic relations will be considered based on the solution of the production-transport type problem. This model will allow solving issues related to the cost-effective distribution of produced (extracted) containerisable products between consumers, taking into account transport costs and the costs of developing terminal and logistics facilities, as well as other factors.

In order to solve the problem in a most efficient way, it is necessary to take into account the volumes of production and consumption with their given geoinformation parameters. The throughput capacity of transport systems, cost of transport services, the costs associated with the development of terminal and logistics facilities necessary for processing the planned cargo flow should also be considered. Accordingly, the optimality criterion will be the total costs associated with production, the development of the throughput capacities of transport facilities, and the transportation process. The following additional criteria can be taken into account when stating and solving the problem: the presence of international transport corridors passing through the territory of particular region; level of container appeal of the region; infrastructure readiness, etc.

As a rule, production and transportation models when solving are reduced to problems of the transport type (matrix and network transport problem), linear programming, integer programming, network planning [1-8,13]. However, as analysis has shown, a feature of these problems is a large dimension, which does not always allow solving them in a nonlinear setting. These models are usually reduced by linear approximation to linear problems.

Solving optimization problems of choosing the location of terminal-logistic objects using graph models and mathematical programming under many combinatorial constraints leads to complex computational procedures of the exhaustive nature, which does not allow their use within the territories of the federal districts or the whole country. An analysis of existing practical methods [4-10, 13] showed that they do not take into account the number of objects, their location relative to industrial production, the volume of cargo flows from individual consignors and consignees, and the existing railroad topology.

In conjunction with the foregoing, in order to solve the problems of optimizing production and transportation systems, a procedure for clustering objects is proposed. The universal methodology for dividing a set of objects into subsets with their centers possessing optimal properties was implemented. According to the methodology, the use of proximity metrics of points used in clustering models the minimization of distances during transportation, and if the volume of produced/extracted products of a production point was adopted as the “weight” of each point, then the problem of minimizing transportation costs can be solved as a problem of optimizing clusters and their centers.

2 Two-tier model of container transport system

The conceptual model of the country's CTS is proposed in works [3,23,24]. The main idea is the formation of a two-tier infrastructure with service centers at each tier (Fig. 1). This will allow not only creating and selling customer-oriented transport products, such as organized container trains, for example, but also will predetermine a new stage in the development of CTS. A two-tiered CTS will allow concentrating the volumes of containerisable products necessary for the formation of container trains, excluding long periods of their accumulation, and also increasing the delivery speed of goods in containers.

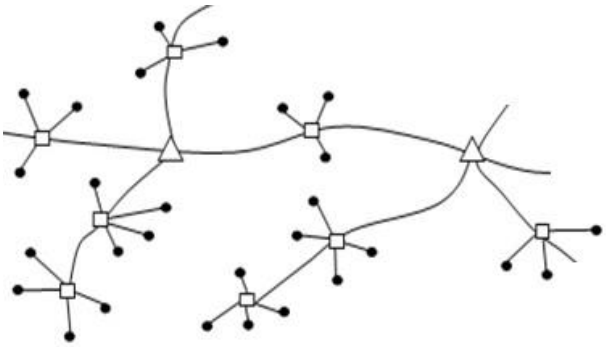

Fig. 1. Model of a two-tiered container transport system (points – producing units, triangles - CSDC, squares - CT).

At the 1st tier the CT network provides for the collection and concentration of goods from production points. Here the task of choosing CT locations arises (linking production points to their CTs). At the 2nd tier the CSDC network carries out the concentration of container flows received from the CT, where accelerated container trains going to other CSDC are formed. This also raises the problem of selecting the location of the CSDC from the optimality condition of some economic cost criterion.

The mathematical model of such two-tiered network presumes that the geographical coordinates of the production points (x_i, y_i), the cargo volumes of these industries v_i and the coordinates of the regional railway stations (x_j, y_j) are given. This allows stating the problem of the optimal location of CT and CSDC as a two-tiered clustering problem, where the criteria are the total costs of freight transportation and the costs of creating a network of CT and CSDC. In the current research, the idea of the famous McQueen k-means clustering algorithm was taken as basis [20], which, for a given number of clusters, finds clusters and their centers by the criterion of the minimum sum of squares of distances from points to their centers.

In order to solve the problem of choosing the location of CT or CSDC, it is necessary that the cluster centers are not at any geographical point, but necessarily in one of the given railway stations. Thus, a new algorithm called k-means pro is proposed, in which the resulting geometric center is projected to the nearest train station at each iteration.

3 K-means pro algorithm

Here is a description of the *k-means pro* algorithm. Let us adopt an initial set of objects J ($j=1,n$) to be clustered, characterized by their coordinates $X = \{x_1,\ldots,x_n\}$, their weights $V = \{v_1,\ldots,v_n\}$ and an admissible set of projections P ($r=1,p$) characterized by their coordinates $Y = \{y_1,\ldots,y_p\}$. Each j-th object and each r-th permissible projection point are

given in the G-dimensional space R^G, i.e., $x_j = (x_{j1},\ldots,x_{jG})$ and $y_r = (y_{r1},\ldots,y_{rG})$. Let us denote the division of the original set into k clusters as a set of subsets $S = \{S_1,...,S_k\}$.

The parameter k is given, which is the number of clusters the X set is divided into. As a result, it is necessary to obtain the optimal partition $S^* = \{S_1^*,...,S_k^*\}$, which centers are the optimal set of projections $C^* \subseteq Y$.

Let us denote the following: n is the number of clustering objects, p is the number of points of the admissible set of projections, i, i' is the cluster number, j is the number of the object, r is the point number of the the set of projections, l is the number of the point coordinate, m is the current iteration, G is the dimensionality of space in which clustering is performed.

The distance between the points t_1 and t_2 in a G-dimensional space according to the Euclidean metric is as follows:

$$d(t_1,t_2) = \sqrt{\sum_{l=1}^{G}(t_{1l}-t_{2l})^2}. \tag{1}$$

1. Select the initial division $S^0 = \{S_1^0,...,S_k^0\}$:

$$S_i^0 = \left\{x_{i1}^0,...,x_{in}^0\right\}, \bigcup_{i=1}^{k} S_i^0 = X, S_i^0 \cap S_{i'}^0 = \emptyset, i \neq i'. \tag{2}$$

2. For each m-th division $S^m = \{S_1^m,...,S_k^m\}$, starting from $S^0 = \{S_1^0,...,S_k^0\}$, a set of average vectors is calculated as $E^m = \{e_1^m,...,e_k^m\}$, i.e., $e_i^m = (e_{i1}^m,...,e_{iG}^m)$,

$$e_{il}^m = \frac{\sum_{j=1}^{n_i} v_j x_{jl}}{\sum_{j=1}^{n} v_j}, \tag{3}$$

where n_i is the number of points of i-th cluster.

3. Let us calculate the set of projections of means for the current partition:

$$C^m = \{y \in Y : \forall i, \quad d^*(y,e_i^m) = \min_{1\le r\le p} d(y_r,e_i^m)\}. \tag{4}$$

4. Let us calculate the partition C^m generated by the set and take it as $S^{m+1} = (S_1^{m+1},...,S_k^{m+1})$, i.e.

$$S_i^{m+1} = \left\{x \in X : d(x,c_i^m) = \min_{1\le i'\le k} d(x,c_{i'}^m)\right\}, \ 1 \le i \le k. \tag{5}$$

5. Checking-up: if $S^{m+1} \neq S^m$, then we proceed to the section 2, replacing m by $m + 1$, and if $S^{m+1} = S^m$, then the following is supposed: $S^m = S^*$, $C^m = C^*$. This is the end of the algorithm.

The optimization criterion in the classical k-means algorithm is the functional

$$F(S) = \sum_{i=1}^{k} \sum_{X\in S^i} \left\|X - e_i(S)\right\|^2. \tag{6}$$

The functional $F(S)$ does not increase on the sequence of divisions S^0, S^1,…,S^m,…, and $F(S^m) = F(S^{m+1})$ only if $S^m = S^{m+1}$. Therefore, after a finite number of steps, algorithm ends for any initial division S^0. In our case, the achieved optimization criterion for the found centers c_i^* has the following form:

$$F'(S) = \sum_{i=1}^{k} \sum_{X \in S^i} \left\| X - c_i^*(S) \right\| . \quad (7)$$

As a result of the algorithm operation, each time local minimum $F(S)$ is obtained, and the clustering result depends on the selection of the initial standard e^0. The e^0 coordinates can be obtained in various ways. For example, they can be taken as random numbers evenly distributed among the possible coordinates of the starting points. In order to check the stability of the results and to obtain various averaged dependencies, the choice of e^0 can be changed.

It should be noted that solving the 2nd tier problem, which is selecting the location of the CSDC, is also done using this clustering algorithm. Here the points for clustering are CTs, and the resulting cluster centers determine the places of developing the CSDC. For the final choice of locations for the CSDC, it is advisable to supplement the analysis of these already selected locations based on additional factors [23, 24].

It is natural to assume that the number of clusters is also unknown and K must be determined according to the condition of minimum costs.

It was shown in [23–25] that with an increase in K, the F value decreases sharply at first, and then more slowly. That is, transportation costs are reduced with an increase in the number of constructed CTs and CSDCs (Fig. 2).

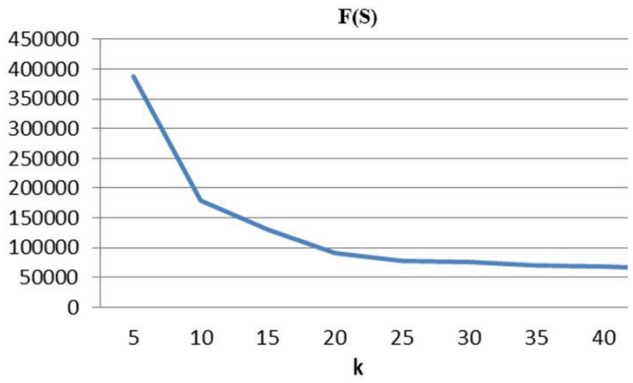

Fig. 2. The correlation between the F value and the number of clusters for the data from Volga Federal District of Russia.

But at the same time, the costs of developing CTs and CSDCs will increase.

The total costs associated with developing the first tier of the container transport system can be expressed as follows:

$$E(K) = E_1 + E_2 = \sum_{l=1}^{K} \sum_{i \in S_l} d_{il}\, v_i + ck \rightarrow \min , \quad (8)$$

where c is the standardized reduced cost of developing one CT.

A similar formula will be used for developing a second tier of CTS. The function *E(K)* is shown in Fig. 3.

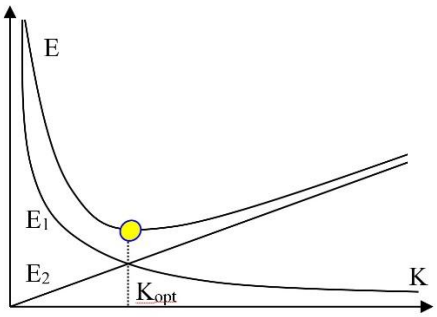

Fig. 3. The dependence of total costs on the number of objects in CTS terminal and logistics infrastructure.

A minimum of total costs can be obtained by varying the *K* value, for which it is necessary to solve the problem again for each *K* [25].

Based on the developed software, calculations were carried out for the production and transport complex of the Volga Federal District of Russia, and several specific tasks for creating a two-tiered CTS were considered.

4 Clusterization algorithm based on proximity matrix

Despite the great advantages of the cluster approach and the developed *k-means pro* algorithm (the mathematical validity of optimality, the low growth of computational complexity for large problem dimensionality), there is a drawback in the suggested model that makes it difficult to put into practice. The efficiency function uses the proximity measure during optimization as the Euclidean distance between two points *i* and *j*:

$$d_{ij}=\sqrt{(xi-xj)^2+(yi-yj)^2}\,. \tag{9}$$

The actual distance between the real production points and the point stations does not coincide with the Euclidean, because it does not take into account the tortuosity of roads, the presence of rivers, bridges, traffic congestion, the need for increased delivery speeds and more. The use of other metrics, in particular the "city-block distance" metric (Manhattan distance) is not suitable for road and rail traffic.

A discrete model for creating a CTS is developed and solution algorithms eliminating this drawback are suggested.

Let us assume that a two-tiered CTS is determined by the set of production points I with their numbers *i=(1,m)*, the volumes of containerisable products of production v_i, the set of points of stations J with numbers *j=(1,n)* and the proximity matrix D = $\{d_{ij}\}$. Each element of this matrix expresses the value of the criterion by which the costs (losses) are measured when transporting goods from *i* to *j*. In the simplest case, it can be distances in kilometers, but the whole variety of factors can be taken into account, affecting the costs of transportation from *i* to *j* (qualitative characteristics of the delivery route, non-linearity of the volume, delivery speed, seasonality, etc.). Further, the quantity *dij* will be called distance.

Then the task of choosing the locations of the centers (clusters) is stated as follows. The disjoint subsets of I set points, which are S_l clusters, need to be determined so that the total weighted distance from the production points to their cluster centers, belonging to the set of stations J, is minimal

$$F=\sum_{l=1}^{K}\sum_{i\epsilon S_l} d_{il}\, v_i \to \min\,, \tag{10}$$

where d_{il} is the distance between the i-th production point, which is part of the l-th cluster, and one of the points of the railway stations, which is the center of the l-th cluster.

The F value depends on dividing into clusters S_l and on selecting the numbers of railway stations (K in total), which will be the centers of the found clusters.

Let us consider a clustering algorithm based on the distance matrix D=$\{d_{ij}\}$ for each fixed K. Initial data for the algorithm operation are the following: the set of production point numbers I, i =*(1,m)*; the set of railway station points numbers J, j=*(1,n)*; distance matrix D=$\{d_{ij}\}$; "weights" of production points v_i, representing, for example, the volume of containerisable products produced/extracted by the i-th production; station distance matrix B=$\{b_{jj'}\}$.

Algorithm:

1. K stations are selected and declared as the centers at the first iteration, the standard. The choice of this first standard, strictly speaking, affects the result of the algorithm. In general, as suggested in [25], the composition of the standard can be randomly selected. (a more advanced method of developing a standard will be considered below).

2. The each i number is selected and the center from the set of centers m is determined, to which the distance d_{il} is minimal. Thus, all production points are tied to the centers and S_l clusters are obtained. l=*(1,K)*.

3. For each cluster l, a new center is found, for that, a railway station from the set J is found in each cluster for $i \in S_l$, for which the total weighted distance d_j is minimal

$$d_j = \sum_{i \in S_l} d_{ij} v_i \ . \tag{11}$$

Thus, new centers for each cluster are found.

4. Having ontained new centers, proceed to step 2 and obtain new clusters, etc. Repeat procedures 2,3,4 until the resulting clusters and their centers begin to repeat.

The end of the algorithm.

5 The algorithm for obtaining the first standard

Achieving the best option will be more systematic if the initial centers are as far apart as possible.

In order to improve convergence, the following algorithm for obtaining the first standard can be implemented.

Algorithm:

1. Choose a random number from the set of numbers of railway stations J and find another railway station, which is the farthest from the first largest d_{jj}. These two railway stations are already elements of the initial standard.

2. For each railway station that is not included in the standard, the nearest railway station from the standard can be found using d_{jj} value. Two clusters of stations are obtained.

3. In each such cluster the farthest railway station from the element of the standard is found and the station, at which the distance to the standard is the largest in the cluster, is included in the standard. The third standard railway station is determined.

4. We proceed to point 2,3 and so the 4th, 5th and k-th railway station of the standard are obtained.

The end of the algorithm.

Figure 4 shows the testing results of this algorithm in comparison with a random selection of standard. It can be seen from the presented correlation that, on average, the number of iterations of the main clustering algorithm decreases when using the algorithm for obtaining the first standard.

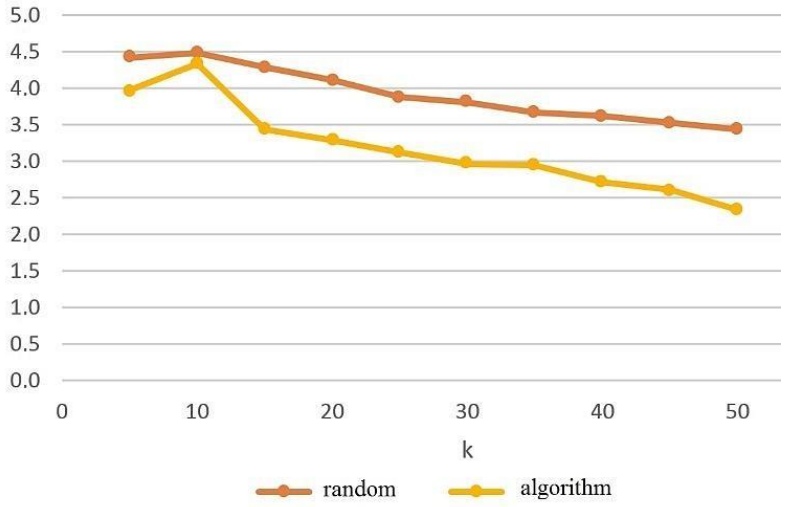

Fig. 4. The correlation between the average number of iterations and *k*.

The developed applied toolkit as a software product for designing the terminal and logistics infrastructure allows implementing the proposed models and algorithms for the optimal selection of the location for the CTS infrastructure for its organization and cost-effective functioning under various optimization criteria. It also allow receiving both tabular digital data for outline design and a graphic image of locations of designed terminal and logistics facilities on a territory map.

Thus, the proposed model and algorithms for solving the CTS optimization problem using cluster analysis methods for the long-term planning and development of the terminal and logistics infrastructure make it possible to solve the problems of optimizing the placement of CTS infrastructure facilities, determining their required quantity and capacity, taking into account minimization of the transporting cost for goods and investments.

References

1. P.A. Kozlov, N.A. Tushin, World of Transport, №2, 22-25 (2011)
2. B.A. Levin, E.A. Mamaev, V.V. Baginova, Science and Technology of Transport, № 4, 8-17 (2003)
3. S.M. Rezer, O.V. Moskvichev, E.E. Moskvicheva, Transport: Science, Technology, Management, № 7, 3-7 (2016)
4. P. Arnold, D. Peeters, I. Thomas, Transportation Research. Part E: Logistics and Transportation Review, **40(3)**, 255–270 (2004) doi:10.1016/j.tre.2003.08.005
5. C.-C. Lin, Y.-I. Chiang, and S.-W. Lin, Computers & Operations Research, **51**, 41–51 (2014) doi:10.1016/j.cor.2014.05.004
6. R. Wang, K. Yang, L. Yang, Z. Gao, Engineering Applications of Artificial Intelligence, **72**, 423–436 (2018) doi:10.1016/j.engappai.2018.04.022
7. M. Rabbani, S.M. Kazemi, International Journal of Industrial Engineering Computations, **6(3)**, 405–418 (2015) doi:10.5267/j.ijiec.2015.2.002
8. R. Ishfaq, C.R. Sox, European Journal of Operational Research, **210(2)**, 213–230 (2011) doi:10.1016/j.ejor.2010.09.017

9. K. Yang, L. Yang, Z. Gao, Information Sciences, **402**, 15–34 (2017) doi:10.1016/j.ins.2017.03.022
10. M. Zhalechian, R. Tavakkoli-Moghaddam, Y. Rahimi, Engineering Applications of Artificial Intelligence, **62**, 1–16 (2017) doi:10.1016/j.engappai.2017.03.006
11. V.G. Galaburda, World of Transport, № 1, 96-100 (2014)
12. A.V. Kirichenko, N.N. Maiorov, V.A. Fetisov, Vestnik GUMRF, № 5 (33), 26-33 (2015) doi: 10.21821/2309-5180-2015-7-5-26-33
13. A.G. Kirillova, Transport: Science, Technology, Management, № 10, 22-25 (2010)
14. A.L. Kuznetsov, S.S. Pavlenko, V.N. Shcherbakova-Sliusarenko, Vestnik GUMRF, № 5 (33), 33-42 (2015) doi: 10.21821/2309-5180-2015-7-5-33-42
15. P.A. Kozlov, World of trensport, № 2, 22-25 (2011)
16. P.A. Kozlov, I.V. Osokin, V. Iu. Permikin, World of Transport, № 5, 18-23 (2011)
17. A.L., Kuznetsov, A.V. Kirichenko, A.S. Tkachenko, G.B. Popov, Vestnik ASTU: Marine Engineering and Technology, № 1, 100-108 (2018) doi: 10.24143/2073-1574-2018-1-100-108
18. S.S. Pavlenko, Vestnik GUMRF, № 6 (34), 59-71 (2015)
19. V.M. Sai, D.I. Kochneva, Vestnik USURT, №4(32), 65-75 (2016) doi: 10.20291/2079-0392-2016-4-65-75
20. S.A. Aivazian, V.M. Bukhshtaber, I.S. Eniukov, et al., I.S., Finance and Statistics, 607, Moscow (1989)
21. B. Diuran, P. Odell, Statistics, 128, Moscow (1977)
22. B.G. Mirkin, High School of Economics, 88, Moscow (2011)
23. O.V. Moskvichev, Klientoorientirovannaia konteinernaia transportnaia sistema, 186, VINITI RAN (2018)
24. O.V.Moskvichev, World of Transpot, **15**, № 5 (72), 158-173 (2017)
25. B.A. Esipov, O.V. Moskvichev, et al., Advanced Information Technologies, 633-637, Scientific Center of the Russian Academy of Sciences, Samara (2017)