DOI: 10.15439/2023F9176
ISSN 2300-5963 ACSIS, Vol. 37

Evolutionary k-means Graph Clustering Method to Obtain the hub&spoke Structure for Warsaw Communication System

Barbara Mażbic-Kulma
0000-0002-7668-4694
Warsaw School of Information Technology
ul. Newelska 6
01-447 Warsaw, Poland
Email: kulma@wit.edu.pl

Jan W. Owsiński, Jarosław Stańczak, Aleksy Barski,
Krzysztof Sęp
0000-0002-2750-6584, 0000-0003-3722-0085
0000-0003-3746-1778, 0000-0002-1453-4148
Systems Research Institute
Polish Academy of Sciences
ul. Newelska 6 01-447 Warsaw, Poland
Email: {owsinski, stanczak, aleksy.barski,
sep}@ibspan.waw.pl}
Warsaw School of Information Technology
ul. Newelska 6
01-447 Warsaw, Poland

***Abstract*—The k-means method is one of the most frequently used clustering methods due to its efficiency and ease of modification and adaptation to the problem being solved. This paper presents modification of k-means method used for clustering in graphs. The method is presented on the example of generating the hub&spoke structure in the graph of public transport connections in Warsaw. Optimization of the public transport is one of the most important tasks for large cities. An efficient transport system is very important for its inhabitants. One of possible solutions is introducing the idea of hub&spoke to transport system. In this approach it is important to detect main stations, called hubs, which will create axes of high-speed connections (city trains, metro, high-speed trams), from which passengers can transfer to slower local connections to get to their rather close destinations. In the presented approach we propose to find locations of such main changeover stations using an evolutionary k-means algorithm.**

***Index Terms*—hub&spoke, evolutionary k-means algorithm, city transport system.**

I. Introduction

THE BASIC k-means algorithm became the starting point for the construction of many of its modifications adapted to different needs. In this paper we present its modification used for clustering in graphs, applied to obtain the hub&spoke structure ([2], [11]) of public transport system in Warsaw.

Obtaining high efficiency of urban transport is a very big challenge. This can be achieved by high financial outlays for building new fast connections (metro lines, fast trains or fast trams) or to some extent by optimizing the existing system. In this work we propose significantly cheaper approach, which requires (probably slow and well thought out) rearranging the transport in the city using the hub&spoke structure. The hub&spoke structure was successfully used in the 1970s to reorganize air traffic [6]. Currently, the air transport is so developed (or there are so many possibilities now) that this paradigm of connections is often abandoned (an example of which is the announced cessation of production of the A380 aircraft, designed mainly for the mass transportation of passengers between hub airports), which does not mean that the method will not be useful in other areas of transport. It seems that public transport in big cities can be such an application of the hub&spoke structure.

The properties and definitions of the hub&spoke structure were presented in works: [1], [7], [11], [12], [13] [14] and [16]. The basis of the idea of arranging urban transport as a structure hub&spoke is that individual parts of the city are connected by a network of fast means of transport connections, mainly rail, metro and fast trams. Selected stops of these fast means of transport, can become communication hubs where one can change to slower, local means of transport (buses, ordinary trams or even bicycles,...), but the whole journey usually lasts shorter, because the main burden of transport has been transferred to fast and high-capacity means of transport. Of course considered city should have such a fast means of transport.

The proposed ideas for changing the concept of urban transport do not assume a revolutionary removal of stops and connections (which may cause passenger protests), but rather a slow reorganization of the system so that it evolves towards a more effective hub&spoke structure, while maintaining many connections that break this structure due to the habits of users.

Our new approach to this problem is based on the described further evolutionary k-means method for graphs clustering (EKMG), which finds groups of strongly connected stops and designates a central one as a communication hub. The data for the EKMG algorithm come from the preprocessed timetable for the city transport system.

II. Basic Notions, Definitions and Algorithms

A. Clustering methods

Clustering is a disjunctive partitioning of set of data X, containing *n*-dimensional elements *x* into *p* nonempty subsets called clusters, containing elements similar in some sense or measure, while the elements belonging to different clusters should be highly dissimilar in the same sense or measure.

This aim can be obtained using many methods, one of the most commonly used is the k-means method ([4], [5] and [20]), which is presented in the Algorithm 1:

```
1. Choose the number of sought clusters.
2. Generate starting positions of cluster centroids
3. Calculate distances of all clustered objects to all cluster centroids.
4. Assign objects to clusters with the closest centroids.
5. Update cluster centroids as geometric centers of their clusters.
6. If the assignment of objects to clusters in the subsequent two steps does not change, then go to 7, else go to 3.
7. End.
```

Algorithm 1. The basic k-means method algorithm.

The properties of this algorithm strongly depend on the accepted minimized distance measure:

$$C_D = \Sigma_q \Sigma_{i \in Aq} d(x_i, x_q), \qquad (1)$$

where:

$d(.,.)$ - denotes the Euclidean (or different) distance;

x_i – clustered data items;

x_q – centroids (center or mean points) of clusters A_q, q=1,... p.

Very important parameter of this method is *p* – the number of clusters which is imposed by algorithm users but is not tuned by the method. Mentioned earlier and described more precisely in section 4 the EKMG method can deal with this problem.

B. Evolutionary algorithms

The standard evolutionary algorithm (EA) works as this is shown below ([2]):

```
1. Random initialization of the population of solutions.
2. Reproduction and modification of solutions using genetic operators.
3. Evaluation of obtained solutions.
4. Selection of individuals for the next generation.
5. If stop condition is not satisfied go to 2, else go to 6.
6. End.
```

Algorithm 2. The standard evolutionary algorithm scheme.

As it is known from further works ([9], [10]) this simple algorithm requires several improvements in order to work efficiently:

- the invention of a proper encoding of solutions,
- development of specialized genetic operators, appro priate for the accepted solution encoding (if standard ones are not proper),
- formulation of the fitness function to be optimized by the algorithm.

The stop condition is usually described by a certain number of iterations.

C. Basic graph notions

We treat the city transport system as a graph with stations as graph nodes and transport lines as edges, thus some basic notions from graph theory are presented here, following [21].

A graph is a pair $G = (V, E)$, where V is a non-empty set of vertices and E is a set of edges. Each edge is a pair of vertices $\{v_1, v_2\}$ with $v_1 \neq v_2$.

In our problem we can consider also a directed graph, which is an ordered pair $G = (V, A)$ where

- V is a set whose elements are called vertices or nodes;

- A is a set of ordered pairs of vertices, called arcs (directed edges).

A simple non-weighted graph can be described using a neighborhood matrix with elements a_{ij}, which describe the connection between vertices *i* and *j* of the graph, $a_{ij} \in \{0, 1\}$, 0 - no connection, 1 - presence of connection.

In our work we consider mainly generalization of the neighborhood matrix for weighted graphs, where elements a_{ij} describe not only the presence or no of the connection, but also its strength (for instance the capacity or travel time of connection).

A hub and spoke structure (proposed in [11] and [12]) is a graph $H_s = (G_h \cup G_s, E)$ where the subset G_h corresponds to at least a connected graph (of hubs) with the relevant subset of set E, each vertex of subset G_s (of spokes) has degree 1 and is connected exactly with one vertex from subset G_h.

III. Description of Warsaw's Transport System

The timetable describing the urban transport operation in Warsaw can be downloaded from https://www.wtp.waw.pl/rozklady-jazdy/ and ftp://rozklady.ztm.waw.pl. The public transportation system is presented in Fig. 1.

As it can be seen, this network is quite well developed and consists of:

- metro - 2 lines,
- high-speed city rail (SKM), suburban railway (WKD) and rail (KM) - 12 lines,
- trams - 26 lines,
- city, suburban and night buses - 303 lines,
- and over 10,000 stops for all mentioned means of transport.

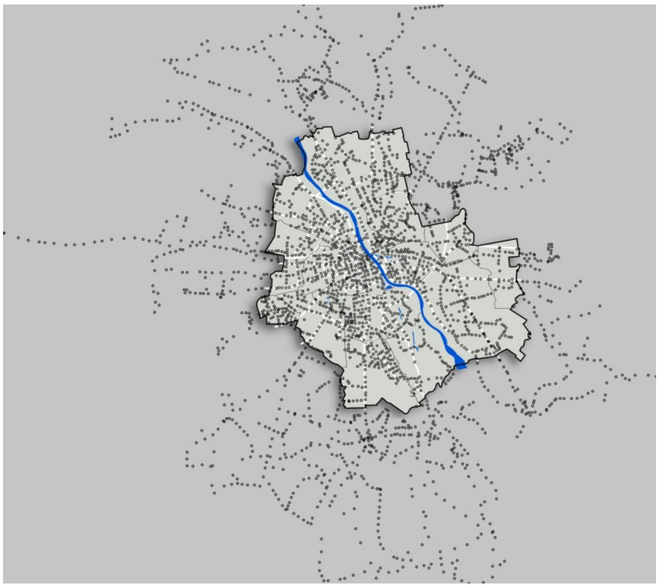

Fig. 1. Warsaw passenger transport system – 2840 unified stops (the state of the timetable as of November 25, 2021).

WTP (Warszawski Transport Publiczny, Warsaw Public Transport, the former abbreviation ZTM is still often used) conducts transport on most of the public transport lines in Warsaw. Other means of transport like private carriers and taxis were not included here due to the lack of possibility to know and to influence their transportation systems, also long-distance buses and trains were not considered to prepare the data for computations, because they are rarely used as urban mean of transport.

The public urban transport system has a large amount of about 10,000 stops, but for calculation purposes it has been reduced to 2,840 stops (graph nodes) as a result of combining into one stop stops in opposite directions and stops divided into "substops" in places with heavy passenger traffic (railway stations, bus terminals) or stops where different means of transport meet (for instance metro and bus or tram).

Our graph model of Warsaw transport system was built only on the basis of the timetable data taken from WTP/ZTM. In the constructed simplified graph of connections, we assumed that its vertices (communication stops), have a direct connection, as long as there is at least one running communication line that connects them. Thus the graph consists of overlapping blocks, because different communication lines have common stops.

The processed data obtained from the timetable may present several properties of the transportation system:

- presence or not the direct connections between the stops,
- frequency or the number of courses in a certain unit of time,
- travel time,
- potential capacity of means of transport in a certain unit of time (data about capacity of vehicles serving particular connections can be found in WTP/ZTM websites).

In the case of stops and connections common to many communication lines, the final values taken for computations are, in our case, appropriately modified (aggregated), so tha e.g. frequencies or capacity are added and the travel time is averaged in order to consider of connections from more lines at the same destination.

IV. Evolutionary k-means Method for Graphs Clustering (EKMG)

The EKMG method is based on evolutionary k-means method (EKM), which is described in detail in work [17]. In short words it can be summarized as follows:

1. Random initialization of the population of solutions (different centroids and numbers of clusters in solutions).

2. Reproduction and modification of solutions using genetic operators.

3. Evaluation of obtained solutions:

a) total minimized distance (2) is equal to infinity, the number of steps is equal to 0

b) take the number and centers of sought clusters from evaluated solution,

c) calculate distances (meant as in formula (3)) of all clustered objects to all cluster centroids,

d) assign objects to clusters with the closest centroids,

e) update cluster centroids as geometric centers of their clusters,

f) if calculated total distance for new data clustering (2) is less than calculated in previous step and number of steps is less than 5, then go to b).

g) the last value computed of the criterion (2) is the value of fitness function of the evaluated solution.

4. Selection of individuals for the next generation.

5. If stop condition of EA is not satisfied go to 2, else go to 6.

6. End.

Algorithm 3. The evolutionary k-means algorithm.

In this approach “solutions” are different instances of k-means algorithm with different numbers of clusters. Numbers of clusters and locations of their centroids can be modified by genetic operators of EA.

The minimized fitness function is similar to (1):

$$C_{Dr} = \Sigma_q \Sigma_{i\in Aq} d_r(x_i,x_q), \qquad (2)$$

where:

$d_r(.,.)$ - denotes a modified distance (Euclidean or different), described further by equations (3) and (4),

x_i– clustered data,

x_q – centroids of clusters A_q, $q = 1,... p_t$, the value of p_t (the number of clusters) is variable.

The modified distance $d_r(.,.)$, as used in (2), is calculated as follows:

$$d_r(x_i, x^q) = \begin{cases} d(x_i, x^q) & \text{if } d(x_i, x^q) \geqslant R \\ R & \text{if } d(x_i, x^q) < R \end{cases} \qquad (3)$$

and

$$R = (1 - r) \cdot d_{min}(x_i, x_j) + r \cdot d_{max}(x_i, x_j) \qquad (4)$$

where:

R – is the threshold value computed used threshold parameter r,

$d_{min}(x_i, x_j)$ - is the minimum value (but bigger than zero) of the accepted distance measure method among grouped different data items x_i, x_j,

$d_{max}(x_i, x_j)$ - is the maximum value of the accepted distance measure method among grouped data items x_i, x_j.

As it can be seen, the value of the threshold R is calculated on the basis of the properties of the grouped data and the given threshold parameter r, $r \in [0,1]$, which is meant to control the degree of detail of the clustering and indirectly the number of detected clusters. The threshold value also prevents the algorithm to find the trivial solution, where equation (2) is equal 0 and all data become centroids of their own one-data clusters.

The EKMG method is an application of EKM method to find clusters in graphs. The adjacency matrix of the graph is treated as a set of data about the attribute values of the nodes of the graph.

The algorithm of this method is presented in Algorithm 4:

1. Random initialization of the population of solutions: numbers of clusters and as centroids of clusters are randomly selected existing nodes of the graph.

2. Reproduction and modification of solutions (number and position of centroids) using genetic operators.

3. Evaluation of obtained solutions:

a) total minimized distance (2) is equal to infinity, the number of steps $s = 0$

b) take the number and centers of sought clusters from evaluated solution,

c) calculate distances (meant as $d(x_i,x_q)$) of all clustered objects to all cluster centroids,

d) assign objects to clusters with the closest centroids,

e) if $s < k$ update cluster centroids as graph nodes closest to computed geometric centers of their clusters,

f) if calculated total distance for new data clustering (2) is less than calculated in previous step and number of steps is less than k (k – the number of repetitions of k-means procedure, $k = 0, 1, 2$, bigger values too much slow down computations), then go to b),

g) the last value computed of the criterion (2) is the value of fitness function of the evaluated solution.

4. Selection of individuals for the next generation.

5. If stop condition of EA is not satisfied then go to 2, else go to 6.

6. End.

Algorithm 4. The evolutionary k-means algorithm for graph clustering.

The specialized evolutionary algorithm has in this case 4 genetic operators that modify solutions:

- the number of clusters – q (mutation like operator);
- values of cluster centers (random selection of new centroid among the nodes of the cluster) – A_q (mutation like operator);
- values of cluster centers (random selection of new cen troid among the nodes of the graph) – x_i (mutation like operator);
- uniform crossover (exchange of parameters between solutions).

The mechanism described in [15] was used to manage the genetic operators and select them to modify the solutions.

V. Results of Computer Simulations

New method of graph clustering was tested on Warsaw transport system data, using the time of travel and the capacity of connections as attributes of graph nodes. Simulations were conducted for different values of r parameter, equal 0.01, 0.05, 0.1, 0.3, 0.5, 0.7 and 0.9. Results with different numbers of detected hubs are presented in consecutive Table I, Fig. 2 and Fig. 3.

As you can see in Table I, the method usually selects about 24 hubs. For higher values of r imposed, the number of detected hubs starts to decrease, which is in line with the way the clustering method works: for bigger values of r, the method finds smaller number of more general clusters, for smaller values of r, the method finds bigger number of more detailed clusters. The function of the threshold parameter value r can be compared to the zoom function in a camera lens. Of course, there is no perfect proportion here, because the data parameters of the considered problem are also important and they affect the number and distribution of the clusters found.

Table I.
Numbers of clusters detected using the EKMG method depending on imposed r value

	Number of clusters detected	
r	Criterion: time of connections	Criterion: capacity of connections
0.01	27	24
0.05	25	24
0.10	24	23
0.30	24	25
0.50	25	23
0.70	20	12
0.90	2	4

In the domain of communication hubs the properties described earlier mean that hubs are stronger or weaker connected with their hubs.

Figures 2 - 5 show the results of computations on the map of Warsaw: larger points - hubs and smaller – spokes (ordinary stops) belonging to them, marked with the same color. Presented results are obtained for value of $r = 0.3$ and $r = 0.7$ respectively for the criterion of time and capacity. As it can be seen in the pictures, hub stations are mainly located in central, important communication points of the city.

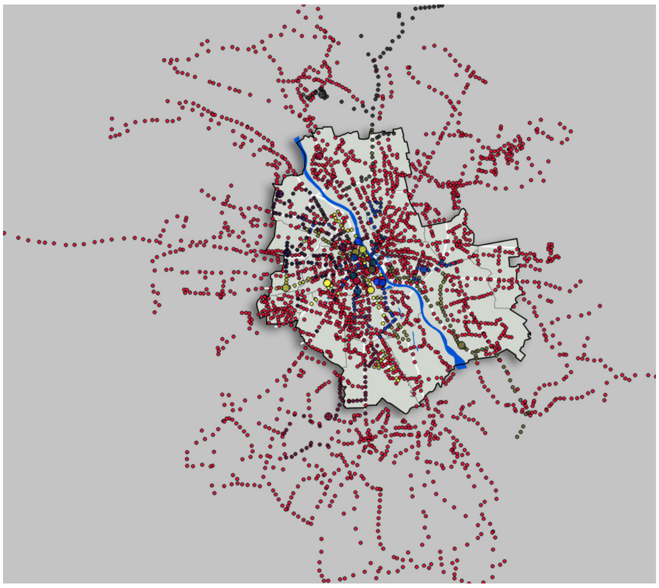

Fig. 2. Warsaw transport system with 24 hubs computed on the basis of time of connections for $r = 0.3$ (the state of the timetable as of November 25, 2021).

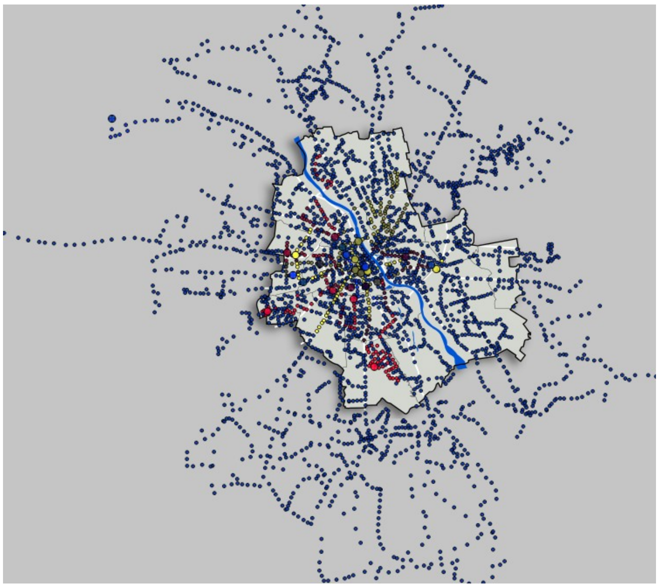

Fig. 3. Warsaw communication system with 25 hubs computed on the basis of capacity of connections for $r = 0.3$ (the state of the timetable as of November 25, 2021).

The method of finding transport hubs in the local transport system presented here is one of many possible ones, more information on other possible methods and results obtained for Warsaw can be found in the works: [13], [18] and [19]. Certainly, the stops indicated as potential hubs by several methods are definitely the best candidates for giving them such a function in reality.

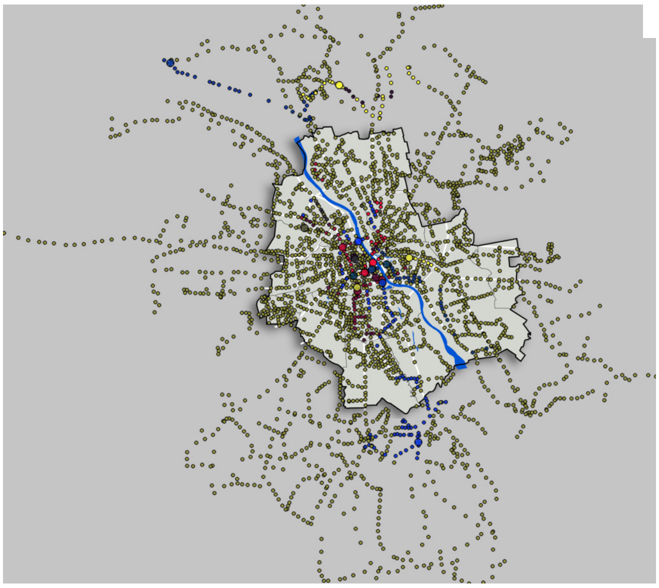

Fig. 4. Warsaw transport system with 20 hubs computed on the basis of time of connections for $r = 0.7$ (the state of the timetable as of November 25, 2021).

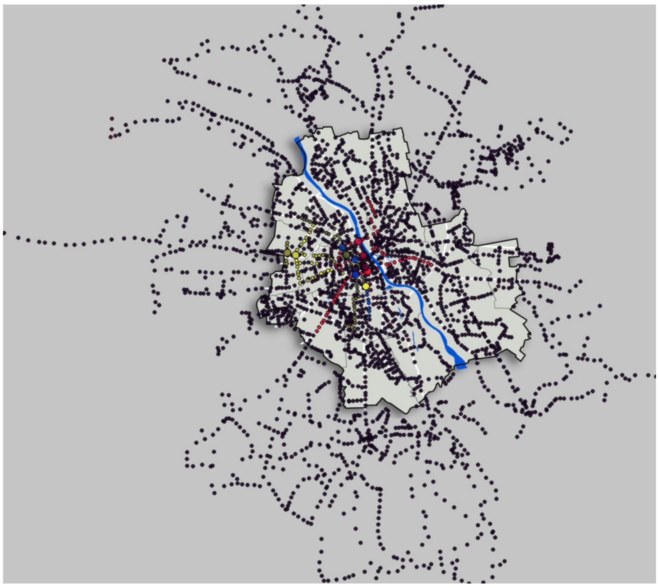

Fig. 5. Warsaw communication system with 12 hubs computed on the basis of capacity of connections for $r = 0.7$ (the state of the timetable as of November 25, 2021).

VI. Conclusions

This work deals with the possibility of applying the well-known k-means clustering algorithm in the problem of graph clustering with the possibility of improving the public transport system by using elements of the hub&spoke idea. The proposed specialized evolutionary method of the communication data processing is quite efficient and can deal with large sets of stops, characterizing big cities. We showed this feature on the example of Warsaw. As a result we obtained several solutions for different values of the threshold parameter r for the evolutionary k-means method for graphs clustering. The calculated transfer points are, of course, indicative proposals and actual transfer hubs may be created in slightly different places due to the influence of many other factors (e.g. existing buildings, land ownership), which the presented algorithm does not take into account, as only data on transport connections have been considered.

References

[1] J. Coyle, E. Bardi, and. R. Novack, "Transportation", Fourth Edition, New York: West Publishing Company, 1994.

[2] J. F. Campbell, and M O'Kelly, Twenty-Five Years of Hub Location Research, Transportation Science, 46(2), 2012, pp. 153-169.

[3] D. Goldberg, "Genetic Algorithms in Search, Optimization and Machine Learning", Addison-Wesley, Massachusetts, USA, 1989.

[4] S. Lloyd, "Least squares quantization in PCM" In: Bell Telephone Labs Memorandum, Murray Hill NJ, reprinted in: IEEE Trans. Information Theory IT-28, Vol. 2, 1982, pp. 129-137

[5] J. MacQueen, "Some methods for classification and analysis of multivariate observations", in: Proc. 5th Berkeley Symp. Math. Statistics and Probability. Vol. 1. 1967. pp. 281-297.

[6] T. Matisziw, Ch. Lee., and T. Grubesic, "An analysis of essential air service structure and performance", Journal of Air Transport Management 18, 1, January 2012, pp. 5–11.

[7] B. Mażbic-Kulma, H. Potrzebowski, J. Stańczak, and K. Sęp, "Evolutionary approach to solve hub-and-spoke problem using α-cliques", Evolutionary Computation and Global Optimization, Prace naukowe PW, Warszawa, 2008, pp. 121-130.

[8] B. Mażbic-Kulma, J. Owsiński, J. Stańczak, A. Barski and K. Sęp, Mathematical Conditions for Profitability of Simple Logistic System Transformation to the Hub and Spoke Structure, in: Atanassov, K., *et al.* Uncertainty and Imprecision in Decision Making and Decision Support: New Challenges, Solutions and Perspectives. IWIFSGN 2018. Advances in Intelligent Systems and Computing, vol. 1081. Springer, Cham., 2021, pp. 398-408.

[9] Z. Michalewicz, Genetic Algorithms + Data Structures = Evolution Programs, Springer Verlag, Berlin Heidelberg, 1996.

[10] Z. Michalewicz, and B. Fogel, How to Solve It: Modern Heuristics, Springer-Verlag, Berlin Heidelberg, 2004.

[11] M. O'Kelly, and D. Bryan, Interfacility interaction in models of hubs and spoke networks, Journal of Regional Science, 42 (1), 2002, pp. 145-165.

[12] M. O'Kelly, A quadratic integer program for the location of interacting hub facilities, European Journal of Operational Research, V. 32, 1987, pp. 392-404.

[13] J. Owsiński, J. Stańczak, A. Barski, and K. Sęp, Identifying main center access hubs in a city using capacity and time criteria. The evolutionary approach, Control and Cybernetics, 45(2), 2016, pp. 207-223.

[14] J.-P. Rodrigue The Geography of Transport Systems, New York: Routledge, 2020

[15] J. Stańczak, Biologically inspired methods for control of evolutionary algorithms, Control and Cybernetics, 32(2), 2003, pp. 411-433.

[16] J. Stańczak, H. Potrzebowski, and K. Sęp, Evolutionary approach to obtain graph covering by densely connected subgraphs, Control and Cybernetics, vol. 41, No. 3, 2011, pp. 80-107.

[17] J. Stańczak, and J. Owsiński, Evolutionary k-Means Clustering Method with Controlled Number of Clusters Applied to Determine the Typology of Polish Municipalities. In: Uncertainty and Imprecision in Decision Making and Decision Support: New Advances, Challenges, and Perspectives. IWIFSGN BOS/SOR 2020. Lecture Notes in Networks and Systems, vol. 338, 33. Springer, Cham, 2022, pp. 436-446.

[18] J. Stańczak, A. Barski, K. Sęp, and J. Owsiński, The problem of distribution of Park-And-Ride car parks in Warsaw, International Journal of Information and Management Sciences, 27(2), 2016, pp. 179-190, http://dx.doi.org/10.6186/1JIMS.2016.27.2.6.

[19] J. Stańczak, K. Sęp, and J. Owsiński, "Evolutionary methods for finding kernel & shell structures in a graph of connections" (in Polish: "Ewolucyjne metody znajdowania struktur typu "kernel & shell" w grafie połączeń"), Instytut Badań Systemowych PAN, Warszawa, 2023.

[20] H. Steinhaus, Sur la division des corps matériels en parties. Bulletin de l'Académie Polonaise des Sciences, Classe 3, 1956, 12, pp. 801-804.

[21] R. Wilson, Introduction to graph theory, Addison Wesley Longman, 1996.