Classification of Urban Focal Points Based on Cluster Analysis

Anton Zedgenizov¹, Evgenii Strizhakov¹

¹ Irkutsk National Research Technical University, Department of Oil and Gas Engineering, 664074, 83, Lermontov St. Irkutsk, Russian Federation

Abstract. The article presents cluster analysis algorithms used to assess traffic demand and to define the urban territories (focal points (FPs)) as separate clusters (groups) allowing proving/disproving the necessity of territories (FPs) merging on the basis of stated indicators with mathematical precision. The relevance of developing the classification of urban FPs within the context of basic parameters of their functioning is defined. The uniqueness boundary of FP is determined which is used as the base of classifying FPs. The cluster analysis is discussed, namely, its advantages and disadvantages. The goal of the cluster analysis is stated as applied to traffic demand assessment. The cluster analysis is applied to each of 6 enlarged types of territories (FPs): warehousing and utilities; industry; offices; education, healthcare, sport, culture, leisure; trading, public catering, service sector; housing sector.

1. Introduction

Current social and economic phenomena, technical progress level, patterns of population behavior in cities have a significant impact on the formation of factors associated with the development of traffic demand; it is especially evident in trips related to cultural and everyday goals [1, 2, 6]. At the same time, it is necessary to note some trends that significantly improve trips for business and even working purposes [18, 19, 20]. Currently, the development of communications and telecommunications allows for video meetings (videoconferences) that do not require a person’s physical presence. Vivid examples of these phenomena are distance education, on-line working meetings in the government of the Russian Federation and in large companies. Many people of creative occupations (designers, writers, scientists, etc.) who do not need complex large equipment for their work do not have to spend the whole working day in the office today. The results of their work can be presented by means of modern telecommunications. Therefore, nowadays, the study of diverse factors that affect the chance of a trip conditions the compulsory use of economic and statistical analysis, which has a wide range of methods, including factor and cluster ones.

Topicality of the Research

The determined algorithms of cluster analysis used to assess traffic demand and to define territories (FPs) as separate clusters (groups) involve the diverse and sufficient set of instruments, procedures and methods that allow proving/disproving the necessity of territories (FPs) merging on the basis of stated indicators with mathematical precision.
Problem Statement

An urban territory is a conglomeration of separate focal points, the combination of which is not always predictable [9, 11, 13]. All the diversity of urban FPs involving the characteristics of their functioning should be presented in the form of a parent entity of FPs; however, these FPs must be classified from the point of view of the features of their functioning. In particular, it can be noted that differences in the functioning of FPs may significantly influence the transport services of these FPs, loads on the street and road network and on public transport. Thus, taking into account the purpose of FPs, their irregular loads during the day, the average time of parking, specific generation of correspondences and other functioning parameters of FPs [5, 7, 10, 15, 17] they should be classified on the basis of cluster analysis.

Theoretical Background

As there are a large number of factors influencing the development of traffic demand only the most significant ones should be taken into account. Factor analysis allows for a more concise and simple description of the response function (dependent variable), as well as the structure and characteristics of relationships between them. For the first time cluster analysis was applied in sociology in 1939. Later, cluster analysis became the most widespread one in all fields of science due to its unique property – it allowed separating homogeneous groups (clusters) from an array of researched objects, that is, building a tree-based network according to the principle “from uniqueness to universality”. One of the main advantages of cluster analysis is its ability to perform analysis based on several criteria simultaneously, and these criteria can be very diverse. To study the case of classifying the factors influencing the volume of correspondences performed to the researched territory of FP, the following criteria may be used: specific generation of correspondence, average time of parking, remoteness of the territory (FP) from the main street, etc. The usage of cluster analysis together with regression-correlation analysis allows confirming the maximal similarity of the cases under consideration or proving the complete absence of correlations between the given cases.

At the same time, cluster analysis still has a number of drawbacks, for example, units of characteristics measurement are to be standardized and used with the same scale, otherwise, a characteristic having a larger absolute value will dominate the others (remoteness of the territory (FP) from the city center (more than 500 m.) will always dominate the average time of parking (more than 10 minutes)). There are some other disadvantages conditioned by the fact that the characteristics used for clustering provoke a “desirable” partition, which does not always coincide with the partitioning based on other random characteristics.

Thus, it is possible to state the main goal of cluster analysis as a whole and in relation to the tasks of assessing traffic demand in particular: to divide the multitude of objects \( G \) into \( m \) clusters relying on the array data of \( X \) so that each object belongs to only one cluster, while the objects belonging to different clusters were heterogeneous.

Cluster analysis like any other branch of scientific direction has its own terminology and features. The center of the cluster is the geometric average locus of points in the space of variables. The cluster radius is the maximum distance between the points and the cluster center. Metric is a quantitative assessment of similarities or differences between the classified objects [3, 4, 8].

Different indicators can be used as metric. Depending on the characteristics of the source data, the following indicators can be taken into account: linear distance, Euclidean distance, squared Euclidean distance, generalized power Minkowski distance, Chebyshev distance, distance of city blocks (the Manhattan distance). Among the metric indicators presented, the most popular one is Euclidean distance; it is a simple geometric distance, however, it integrates the objects that can be described via sphericities in the best way. Minkowski power distance is of interest only as a mathematical universal
metric, Chebyshev distance is the most suitable metric when identifying differences between objects, the Manhattan distance is very similar to Euclidean distance, but it gives less measuring errors and is more resistant to “runs” as the coordinates are not squared.

There are two basic methods in cluster analysis: hierarchical and non-hierarchical [12, 16]. The first group of methods includes agglomerative and divisive methods. Among the non-hierarchical methods there is the method of k-means. The essence of hierarchical methods involves the sequential (step-by-step) merging of objects into clusters. At the first step (iteration) each object is unique; with each next step it loses its uniqueness and becomes more and more generalized, at the same time the number of classes decreases. The advantage of hierarchical methods is their visibility expressed in the possible construction of dendrograms (Fig. 1) which clearly present the hierarchical structure of the array of analyzed data. Besides, the hierarchical methods can involve different procedures of combining objects into clusters: the “Nearest neighbour” distance (single linkage); the “Far neighbour” distance (complete linkage); unweighted pair average; weighted pair average; unweighted centroid method; weighted centroid method (median); Ward’s method.

In the first procedure the distance between classes is estimated by the distances between the nearest objects of classes; this procedure is more suitable for cases when it is necessary to combine the objects with the least degree of iterations. The second procedure has the opposite goal and at each iteration the number of combined clusters will be minimal. The unweighted pair average procedure uses the average distance between all class objects and the center of the class. The procedure is efficient when comparing heterogeneous data. The weighted pair average procedure is identical to the previous one, except for taking into account the number of objects located in every compared cluster, therefore, it must be used if clusters sizes are different. In the unweighted centroid procedure, the distance between clusters is understood as the distance between their centroids. The weighted centroid procedure is more preferable if there are significant differences in cluster sizes. According to Ward’s procedure, the sum of squares of the distance between each object and the average in the cluster containing this object is used. At each new iteration, only those clusters are combined, the combination of which leads to a minimum increase in the intragroup sum of squares. This procedure is more suitable for combining the closely located clusters.

![Figure 1. An Example of Dendrogram for Cultural and Social Objects.](image)

The k-means method is a non-hierarchical method. The essence of this method lies in the fact that a researcher believes that the array of objects (events) in question can be divided into two or more groups (clusters). Based on this assumption, the objects are combined into clusters at each iteration according to the principle of the largest distance between each other or from the center of the cluster. In other words, the research object is referred to the cluster which corresponds to the maximum
distance from the centers of other clusters. At the next iteration the cluster center is recalculated taking into account a new object. Thus, the resulting clusters will contain objects that do not have almost any resemblance to each other.

The method of k-means includes several procedures that adjust the clustering results. In Statistica application, one of the procedures is called: “Choose observations to maximize initial between-cluster distances” – according to this procedure, the first several values are chosen as cluster centers; their number depends on the sample size. This procedure results in maximized distances between clusters. The procedure “Sort distances and take observations at constant intervals” involves determining permanent centers of clusters only after the preliminary analysis of the whole data set.

**Practical Relevance**

The classification of territories (FPs) given their diversity is a very complicated task. However, researchers can use tools and methods of statistical analysis to determine relationships, similarities or lack of them in studied human-made or natural processes and phenomena. In particular, determining similarities between the types of territories (FPs) is possible on the basis of some or all of the available quantitative criteria (area, time of parking, visitors' load at definite peak times, etc.). Numerous experimental comparisons resulted in finding out the most suitable criteria for differentiating the territories (FPs) typology; they are their area and distance from the city center, which indirectly reflect the popularity of the territory (FP). The tree-based diagram of differentiation is presented in Figure 2.

As a measure of intraclass distance, the distance of city blocks (the Manhattan distance) is used.

![Figure 2. Tree Diagram of Differentiation of Territories (FPs) according to the Parking Time.](image)
The data of Table 1 are based on Fig. 2; the table reflects the distribution of the number of objects entering each cluster.

| Territory Code | Territory Name | Clusters |
|---------------|---------------|---------|
|               |               | 1  | 2  | 3  | 4  | 5  | 6  | 7  | 8  |

*Table 1. Distribution of Territories (FPs) by Clusters.*
Table 1: Distribution of Territories (FPs) by Clusters

| Cluster | Territory Description                          | Cases |
|---------|-----------------------------------------------|-------|
| 1       | Housing                                       | 5     |
| 2       | Trading, Public Catering, Service Sector      | 9     |
| 3       | Education, Healthcare, Sport, Culture, Leisure| 6     |
| 4       | Offices                                       | 5     |
| 5       | Industry                                      | 0     |
| 6       | Warehousing and Utilities                     | 2     |

The data of Table 1 are presented in Figure 3.

![Figure 3. Distribution of Territories (FPs) by Clusters.](image)

**Findings**

Thus, the table and the figure clearly show that each cluster corresponds to one of the most significant types of territories. The sixth cluster is more common for the territory (FP) housing (code 1), while for trading, public catering and service sector the second cluster is more common, education, healthcare, sport and culture enter the seventh cluster. Offices are mostly represented by the fourth cluster, etc. It should be noted that there are several prior clusters in the category of warehousing and utilities: the seventh, the second, the third ones; besides, in other integrated territories (FPs) there is no definite distribution by the clusters. This is mainly conditioned by the extremely high uniqueness of each territory (FP), which begins to function with significant differences even if the single factor is changed, for example, remoteness from the city center. Thus, it is necessary to perform the in-depth analysis of the reasons of the territories duplication within the same cluster. The first cluster is represented almost in every integrated type of territories (FPs) which indicates the diverse characteristics of objects and inability to refer them to a specific group. The second cluster is represented in the enlarged territory (FP) trading, public catering, service sector (code 2) best of all; besides, it included housing located far from the city center which undoubtedly influenced the
redistribution of the second cluster. The third, fifth, and eighth clusters like the first one are represented almost in all integrated types of territories (FPs). The fourth cluster dominates in the enlarged territory (FP) offices (code 4), although, housing and garages located quite far from the city center were included in this cluster, which undoubtedly influenced the redistribution of this cluster. Housing (code 1) dominates in the sixth cluster, despite there is one logistics center there as well. The seventh cluster is convincingly represented in the enlarged territory (FP) education, healthcare, sport, culture, leisure (code 3), although the cluster also covered such objects as a service station, petrol station, communal warehousing facilities, etc., that have different modes of functioning.

Conclusion

Unfortunately, the parameters used to perform the cluster analysis were not unique for each of the 6 enlarged types of territories (FPs). Therefore, it should be recognized at the arbitrary level that there are such parameters that would allow separating facilities of the communal warehousing and industry (codes 5 and 6) into separate clusters just as it was done with the territories (FPs) (codes 1-4), however, it is impossible to define them at this stage.

References

[1] Alekseev Yu.V., Somov G.Yu. Organization of Urban Planning: textbook. – M.: MGSU, 1996. – 59 p.
[2] Belinsky A.Yu. Laws of Mobility and Resettlement of Residents of Tallinn: Collected Papers Urban Transport. – M.: CSTI for Civil engineering and Architecture. – 1969.
[3] Biryukova L.G., Bobrik G.I. et al. Probability Theory and Mathematical Statistics: textbook / L.G. Biryukova [and others]. – M.: INFRA-M, 2004. – 287 p.
[4] Borovikov V.V. Statistica: The Art of Computer Analysis of Data. For Professionals. – SPb.: Peter, 2001. – 656 p.
[5] Vilner M.Ya. On the Town-Planning Bases of the Russian Innovative Development. Real Estate and Investments. Legal Regulation 4 (37) December 2008.
[6] Vukan R. Vuchik. Transportation for Livable Cities / transl. from English by A. Kalinin, scientific ed. M. Blinkin. Territory of the Future; Moscow; 2011
[7] Urban Planning. Ed. by V.N. Belousova. 2nd edition. – M.: Stroiiizdat, Designer’s Handbook, 1978. – 367 p.
[8] Zaks L. Statistical Evaluation / transl. from German by V.N. Varygin; ed. by Yu.P. Aller, V.G. Gorsky. – M.: Statistika, 1976. – 598 p.
[9] Kudryavtsev A.O. Rational Use of Territory when Planning and Building Cities in the USSR. M., Stroiizdat, 1971.
[10] Lobanov E.M. Transport Planning of Cities. – M.: Transport, 1990. – 240 p.
[11] Mikhalev Yu.A. Basics of Urban Development and Planning. Textbook / Krasnoyarsk State Agrarian University. – Krasnoyarsk, 2012. – 237 p.
[12] Bureeva N.N. Multivariate Statistical Analysis Via STATISTICA Computer Program. Educational materials on the advanced training program Use of Software in Scientific Research and Teaching Mathematics and Mechanics. Nizhny Novgorod, 2007, 112 p.
[13] Romm A.P. Comprehensive Assessment and Functional Zoning of Territory in Urban Design: thesis for a Doctor of Architecture: 18.01.04. – Moscow, 2002. – 206 p.
[14] Guide to Urban Reconstruction. Administrative Centers of Cities. – M.: Stroiizdat, 1979. – 159 p.
[15] Safronov E.A. Transport Systems of Cities and Regions: textbook / E.A. Safronov. – M.: Publishing house ASV, 2005. – 272 p.

[16] Electronic textbook Statistica [Electronic resource] // http://statsoft.ru/ (access date: 10.10.2015).

[17] Trip Generation Handbook, 2nd Edition: An ITE Recommended Practice. Washington, DC: ITE, 2004.

[18] Zedgenizov A.V. The Estimation of Correspondence Generation to the Residential District of City Centre. Nauka i studia, 2014. – № 16 (126). – С. 76-83.

[19] Zedgenizov A.V., Burkov D.G. Methods for the Traffic Demand Assessment Based on the Quantitative Characteristics of Urban Areas Functioning. 12th International Conference Organization and Traffic Safety Management in Large Cities, SPbOTSIC-2016, 28-30 September 2016, St. Petersburg, Russia. Edited by Ulrich Brannolte, Pavel Pribyl and Valentin Silyanov. Transportation research pro-cedia, P. 724-730.

[20] Zedgenizov A.V. Location-Based Transport Demand Forecasting Methods for Suburbanized Areas. Proceedings of the international conference Aviamechanical Engineering and Transport (avent 2018) AER-Advances in Engineering Research volume 158, 2018, P. 458-461.