# ROAD NETWORK BASED COMMUNITY DETECTION. CASE STUDY FOR AN EASTERN REGION OF AUSTRO-HUNGARIAN MONARCHY

*Zsolt **MAGYARI-SASKA***[1]

DOI: 10.21163/GT_2019. 141.06

**ABSTRACT**. The aim of the study was to find ways of measuring the relationship between clusters resulted on road networks topology and regional communities, before modern transport vehicles had widely spread. We have chosen as investigation area an eastern region of Austro-Hungarian Monarchy because at the beginning at the XX. century it had a modern transportation infrastructure for its time and a multicultural environment where different local or regional communities were present. Using data of that period we constructed a network model, attaching the populations' mother tongue as attribute data for settlements. After applying community detection algorithms considering only road network topology and later also the attribute data, we tried to compare the resulted clusters. We observed that in some cases mother tongue-based communities form sub-clusters for solely road network base clusters, while in other cases mother tongue-based communities restructure the road network based clusters. To quantify the differences, we identified three similarity/dissimilarity comparing methods, one based on clusters' spatial extends and two on statistical approaches.

***Key-words***: *Heterogeneous network, Modeling reality with graphs, Clustering, R, Spatial relation, Comparing communities*

## 1. INTRODUCTION

In society from the very beginning, communities have appeared, that goes beyond the family, friends or acquaintances and even beyond local and settlement-level communities. These communities hold a local or regional identity that distinguishes them from the neighboring environment (Pohl, 2001).

The study of the positive and negative effects of these communities has inspired many studies (Semian & Chrmy, 2014; Chrmoy & Janu, 2003), which all confirm the social importance of them. Most of these researches focus on the present, but at the same time also there also studies on past research and historical exploration (Baker & Biger, 1992).

There are many types of communities which are based on different factors (McMillan & Chavis, 1986). This study attempts how can be explored the relationship between human communities and transport routes, and how these two concepts are related. As a test site an eastern region century of the Austro-Hungarian Monarchy (**Fig. 1**) at the beginning of the 19$^{th}$ century was selected, which had a relatively developed infrastructure at that time, and based on its geography and history had a multiethnic and multicultural society (Brie, 2014), which offered an opportunity for the appearance of several different regional communities.

Similar type of researches that explores spatial homogeneity and diversity and which studies populations clustering, is not new (McDoom & Gisselquist, 2015; York et al., 2011), but it is a constantly evolving trend where new methodological approaches have their place (Páez et al., 2012). The importance of networks is also highlighted in different studies as an important factor in analyzing social phenomena (Bobkova & Holesinska, 2017; Cadar et al., 2017).

[1] *Babes-Bolyai University, Faculty of Geography, 535500, Gheorgheni, Romania, zsmagyari@gmail.com*

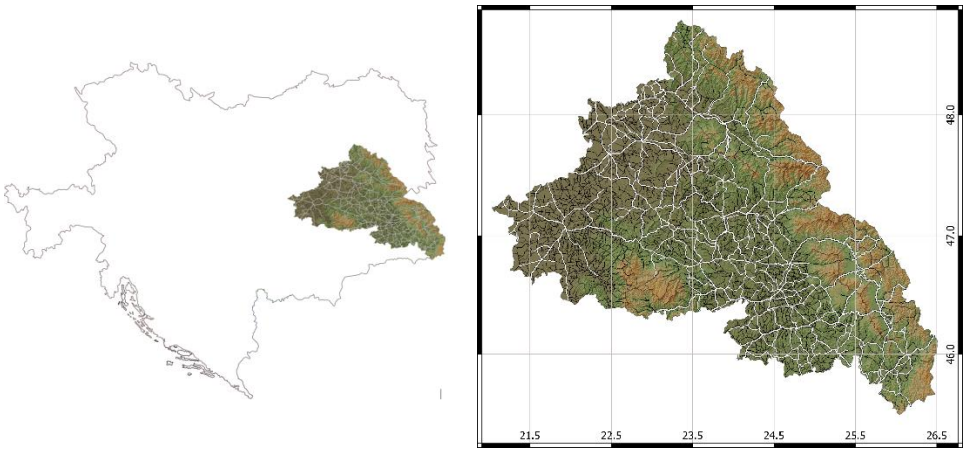



**Fig. 1.** The study area within the Austro-Hungarian Monarchy (left) and its road network (right)

This research is made up of several interconnected processes. The basic tasks that model these processes are in fact answers to the following questions: how a weighted network model based on historical maps can be created; how the homogeneity of two connected settlements based on mother tongue can be expressed and measured; how the similarity of two different cluster map can be evaluated. The backbone of this research is answering these questions, projecting the results into the target area.

## 2. BUILDING THE DATA MODEL

### 2.1. Data sources and their limitations

Our first task was to create a network topology where nodes represent the settlements and the edges the existing directly connecting roads between them. A directly connecting road should not pass through other settlements but can go through multiple crossroads.

The input data we had was the point layer with settlements' position, and the line layer containing the road network, based on ten pieces of late XIX, early XX century Austro-Hungarian Empire map sheets. Their scale varied between 250.000 to 400.000. Another, but non-spatial data source was a tabular record of the 1910 census including the mother tongue of the population.

To build the desired network we had to properly snap the settlements on the road network. This step theoretically could have been done in the time of vectorization, we intentionally don't want to do this for two reasons. Firstly, not all settlements on the map are placed near roads, in some cases, there are several kilometers between the settlement and the nearest road. Secondly, the basic idea of the vectorization was to preserve as much as possible the original map layout and content and a forced snapping would have altered this principle. That's why we had to find post-processing techniques to achieve our goal.

Latter automatic snapping was not a possibility as snapping operation considers just vertices which is not necessarily the best solution as it disregards the minimum distance between the settlement point and the road. In addition, the automatic snapping would have neglected the distance between the settlement and the road (sometimes considerable).

Even in the first steps, it was easy to identify that in first phases we can build a heterogeneous network, which means that not all nodes have the same role. Some of them

are crossroads with routing possibility while others representing the settlements which are the true data holder components of the network.

To build a usable network model, our processing algorithm used QGIS to create the heterogeneous network, Gephi for basic visualization and to eliminate non-connected components and R for homogenous network building, in which all nodes will represent settlements, maintaining the effective distance between them.

### 2.2. Building the heterogeneous network model

After the points representing the settlements were vectorized, the census data had to be geocoded based on the settlements name as no other uniquely identifying element exists. The used census data was available in electronic format for the study area from Varga E. Árpád database (http://varga.adatbank.transindex.ro).

One of the first problems we had to overcome was that the road network was composed by arbitrary road sections, the junctions were in the middle of road segments disregarding any relationship between road network and settlements position. We had to transform in such a way that every road section to end at a crossroad. For this, the *Split with lines* command was used.

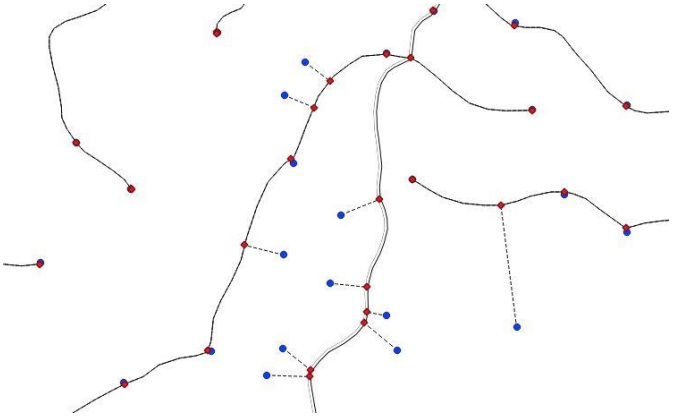


**Fig. 2.** Road network model in QGIS (red marks – junctions/crossroads, blue marks - settlements, dash lines – added segments to road network)

As mentioned the settlements position initially was not snapped to roads or crossroads. That means that every settlement had to be connected with a separate line to the road network. In QGIS the *Connect nodes to lines* from *Network* packages could do this. For every settlement a new road section is added to the nearest road, splitting the initial road segment into two and creating a junction (**Fig. 2**). After this operation, the *Network* package in QGIS is capable to build the node and edge tables representing the graph. The edges endpoints are the ending vertices of road sections. In such way, we had a network in which some of the vertices marks junctions or crossroads, while others vertices mark settlements, both without any data. The last step was to add the geolocated census data to the proper nodes with *Join attributes by location* command. The length of the road segments was also calculated and added to edges.

The resulted two layers (nodes and edges) were imported in Gephi (**Fig. 3**) where a component analysis was performed, eliminating 4 small parts of the graph which were not connected (as not the whole monarchy's map was used) to the main core. After this operation, we had 10266 edges and 9157 nodes, 3211 of them representing settlements.

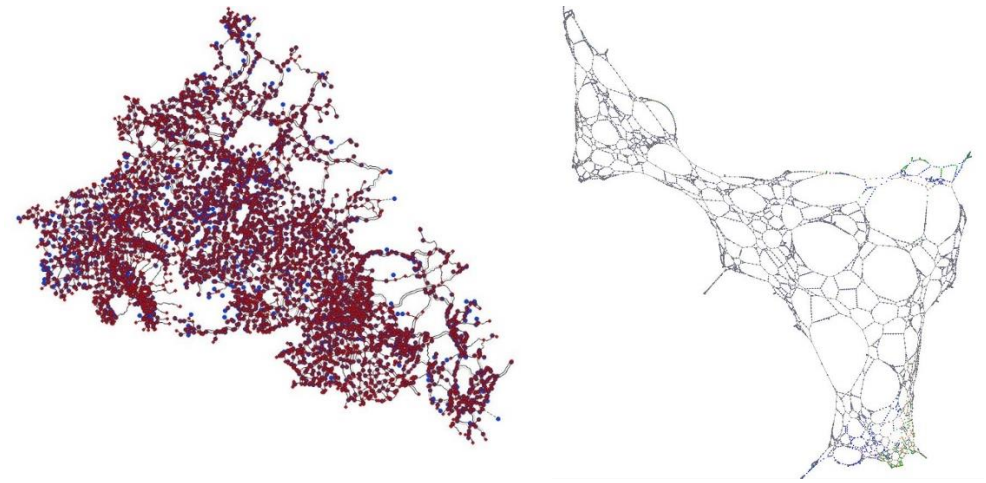


**Fig. 3.** Settlements connectivity GIS data model (left) / network data model (right)

### 2.3. Building the homogeneous network model

The homogeneous network model, which should connect only settlements, if between them exists a direct road was built based on the following algorithm implemented in R.

***Deleting of some singular nodes***. As not all nodes represent settlements, those which has only one connection and are not settlements could not participate in any routing sequence, they should be eliminated.

***Inserting the settlement nodes on the road network***. In fact, this operation is composed of three steps: 1) adding the edge's length which connects the settlement node and a junction node to both other edge of the junction node, 2) copying the settlement data to the junction node and 3) deleting the original settlement node.

***Putting some nodes in crossroads***. Now all settlement nodes are on roads but in many cases, the settlements are not positioned in crossroads but are very close to them. This situation causes incorrect routing data between settlements as it seems that one of them is connected directly to much more than in reality is. In this step, we overcame this situation by "rearranging" the characteristics of the network if a settlement node is closer to a junction node than a predefined distance. In such case, the settlement node is "moved" to the junction node adjusting the length of connecting edges, in such way that all implied edges to reflect the real length (**Fig. 4**).

The result so far still contains both types of nodes, therefore a new network should be built containing just the settlement nodes and the connecting weighted direct edges. In this step, the Dijkstra algorithm, implemented in R was used to find the shortest paths between settlement nodes.

In the end, we obtained a connected graph with 3211 nodes and 5279 edges were all nodes represents settlements having as attributes regarding the number of inhabitants with different mother tongues. An edge between two settlements represents the possibility to travel from one to other without crossing other settlements. Edges are weighted by travel distance.

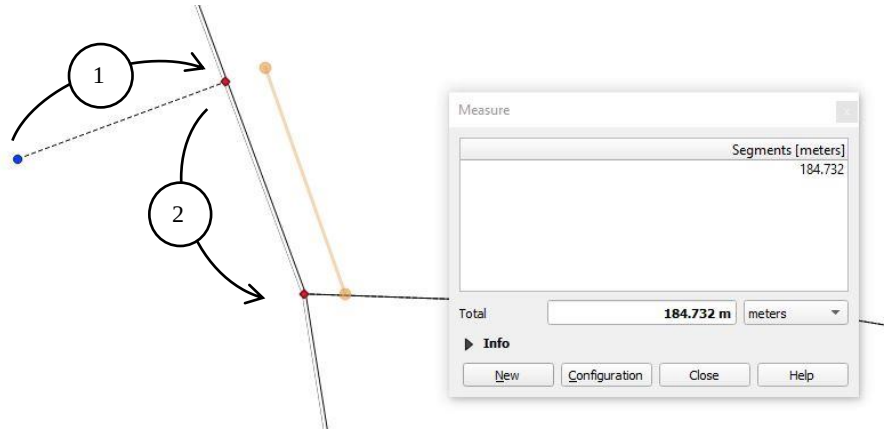



**Fig. 4.** Settlement repositioning to a nearby crossroad in the network model
(1 – inserting settlement on the road network; 2 – moving settlement to crossroad)

## 3. CLUSTER ANALYSIS – METHODS AND RESULTS

Cluster analysis refers to object grouping based on certain criteria. In connectivity models, represented by graph or networks, when the linkage between objects exists the clustering process is equivalent to community detection. This process is based on the similarity between nodes and can take account the topological characteristics of the network or beyond it also other attributes of the network components. In the first case, the most known similarity values are the common neighbors, cosine, Jaccard or min indexes (Fu et al., 2017) although particular indexes could be defined (Cheng et al., 2013).

In our case, we were interested to include in community detection values, related both to vertices having the mother tongue distribution attribute as for edges characterized by their spatial distance. There are several studies which includes such approaches, defining similarity, based on edge attributes (Steinhauser & Chawla, 2008), based on vertex attributes (Dang & Viennet, 2012; Boobalana et al., 2016) or even based on both types of attributes (Zhou et al., 2009; Chakrabarty et al., 2016).

The similarity value between two nodes (settlements) in this research was based on the Herfindahl concentration formulae which is often used in ethnic fractionalization analysis (Posner, D.N., 2004; Bossert et ll., 2011; Fearon, D.J., 2013). In our case the original formula, which was developed to be a measure of fractionalization for each given location, was transformed in such a way to consider the sum of share differences of each mother tongue group for an edge connected settlement pair, as shown below:

$$edge_{AB} = 1 - \sum_{i=1}^{k} (A_i - B_i)^2 \quad (1)$$

where,

$edge_{AB}$ – edge attribute, based on mother tongue similarity
k – the number of different mother tongue categories
$A_i$, $B_i$ – share of *i.* mother tongue in each settlement

It's easy to observe that in case of exact matches between mother tongue shares the edge attribute value will be 1, and as much as there are higher differences between shares

the value decreases to 0. After the similarity was defined, the cluster analysis had to come. There are many methods for community detection which differ in direction, complexity, resulted cluster shape (Neethu & Surendran, 2013; Bindiya et al. 2014). From the implemented clustering algorithms in igraph package of R we used the FastGreedy method. This algorithm uses a bottom-up hierarchical approach, merging smaller communities to maximize the modularity score, which is a measure of internal strength of a network component. When the modularity score could not become higher the algorithm stops.

In our research, we started the cluster analysis on the created road network using different edge weight. In the first case we don't consider the mother tongue attribute, leaving the community detection algorithm to take account just the length of connecting roads. In this case, we obtained 46 structural communities as shown in **Fig. 5a**.

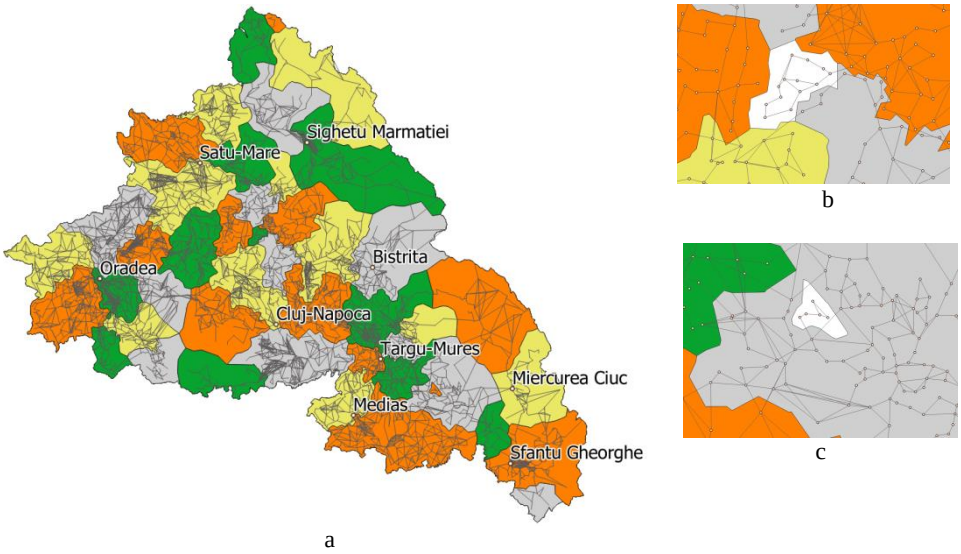



**Fig. 5.** Clusters resulted based on topological relation (a) with examples of isolated clusters (b,c) (lines represent the network model, colors has no special meaning just facilitate interpretation)

The cluster analysis was set up on a settlement network, each settlement receiving an id value corresponding to the cluster number to belongs to. The spatial extension of the clusters, represented as polygons were create with as Voronoi polygons, later merged by their cluster id attribute field. Due to this method, the area of clusters as it's shown on the map does not reflect the size of the cluster. Nonetheless, the spatial extent visualization helps to identify the individual cluster covered area and patterns of isolated clusters as presented in **Fig. 5b, 5c**. In the next phase of the analysis, we considered as edge weight only the similarity based on mother tongue as expressed in formula 1. In this case, we obtained 58 communities, with 26% more than the cluster number obtained at the previous step (**Fig. 6**).

The growth of cluster numbers theoretically suggests the division of initial clusters but that's not true for the whole study area. In the southeast region, we can observe sub-clusters of the road network based clusters, while in the west part we can observe clusters which aggregate partially or totally some road network based clusters.

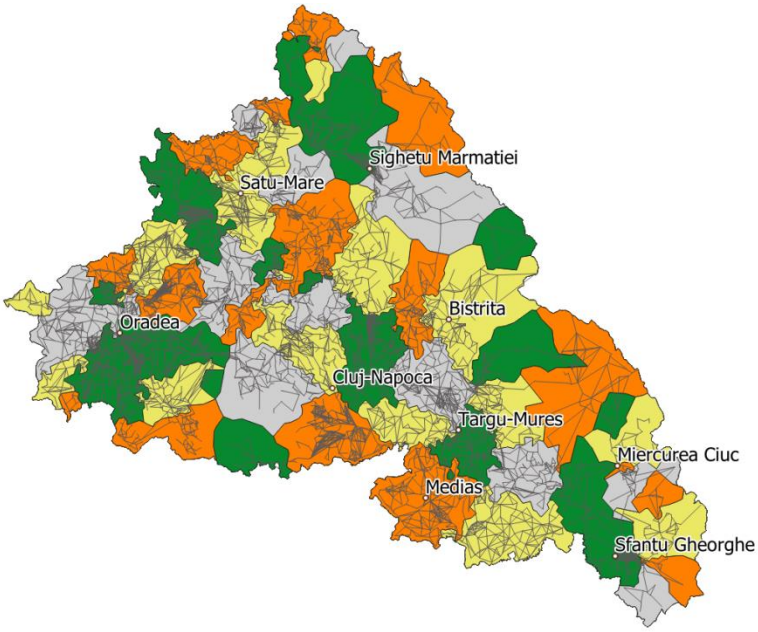



**Fig. 6.** Clusters resulted based on mother tongue similarity
(lines represent the network model, colors has no special meaning just facilitate interpretation)

Furthermore, we wanted to analyze if this cluster number and configuration changes dynamically if we try to limit the search radius at community detection, using a minor change in formula 1. We have repeated the cluster analysis based on mother tongue similarity gradually increasing the search distance from 2km to 100km. The evolution of cluster numbers stabilizes at 48km. The number and spatial extent of these stabilized clusters are exactly the same as those obtained when no distance limit was imposed. Beyond the 48km value, no changes appear in case of mother tongue-based community detection. We were also interested in the first distance value from which the final clustering result starts to appear. We considered a 10% limit against the 58 value representing the cluster numbers. In this case, starting from 16km the resulted clusters tend to be identical with the final ones (**Fig. 7**).

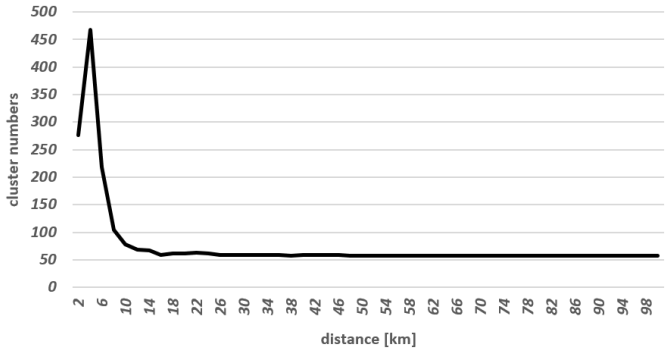



**Fig. 7.** The number of clusters evolution by considered search distance
(cluster analysis based on road network topology and mother tongue similarity)

To measure the similarity of the resulted clusters we used three approaches: based on spatial extension, based on cluster size (settlement number) and based on mother tongue similarity. For the first one, based on the spatial extension we developed a comparison algorithm between two polygon layers, measuring the fragmentation with the following steps:

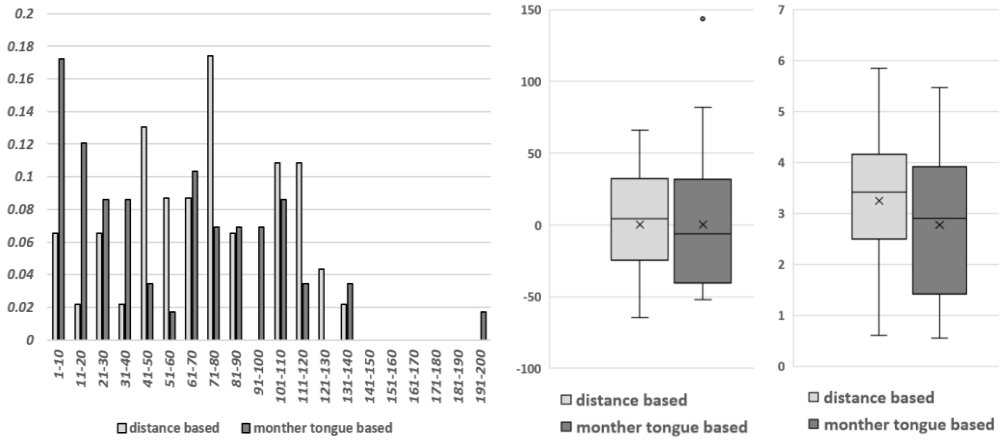
- one of the layers is considered the base layer, to which we will refer to. The other layer is considered the fragmenting layer.
- the fragmenting polygon layer is transformed to cutlines and it will divide the base layer in multiple polygons
- for each cluster value in the base layer, the ratio between the sum of the squared area with the same cluster value and the base layers' squared area is calculated as shown in formula 2, obtaining a value between 0 and 1. As higher the calculated value is there's a higher similarity.

$$cluster\ similarity_c = \frac{\sum_{i=1}^{n_c} {A_i}^2}{A_c^2} \tag{2}$$

where,

c – cluster number for which the similarity is calculated
$n_c$ – the number of polygons in the fragmented layer having c as cluster identifier
$A_i$ – area on the fragmented layer
$A_c$ – area on the base layer

By applying the mentioned algorithm, we have got 46 values, one for each road based cluster. The mean cluster similarity value of the road based clusters by the mother tongue base clusters was 0.6 with a standard deviation of 0.24. There were 6 clusters, 13% of the total, which had the same boundaries in both clusterings.




*a* *b* *c*

**Fig. 8.** Frequency distribution of cluster size (a), the statistical summary of the two clusters data series based on cluster size (b) and mother tongue relative standard deviation (c)

Secondly, we compared the resulted clusters' size. As the number of clusters differs in the two analyzed situations we do not compare the two data series directly, but their standardized values. The standardized values were calculated as differences between each value of a data series and its theoretical average size (settlement number divided by the

number of clusters). In this case even if in the percentile frequency distribution of cluster sizes there are differences, the two series statistical characteristics are similar (**Fig. 8b**). It's notable the presence of a huge cluster with 199 settlements in case of mother tongue-based community detection (**Fig 8a**), which has appeared by aggregating several road based cluster parts near Oradea. The last comparison method was to analyze the homogeneity of the resulted clusters after both clustering methods. For this we used the relative standard deviation (standard deviation divided by the mean) for all mother tongue classes, calculating its mean for each cluster. The statistical characteristics of the resulted two data series are similar with a notable lower mean value, representing a more homogeneous cluster, for mother tongue-based clustering (**Fig. 8c**), as it was expected.

## 4. CONCLUSIONS

Our research followed two main directions: developing a study methodology for comparing two types of clustering results and applying it to a study region. Firstly, we presented an automated way to create a homogenous network, based on routes and settlements even if the two type of objects were vectorized independently with no initial spatial connection between them. The only parameter of the algorithm is the distance threshold for snapping a settlement to a crossroad. With all transformation, the algorithm maintains the real distance between the settlements. Another result of the research was a cluster analysis based on edge attached similarity values, which at data level was based on the share of a whole. For comparing the results of two clustering operations we presented three different approaches, each of them having a specific role in data evaluation.

Regarding the selected study area, the obtained cluster configuration becomes constant starting from the 48km distance limit in cluster analysis. This value can be considered close to the maximum distance that can be covered at that time without motorization. Comparing the clusters obtained with and without considering the mother tongue-based similarity between settlements the spatial similarity value indicates that in approx. 60% of the study area belongs to the same community in both clustering method. Both statistical comparison of the resulted clusters also indicates a notable similarity, especially regarding cluster size.

Considering the results of the three comparison types, we can affirm that road network topology was seriously related but not determinative in forming local or regional communities based on mother tongue till the I World War at the study area. Our study reinforces that human communities are formed considering the transportation network. However, the presented methodology had to be tested also in other location comparing the obtained results.

## ACKNOWLEDGEMENT

The research was supported by the DOMUS scholarship program of the Hungarian Academy of Sciences.

## REFERENCES

Baker, A.R.H. & Biger, G. (1992) *Ideology and Landscape in Historical Perspective: Essays on the Meaning of Some Places in the Past*. Cambridge University Press, Cambridge.

Bossert W., D'Ambrosio C. & La Ferrara E. (2011), A generalized index of fractionalization, *Economica*, 78, 723-750

Bobkova, M. & Holesinska, A. (2017) Networking in a destination from the perspective of virtual relationships and their spatial dimension, *Geographia Technica*, 12(2), 10-19

Brie M. (2014), Ethnicity and politics in the Romanian space. The case of northwestern Transylvania. Published in: No. Sorin Șipoș, Gabriel Moisa, Dan Octavian Cepraga, Mircea Brie, Teodor Mateoc (coord.), *From Periphery to Centre. The Image of Europe at the Eastern Border of Europe*, Editura Academia Română. Centrul de Studii Transilvane, Cluj-Napoca, 158-170.

Cadar, R.D., Boitor, M.R. & Dumitrescu, M. (2017) Effects of the traffic volumes on accidents: the case of Romania's national roads, *Geographia Technica*, 12(2), 20-29

Chakrabarty, A., Chelladurai, J. & Venkateswaran, S.K. (2016) Community detection in citation networks using attribute similarities, *International Journal of advances in cloud computing and computer sciences*, 2(5), 1-5

Cheng, J., Xu, H., Gaybullaev, M., Leng, M. & Chen, X. (2013) Community Detection Algorithm based on Neighbor Similarity, *Telkomnika*, 11(8), 4484-4490

Chromý, P. & Janů, H. (2003) Regional identity, activation of territorial communities and the potential of the development of peripheral regions. *Acta Universitatis Carolinae – Geographica* 38, 105-117.

Dang, T.A. & Viennet, E. (2012), Community Detection based on Structural and Attribute Similarities, *The Sixth International Conference on Digital Society*, 7-12

Fearin D.J. (2013) Ethnic and cultural diversity by country, *Journal of Economic Growth*, 8(2), 195-222

Fu, Y.H., Huang, C.Y. & Sun, C.T. (2017) A community detection algorithm using network topologies and rule-based hierarchical arc-merging strategies, *PLoS ONE* 12(11), 1-30

McDoom O.S. & Gisselquist R.M. (2015) The Measurement of Ethnic and Religious Divisions: Spatial, Temporal, and Categorical Dimensions with Evidence from Mindanao, the Philippines., *Social Indicators Research*, 129(2), 863-891

McMillan, D.W. & Chavis, D.M. (1986) Sense of Community, *Journal of Community Psychology*, 14, 6-23

Neethu C.V. & Surendran, S. (2013) Review of Spatial Clustering Methods, *International Journal of Information Technology Infrastructure*, 2(3), 15-24

Páez, A., Ruiz, M., López, F. & Logan, J. (2012) Measuring Ethnic Clustering and Exposure with the Q statistic: An Exploratory Analysis of Irish, Germans, and Yankees in 1880 Newark, *Annals of the Association of American Geographers. Association of American Geographers*, 102(1), 84-102.

Pohl, J. (2001) Regional Identity, *International Encyclopedia of the Social & Behavioral Sciences*, 12917-12922

Posner D.N., (2014) Measuring ethnic fractionalization in Africa, *American Journal of Political Science*, 48(4), 849-863

Semian, M. & Chromý, P. (2014) Regional identity as a driver or a barrier in the process of regional development: A comparison of selected European experience, *Norsk Geografisk Tidsskrift - Norwegian Journal of Geography*, 68(5), 263-270

Steinhaeuser, K. & Chawla, N.V. (2008) Community Detection in a Large Real-World Social Network, In: Liu H., Salerno J.J., Young M.J. (eds) *Social Computing, Behavioral Modeling and Prediction*, Springer, 168-175

Varga E. Á. (2010) Ethnic and denominational statistics of Transylvania (1850-1992), *in hungarian* (Erdély etnikai és felekezeti statisztikája (1850-1992)). Erdélyi Magyar Adatbank. http://varga.adatbank.transindex.ro

Varghese, B., Unnikrishnan, A. & Jacob, P.K. (2013) Spatial Clustering Algorithms- An Overview, *Asian Journal of Computer Science and Information Technology*, 3(1), 1-8

York, A. M., Smith, M. E., Stanley, B. W., Stark, B. L., Novic, J., Harlan, S. L., Cowgill G.L. & Boone, C. G. (2011) Ethnic and Class Clustering through the Ages: A Transdisciplinary Approach to Urban Neighbourhood Social Patterns. *Urban Studies*, *48*(11), 2399–2415

Zhou, Y., Cheng, H., Yu, J.X. (2009) Graph Clustering Based on Structural/Attribute Similarities, *Proceedings of the VLDB Endowment*, 2(1), 718-729