

COMPARATIVE ANALYSIS OF DIFFERENT METHODS AND OBTAINED RESULTS FOR DELINEATION OF FUNCTIONAL URBAN AREAS

Vojkan Gajović¹, GDi GISDATA LLC Belgrade, Serbia

European Spatial Planning Observation Network (ESPON) recognizes Potential Urban Strategic Horizons (PUSH) and Potential Polycentric Integration Areas (PIA) as territory of one or more neighboring Functional Urban Areas (FUA).

Delineation of FUA territory can be done by using general ESPON methodology, based on a 45-minute car travel time from the center of respective FUAs. This approach is based on network proximity by using shortest path in road network between two nodes. Later, results are approximated on administrative or statistical territorial units, so that PUSH areas are determined. However, other methods for delineation of FUA territory can be used.

This paper deals with other methods that can be used for delineation of FUA territory. Some of those methods are based on machine learning, a branch of artificial intelligence which develops algorithms that take as input empirical data, such as that from sensors or databases. Created algorithms identify complex relationships thought to be features of the underlying mechanism that generated the data, and engage these identified patterns to make predictions based on new data. Clustering and artificial neural networks are some of approaches that can be undoubtedly used for delineation of FUAs territory, based on unsupervised learning and statistical data analysis. This is statistical approach, which clusters administrative or statistical territorial units based on statistical data, and not by network proximity. Such methods involve usage of Self Organizing Maps (SOM) which implies usage of neighborhood function to preserve the topological properties, or using k-means clustering, which partition observations into clusters by dividing space into Voronoi cells. Results obtained from both approaches will be analyzed in order to define the most appropriate method for FUAs territory delineation in Serbia.

Key words: Functional Urban Areas (FUA), Potential Urban Strategic Horizons (PUSH), Self Organizing Maps (SOM), k-means clustering, Serbia.

INTRODUCTION

Functional urban areas (FUA) represent basic building blocks of polycentrism – a principle of organization of a region around several political, social or financial centres (Wikipedia, 2013). The concept of polycentrism has been introduced as a basis for spatial development through European Spatial Development Perspective (ESDP), where it has been highlighted as key goal of spatial development which would contribute to more balanced regional development, increase of competitiveness, more complete regional integration and sustainable development (ESPON, 2005). Polycentrism has two aspects determining

number and size of FUA: morphology, which deals with distribution of urban cores – cities and their hierarchy, and networking, which deals with relations between cities.

Based on these two aspects, FUA can be defined as an area consisting of urban core(s) and surrounding area that is economically integrated with core. European Spatial Planning Observation Network (ESPON) defines FUA as urban core with population size of at least 15,000 inhabitants and more than 50,000 inhabitants in total population size of and country with more than 10 million inhabitants, while for countries with smaller population size, FUA has to have urban core of at least 15,000 inhabitants and area with more than 0.5% of population size of country, and functions of national or regional importance as well. To

analyze territorial context of cities, ESPON uses two additional aspects (ESPON, 2005):

- Potential Urban Strategic Horizon (PUSH) – it encompasses all territorial units (such as municipalities or settlements) which have at least 10% of their territory available with 45 minute travel by car from center of FUA, so number of PUSH actually is number of FUA, while neighboring PUSH can overlap;
- Potential Polycentric Integration Areas (PIA) – they represent join of neighboring PUSH areas, if cities smaller in population share at least 1/3 of

¹ Bulevar Mihaila Pupina 10a, 11070 Beograd, Serbia
vojkan.gajovic@gdi.net

This paper was presented at the 2nd International Scientific Conference „REGIONAL DEVELOPMENT, SPATIAL PLANNING AND STRATEGIC GOVERNANCE – RESPAG”, Belgrade, 22nd-25th May 2013, organised by the Institute of Architecture and Urban & Spatial Planning of Serbia.

their PUSH area with neighboring larger city.

Each PUSH area participates in one PIA area, and if there are multiple overlaps, the largest city is chosen as center of belonging. PUSH area actually represents functional border of city.

Determination of PUSH area includes delineation of 45 minutes isochrone of car travel from center of FUA towards periphery via street and road network. This approach is based on network connection and proximity, usage of the shortest paths in road network between two nodes. One of most often used algorithms is algorithm of Edsger Dijkstra (Wikipedia, 2013).

However, different European countries have different definitions of FUA. Some examples include identification of FUA with area of daily migrations, commuter areas, etc. (Korcelli, 2008)

Beside ESPON standard method for delineation of FUA, this paper deals with other methods which can be used for this problem. Such methods are based on machine learning – branch of artificial intelligence which develops algorithms which use empirical data as input, such as sensor data or databases (Wikipedia, 2013). Created algorithm identifies complex relations which consider being part of mechanism which have created data, and then uses identified patterns for prediction based on new data. Clustering and artificial neural networks are representative approaches (Wikipedia, 2013), which can be used for delineation of FUA: both the urban core and PUSH area. These approaches are based on unsupervised learning and statistical data analysis. Clustering is performed on administrative or statistical territorial units based on statistical data, and not on network connections and proximity. Self-Organizing Maps (SOM) and *k*-means clustering are some of unsupervised learning methods which will be discussed in this paper, and whose results will be compared with standard ESPON method.

METHODS OF MACHINE LEARNING AS BASIS FOR DELINEATION OF FUNCTIONAL URBAN AREAS

Machine learning is defined as area of exploration which gives possibility to computers to learn without explicit programming (Samuel, 1959). More complete and more formal definition of machine learning was given by Tom M. Mithcell, in which he supposes that computer program is told to learn from experience *E* with respect to some classes of task *T* and measured performances *P*, if performances from task *T*, measured through *P*, are improving with experience *E*.

Such definition enables that famous Alan Turing's question "Can machines think?" (Turing, 1950) can be replaced with question "Can machines do what we (as thinking entities) can do?" (Harnad, 2008)

Machine learning is based on two concepts, from which one focuses on prediction based on known characteristics learned from samples, while other known as data mining is based on discovering previously unknown characteristics of data. These two concepts overlap in many areas, so it is hard to divide them in practical usage, although in research communities distinct division between these two concepts exists (Wikipedia, 2013). Algorithm types that are used in machine learning can be divided into categories based on wanted output or input learning data, so we have categories of supervised learning, unsupervised learning, semi-supervised learning, reinforcement learning and learning to learn.

Unsupervised learning implies finding hidden structures in unlabeled data. Some of approaches in unsupervised learning are: clustering, neural networks, principal component analysis, independent component analysis, etc.

In this paper, two methods of unsupervised learning will be used: *k*-means clustering and self-organizing maps.

K-means clustering

K-means clustering represents one of the simplest algorithms of unsupervised learning (MacQueen, 1967), whose goal is to divide *n* observations into *k* clusters, where each observation belongs to cluster with values closest to mean. Each cluster has its own center, which is called centroid. As a result, division of space between centroids into Voronoi cells is achieved. Algorithm is expressed through following objective function (1):

$$J = \sum_{j=1}^k \sum_{i=1}^n \|x_i^{(j)} - c_j\|^2 \quad (1)$$

where

$$\|x_i^{(j)} - c_j\|^2 \quad (2)$$

represents chosen distance between point $x_i^{(j)}$ and centroid of cluster c_j . Function is actual indicator of distance of *n* points from centroid of cluster they belong to (Matteucci, 2012). Standard algorithm performs in such manner so centroid for *k* number of points in space is determined, and each cluster receives attribute of belonging to closest centroid. When each point gets its own attribute of belonging, the recalculation of centroid position is performed for each cluster until new centroid is determined. This step is repeated until convergence is achieved and centroids become fixed (Figure 1).

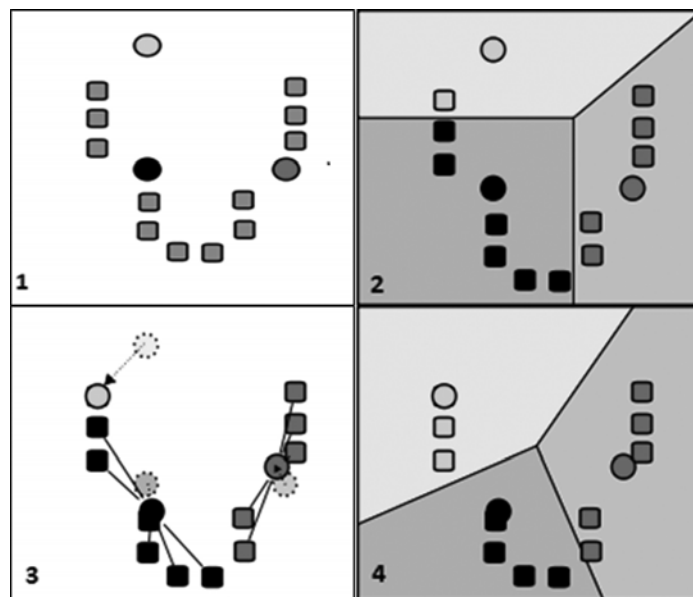


Figure 1. Functioning of *k*-means algorithm with goal to create 3 clusters:

- 1) initial determination of centroids based on spatial domain;
- 2) cluster creation with Voronoi cells while associating each point to closest centroid;
- 3) each clusters centroid becomes new centroid, so re-clustering performs based on distance;
- 4) steps 2 and 3 repeat until achieving convergence – centroid inaction.

K-means is most popular clustering algorithm. It has its own strengths and weaknesses. Strength is seen in simplicity in understanding and application, and in efficiency as well. However, basic components, initial centroids, distances and cluster numbers, are its greatest weaknesses. Initial centroids have massive influence on results, so large number of pre-application analysis must be performed to determine best initial centroids. Users must choose predefined number of clusters, which limits research on chosen number. Distance is represented by Euclidean distance between two points, while variance is used as a measure of dispersal. There is also problem of sensitivity of algorithm on points which are isolated from other points, which can influence spatial disposition and size of clusters. Algorithm is not recommended for clustering of points which do not have hyper-ellipsoidal or hyper-spherical dispersion (Liu, 2007). Modern tools which execute *k*-means algorithm can reduce the influence of these weaknesses. One is Grouping Analysis tool within Esri ArcGIS for Desktop software package, which has ability to use user's defined initial centroids, or software can find optimal centroids, or centroids are defined by random sampling (Esri ArcGIS Resources, 2013). Algorithm can be executed multiple times with different combination of inputs, so different results can be expected. Other tool is CrimeStat (Levin, 2008), primarily used as statistical program for spatial analysis of crime, and which can perform *k*-means clustering with user's defined initial centroids.

Algorithm has been executed within CrimeStat version 3.0, on point representation of settlements in Serbia, in two iterations (Figure 2): in first one, the software alone has determined most optimal initial centroids, while in second one initial centroids are represented by 25 centers for settlements which are significant urban cores of administrative importance. For weight factor, level of urbanity calculated on model of Tošić (2012) has been used, while as intensity number of inhabitants according to 2002 census has been used.

Self-Organizing Maps (SOM)

Self-Organizing Maps (SOM), known also as Kohonen network, represents a method for visualization and analysis of multi-dimensional data, especially those which are experimentally collected (Kohonen, 1982; Kohonen, 2001). In addition, they can be used for clustering (Vesanto and Alhoniemi, 2000), reduction of dimensionality, classification, sampling, vector quantization and data mining (Kohonen, 2001).

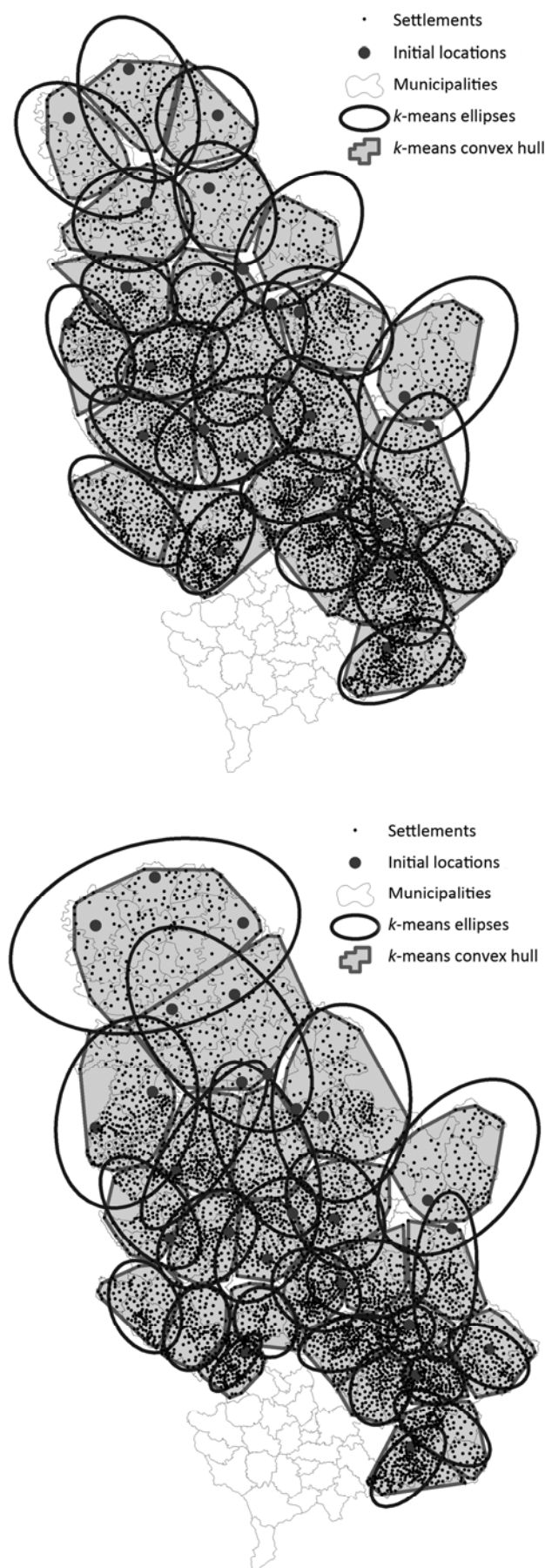


Figure 2: Results of *k*-means clustering with CrimeStat 3.0 – upper with initial centroids defined by user, lower with initial centroids defined by software

SOM is defined as projection of data on regular, usually two-dimensional grid of nodes, while data are mapped on node whose model has been calculated through SOM algorithm and whose model is most similar to data (Figure 3), for example it is closest to data in some measures (Scholarpedia, 2013).

Model usually represents mean arithmetic value of data in space, while SOM algorithm groups similar models on neighboring nodes. SOM was originally developed for visualization of distribution of vectors with some values. It grew out of earlier models of neural networks developed by Kohonen (Kohonen, 1984). Kohonen introduced systematic model which is consisted of at least two interactive subsystems of different nature. One of them is competitive neural network which is executed with function „winner takes it all“, while other subsystem is controlled by neural network and it changes local synaptic plastics of neurons while learning. Learning itself is limited by space to neighborhood of most active neurons (Scholarpedia, 2013).

SOM can be explained mathematically in purely abstract form, without any kind of reference to neural or other components that it is made of. Learning is executed through series of iteration based on formula (1):

$$m_i(t+1) = m_i + \alpha(t)h_{ci}(t)[x(t) - m_i(t)] \quad (1)$$

where $m_i(t+1)$ represents new value iteratively originated from old value $m_i(t)$ and new data $x(t)$; $\alpha(t)$ is scalar factor that defines size of correction and whose values decrease with step t . Index i is model for sub processing, while c is index of model which is closest to $x(t)$ by Euclidean distance. Factor $h_{ci}(t)$ is function of neighborhood, and it decreases when distance between models grows (Kohonen and Honkela, 2011). Mathematical theory of SOM is very complex and complicated, and only one-dimensional case has been analyzed entirely (Fort, 2006). Algorithm is being executed by following steps:

1. Calculation of distance between data and all neurons of SOM;
2. Choosing the closest neuron as winner;
3. Refreshment of each neuron by iterations;
4. Repeat steps 1, 2 and 3 and refreshment of parameters of learning, while certain criteria for stopping algorithm is achieved.

There is large number of SOM variants. Some of them represent basic SOM algorithm with some small changes, while others suffered great changes and do not have same

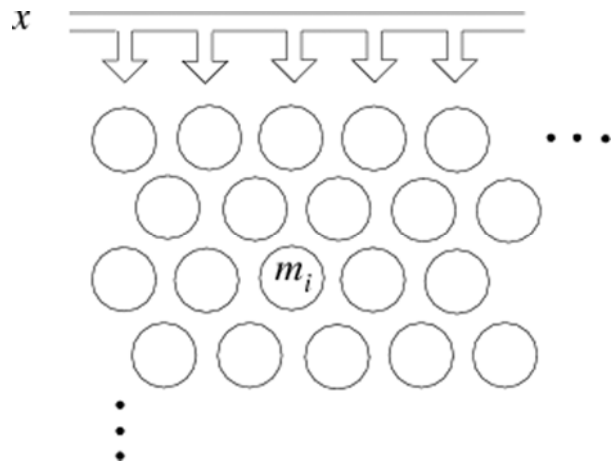


Figure 3: Grid of nodes in two-dimensional SOM, where x are data projected onto models m_i , where each model is associated to one node

characteristics of mapping and visualization (Lobo *et al.*, 2004). Some include geographical characteristics of data. Introduction of Tobler's first law of geography "Everything is related to everything else, but near things are more related than distant things" (Tobler, 1970) into learning of SOM would imply search for best neuron inside SOM for specific data set, while only geographically close neurons would have been searched. Close to this approach are Hypermaps (Kohonen, 1991), where only part of input data are used for finding best neurons, or Kangas architecture (Kangas, 1992), where only small number of neurons close to best previous neuron are being searched. Combination of these two ideas led to Spatio-Temporal Feature Map (Chandrasekaran and Palaniswami, 1995), which use function of spatial gates, together with similar function of time gates, to choose next neuron (Lobo *et al.*, 2004). There are also hierarchical SOM, which instead of one SOM use larger number of different SOMs, while each of them uses part of data. Results of each of these lower level SOMs are projected on higher level SOM. This enables each lower level SOM to specialize for specific aspect of data, delivering more intuitive interpretation of data (Lobo *et al.*, 2004). With georeferenced data it is wise to use geographical coordinates as direct input to SOM of highest level together with summarized information of lower level SOMs.

Idea of using geography with SOM is not brand new. There are several examples of SOM algorithms that introduce geography as basic characteristic of data. One of successful ones is work of scientists from School of Statistics and Information Management of New University of Lisbon (Portugal). Bação, Lobo and Painho

engaged with possibilities of using SOM for determination of homogenous regions and detection of spatial patterns (Bação *et al.*, 2008). They have explored idea of creating tool that enables user to explore spatial data, with emphasis on fuzzy nature of most classification methods. Tool would enable „What if?“ types of analysis and would have visualization of results. Based on these settings, they developed GeoSOM (Bação *et al.*, 2008), tool that have several variants of basic SOM algorithm incorporated: classical SOM, hierarchical SOM and GeoSOM, adjusted for processing of spatial data.

Problem of determination of regions opened both problematic and opportunistic questions in geography. On one hand, bad effects of strict regionalization with administrative and statistical regions are great ballast for geographic research (Openshaw 1984; Fotheringham and Wong 1991; Amrhein 1995), while on the other hand possibilities have opened to use them for the development and evolution of informatics and math, and transformation of problem into significant tool for spatial analysis (Openshaw 1984; Wise *et al.* 1997; Guo *et al.* 2003). Basic premise for definition of homogenous regions is typology of regions. Hagget *et al.* (1977) suggest three types of regions: uniform regions, nodal regions and planning regions. Planning regions are designed in advance, with defined purpose, such as census tracts, or planning regions etc., while uniform and nodal regions are of more explorative nature (Bação *et al.*, 2008). Differences between regions can have significant influence on development of algorithms, because algorithms for uniform and nodal regions must enable larger interaction on researcher-system relation (Bação *et al.*, 2008).

The most important step of SOM training algorithm in establishing the system to define which samples are grouped, is the one where Best Matching Units (BMU) are being chosen (Bação *et al.*, 2008). In basic SOM it is accomplished by comparison of all components of input data with all components of each unit, while GeoSOM can include spatial coordinates. The importance of spatial coordinates is defined with parameter k , which represents geographic tolerance of processing neighboring data. If $k=0$, algorithm will recognize the closest geographical location as BMU. If k is equal to size of map, space does not play role in execution of algorithm, so the

result is the same as result of basic SOM.

For experiment of delineation of functional urban areas in Serbia with GeoSOM algorithm, software called GeoSOM 2.1 was used. It was developed on School of Statistics and Information Management of New University of Lisbon (Portugal). GeoSOM is capable of executing algorithm of basic SOM, GeoSOM and hierarchical SOM. For input data, polygons of settlements NUTS5 level were used, with attributes from 2002 census for each settlement: total population, active population, active population that is employed, agricultural population, population that is employed in

primary, secondary and tertiary sector of economy. Matrix of 8000 neurons has been chosen in system where $x=20$ and $y=40$. For geographic component, spatial coordinates of polygon centroid were used. Execution was performed in three iterations: in first basic SOM algorithm was executed (Figure 4), in second GeoSOM algorithm with geographical tolerance $k=0$ was executed (Figure 5), and in third GeoSOM algorithm with geographical tolerance $k=2$ was executed (Figure 6). Finally, hierarchical SOM which combines results of first and second iteration was executed (Figure 7)

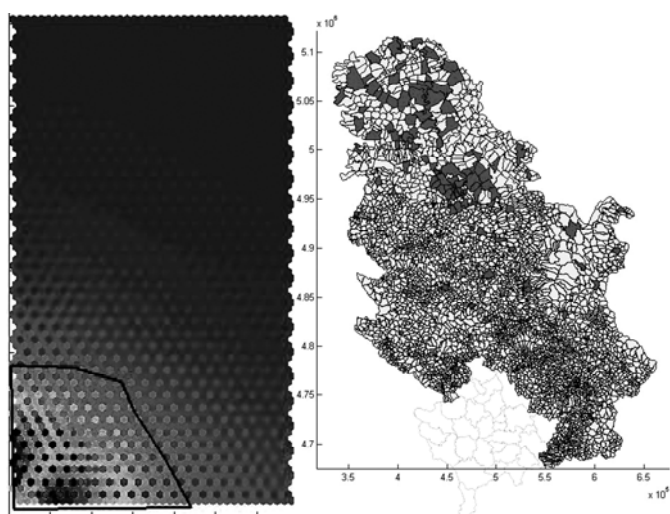


Figure 4: SOM matrix of neurons and resulting cluster pointing to urban centers of Serbia (light grey – processed data, dark grey – results, white – no data)

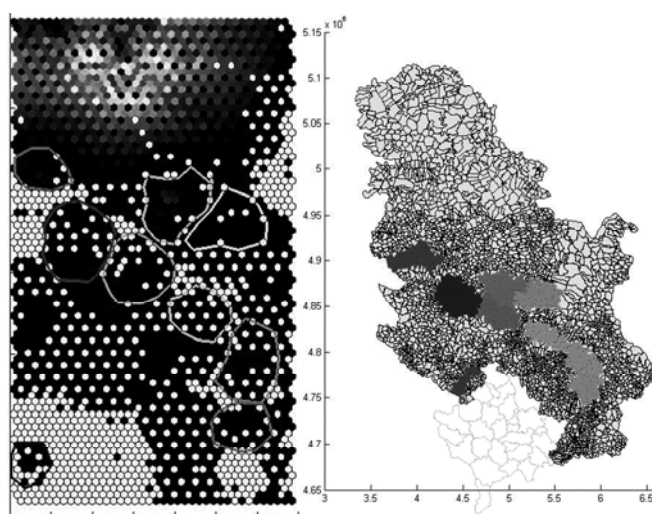


Figure 5: GeoSOM matrix of neurons and some chosen clusters for $k=0$ – full respect of geographical neighborhood, so clusters are spatial continuous in all directions (light grey – processed data, darker shades of grey – results, white – no data)

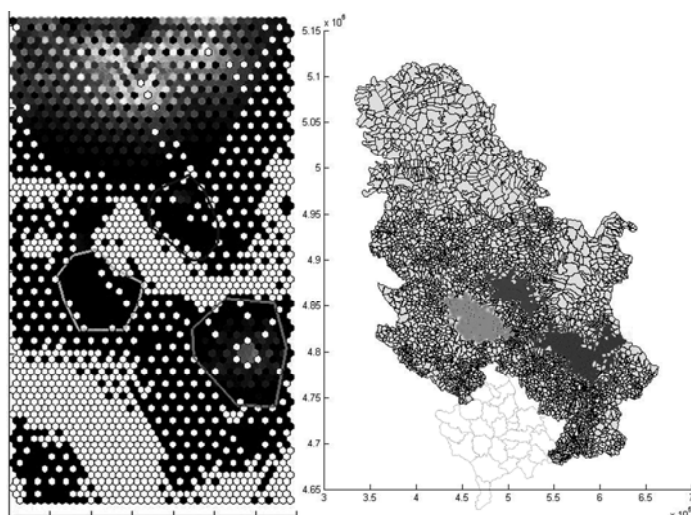


Figure 6: GeoSOM matrix of neurons and some chosen clusters for $k=2$ – less respect to geographical neighborhood, so clusters are dispersed (light grey – processed data, darker shades of grey – results, white – no data)

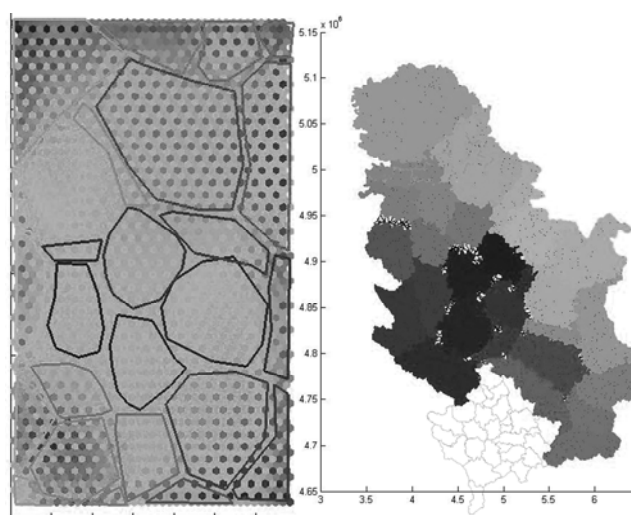


Figure 7: Matrix of neurons for hierarchical SOM which is result of basic SOM - Figure 4 and GeoSOM with $k=0$ - Figure 5 (light grey – processed data, darker shades of grey – results, white – no data)

COMPARATIVE ANALYSIS OF MACHINE LEARNING METHODS AND RESULTS AGAINST ESPON METHOD FOR DELINEATION OF FUNCTIONAL URBAN AREAS

Analysis of *k*-means clustering method

K-means clustering method implies clustering of points around centroid of respective cluster. It represents the most famous and most commonly used algorithm for clustering. Its strength is in its own simplicity and efficiency, while weaknesses are represented through its basis: initial centroids. They can have great impact on results, and it is the essential problem to choose them correctly. The number of initial centroids defines the number of clusters to be created, so the results are limited to that number. Problem is, also, geographical isolation of certain points, which can have impact on size and shape of clusters.

In two iterations of CrimeStat 3.0 software with earlier described settings, results of 25 clusters were generated, for each iteration respectively: one with user-defined initial centroids and other with software's initial centroids. Level of urbanity according to model of Tošić (2012) was used as weight, while 2002 population size was used as intensity. Resulting clusters (Figure 2) from both iterations are different in size and shape, but also in resulting centroid which varies from initial centroid. This is the result of repeating of the algorithm steps until state of convergence is achieved. The result of iteration with user's predefined initial centroids is that those centroids can belong to one or more clusters. Spatially, they can be located on the edge of cluster or in its center, while there are cases that some clusters do not have any of 25 predefined initial centroids – urban cores. The weakness of this method is also in limited amount of features attributes that can be used in algorithm. CrimeStat 3.0 can use only 2 attributes, in this case level of urbanity and population size, while other tools can use more feature attributes, but results become unclear.

Analysis of Self-Organizing Maps method

Self-Organizing Maps (SOM) project data on regular, usually two-dimensional matrix of nodes – neurons, where data are mapped to neuron whose model is calculated with SOM algorithm and is most similar to data, e.g. is closest to data in some measure. Numerous variants of SOM algorithm exist, with small or large changes against basic algorithm. Some

variants include geographical characteristics of statistical data, based on first law of geography. This is how GeoSOM tools have been made, and which were used in this experiment. There are also hierarchical SOM algorithms, which combine results of multiple SOM and those results are being projected to higher level SOM, so lower level SOMs can be specialized for different aspect of data, enabling more intuitive interpretation of data. If georeferenced data are used, it is wise to use geographical coordinates as direct input for highest level SOM together with summarized information of lower level SOM.

Such method has been implemented through GeoSOM 2.1 software. Input data were polygons of settlements with attributes from 2002 census. Several iterations were performed. The first iteration was with basic SOM algorithm, and results were clusters which present statistical characteristics of data, but not geographical. This is how urban centers of Serbia were determined through first cluster, while other settlements were grouped into 6 other clusters, according to statistical data. Second iteration included GeoSOM algorithm: geography has been used as input. Algorithm has been executed against same data with variations of *k* parameter from values 0 to 2. This parameter explains the extent to which the principle of geographical neighborhood will be used, so when value is 0, the principle is absolutely important, and resulting clusters are continuous in all directions. With value of 2, this principle becomes less important and resulting clusters become more dispersed. Finally, hierarchical SOM algorithm has been executed against results of basic SOM (Figure 4) and GeoSOM with *k*=0 (Figure 5). The results are homogenous regions in Serbia, clearly determined in neuron matrix. Using this method, 27 homogenous regions of Serbia were determined. Advantage of this method is that there is no limit in number of attributes used and also more intuitive analysis of specific data in different SOM algorithm, which can ultimately be combined and projected into hierarchical SOM.

ESPON method

ESPON methodology (ESPON, 2005) implies existence of urban core with at least 15,000 inhabitants and area with at least 0.5% of population of a country, with functions of national or regional importance. Isochrone of 45 minutes of car travel through road network is calculated from urban core towards periphery. Resulting isochrone area is approximated on NUTS5 level, so at least 10%

of territory of respective settlement must be under isochrone area to be considered as part of FUA. As a final result there is PUSH area, which can be overlapped with neighboring PUSH areas.

There were 49 FUA identified with this method (Figure 8), whose areas do overlap. Population size and isochrone have great impact on results, but usage of other data should be examined, such as daily migration or settlement typologies. Also, reduction of input parameters of population and isochrone should be examined, e.g. increasing the population size of urban core to 20,000 inhabitants and usage of 30 minute isochrone.

Comparative analysis

Analysis of used methods and results points out significant differences in basic premises of each method. Although each of them treats the same space and data, results are different. Results represent regions whose area and shape are predisposed by spatial component of central places; such is with *k*-means and ESPON method. On the other hand, SOM uses principles of analysis of statistical data with spatial component, without predefined central place as initial core around which delineation of city limits should be done. Example is city of Kraljevo (Figure 9), where results of each method are being shown. SOM method gives area of 3,020 km² and population of 266,000 inhabitants, while *k*-means gives area of 2,145 km² and population of 346,000 inhabitants. ESPON method gives area of 3,415 km² and population of 515,000 inhabitants. These results should be taken into account with reserve, because PUSH area of Kraljevo includes two neighboring large urban centers, Čačak and Kragujevac, and smaller ones – Vrnjačka Banja and Trstenik, so population size is significantly enlarged. *K*-means also includes territory and population size of neighboring centers, while SOM gives the most realistic results, since principle used determines homogenous regions, in this case nodal ones.

Comparison of results shows that methods give different output, but surely there is place for combining methods. This is especially case with SOM and ESPON, since there is a possibility of combining multiple statistical data for same territorial unit– NUTS5 level, and intuitive analysis of data over multiple dimensions can be used. Also, different principles of spatial analysis can be combined: first law of geography, and principles of mobility and accessibility. It is necessary to continue research in area of Self-Organizing

Maps and determination of homogenous regions with focus on enlargement of used attributes and fine tuning of algorithm, and in combination with other methods of spatial statistics and artificial intelligence.

CONCLUSION

Functional urban areas as basic blocks of polycentrism in EU space are determined by using different methods from country to country. ESPON gave standard method which presumes existence of urban core with 15,000 inhabitants at least and important functions, around which isochrone of 45 minutes travelling by car is delineated. Resulting area is approximated to NUTS5 level, where the principle where at least 10% of NUTS5 unit must be covered by 45 minute isochrones, so it can belong to respective PUSH area. Neighboring PUSH areas can overlap and that overlap represents area with high potential for integration.

In addition to this method, there are other methods that can be used for delineation of FUA. Some of them originate from artificial intelligence, grouped into machine learning tools. The best known are clustering algorithms and artificial neural networks.

K-means clustering algorithm is one of most famous, which works on principle of grouping points around centroid of respective cluster. Its strength is in simplicity and efficiency, but weakness is in its basis: initial centroids which have great impact on results, so they must be chosen well. Number of initial centroids defines the number of resulting clusters, so method is limited to that number. Also, position of geographically isolated points can impact shape and size of clusters. K-means algorithm was executed within CrimeStat 3.0 software on centroids of settlements of Serbia with integrated data of 2002 census and some calculated indexes, such level or urbanity is. Two iterations have been performed: one with 25 user predefined initial centroids of settlements with important administrative functions, and second one where software calculated initial centroids. Results gave 25 clusters – polygons of different size and shape, which in most cases do not take into account the importance of statistical data, but just principle of proximity.

From group of neural networks, algorithm of Self-Organizing Maps (SOM) was explored. It performs projection of data on matrix of nodes-neurons, while data are mapped to neuron whose model is calculated with SOM algorithm and is most similar to data, e.g. is closest to data in some measures. There are numerous

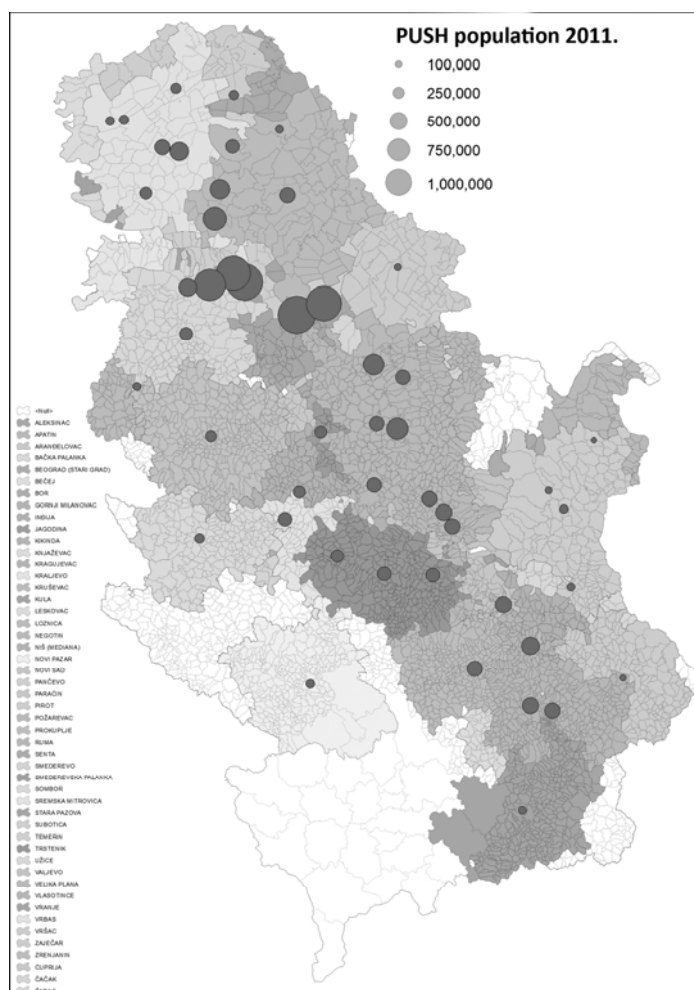


Figure 8: PUSH areas of Serbia and population size according to 2011. Census

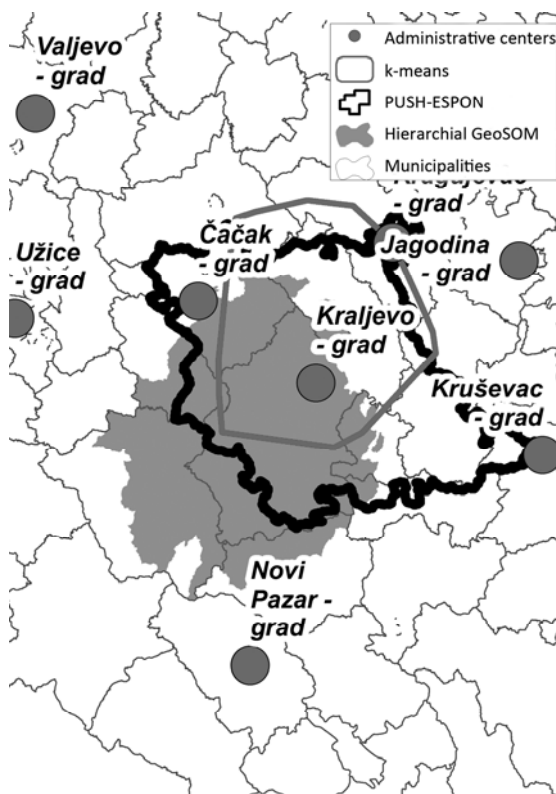


Figure 9: Comparison of FUA Kraljevo gained with methods ESPON, k-means clustering and GeoSOM

variants of SOM, with smaller or greater changes of basic algorithm. Variant which includes geographical characteristics of statistical data based on first law of geography was used in this paper. Those are called GeoSOM tools. There are also hierarchical SOM algorithms, which combine results of multiple different SOMs, so resulting data are projected on higher level SOM, which enables that each lower level SOM can be specialized of specific aspect of data, and give more intuitive interpretation of data. If input is represented with georeferenced data, it is wise and recommended to use geographical coordinates as direct input to higher level SOM together with summarized information of lower level SOMs. SOM has been executed through variant of basic SOM algorithm with input of settlement polygons (NUTS5 level) with 2002 census data incorporated. Then, data were projected onto GeoSOM where variance of parameter k between values 0 and 2 has been used to define level of importance of geographical neighborhood. Results from both iterations were then executed within hierarchical GeoSOM, which gave output of 27 homogenous nodal regions.

Used methods and respective results were compared, and conclusion is that neither method gives exclusively good results. K -means should not be applied for purpose of delineation of functional urban areas, since it is limited to attribute data that can be used and is simple expression of proximity of points in space, while central point is moveable. SOM algorithms represent excellent basis for purpose of delineation of FUA, especially if they are combined into hierarchical GeoSOM, where it is possible to specialize multiple different SOM algorithms for specific sets of spatial and statistical data. Research in this direction will be continued, together with fine tuning and change of algorithm, and combination with other methods of spatial and statistical analysis.

References

- Amrhein, C. G. (1995) *Searching for the elusive aggregation effect: evidence from statistical simulation*, Environment and Planning A 27, pp. 105-120
- Baço F., Lobo, V., Painho, M. (2004) *Geo-Self-Organizing Map (Geo-SOM) for Building and Exploring Homogeneous Regions*, in Egenhofer, M.; Miller, H.; Freksa, C.(Eds.); GIScience 2004 Chandrasekaran, V., M. Palaniswami. (1995): *Spatio-temporal Feature Maps using Gated Neuronal Architecture*, IEEE Transactions on Neural Networks 6(5), pp. 1119-1131.
- ESPON (2005), *ESPON 1.1.1 Potentials for polycentric development in Europe - Project report*, Luxembourg, The ESPON Monitoring Committee: Luxembourg.
- Esri ArcGIS Resources (2013) <http://resources.arcgis.com>
- Fort, J.C. (2006), *SOM's mathematics*, Neural Networks, 19, pp. 812-816.
- Fotheringham, A. S., Wong, D. W. S. (1991) *The modifiable areal unit problem in multivariate statistical analysis*, Environment and Planning A 23, pp. 1025-1044.
- Haggett, P., Cliff, A. D., Frey, A. E. (1977) *Locational Analysis in Human Geography*, Second Edition, London, Arnold.
- Harnad, S. (2008) *The Annotation Game: On Turing (1950) on Computing, Machinery, and Intelligence*, in Epstein, Robert; Peters, Grace (eds.), *The Turing Test: Sourcebook: Philosophical and Methodological Issues in the Quest for the Thinking Computer*, Kluwer.
- Kangas, J. (1992) *Temporal Knowledge in Locations of Activations in a Self-Organizing Map*, Proceedings of the International Conference on Artificial Neural Networks, Brighton, England, pp. 117-120
- Korcelli, P. (2008) *Review of Typologies of European Rural-Urban Regions*, PLURAL
- Kohonen, T. (1982) Clustering, Taxonomy, and Topological Maps of Patterns, *Proceedings of the 6th International Conference on Pattern Recognition*, Munich, pp. 114-128
- Kohonen, T. (1991) The Hypermap Architecture, in: T. Kohonen, K. Mäksä, O. Simula, J. Kangas (eds.) *Artificial Neural Networks*, 1: Helsinki, Elsevier pp. 1357-1360
- Kohonen, T. (2001), *Self-Organizing Maps*, Springer-Verlag New York
- Kohonen, T., Honkela, T.(2011), *Kohonen network*, Scholarpedia
- Liu, B. (2007) *Data Mining and Text Mining*, lecture on Department for Computer Sciences at University of Illinois, (UIC), USA.
- Lobo, V., Baço, F., Painho, M. (2004) Regionalization and homogeneous region building using the spatial kangas map, in F. Toppen and P. Prastacos (eds.) *7th AGILE Conference on Geographic Information Science*, Heraklion, Greece, pp. 301-313.
- MacQueen, J. B. (1967) Some Methods for classification and Analysis of Multivariate Observations, Proceedings vol 5., *Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, University of California Press, 1, pp. 281-297.
- Matteucci, M. (2012) *A Tutorial on Clustering Algorithms*, http://chrome.ws.dei.polimi.it/index.php/Matt%27s_Home_Page
- Moore, A. (2004) *K-means and Hierarchical Clustering Tutorial*, School of Computer Science, Carnegie Mellon University, <http://www.autonlab.org/tutorials/kmeans.html>
- Openshaw, S. (1984) *The Modifiable Areal Unit problem*, CATMOG, Geo-abstracts, Norwich, UK.
- Samuel, A. (1959) Some Studies in Machine Learning Using the Game of Checkers, *IBM Journal* 3 (3), pp. 210-229
- Tobler W. (1970) A computer movie simulating urban growth in the Detroit region, *Economic Geography*, 46(2), pp. 234-240
- Tošić, D. (2012) *Principi regionalizacije*, Geografski fakultet Univerziteta u Beogradu: Beograd.
- Turing, A.M. (1950) Computing machinery and intelligence, *Mind*, 59, pp. 433-460
- Vesanto, J., Alhoniemi, E. (2000) Clustering of the Self-Organizing Map, in *IEEE Transactions on Neural Networks*, Volume 11, Number 3, pp. 586-600. © 2000 IEEE
- Wikipedia (2013), article: Dijkstra's algorithm, http://en.wikipedia.org/wiki/Dijkstra%27s_algo_rithm
- Wikipedia (2013), article: Machine learning, http://en.wikipedia.org/wiki/Machine_learning
- Wikipedia (2013), article: Polycentrism, <http://en.wikipedia.org/wiki/Polycentrism>
- <http://www.gdi.net>
- <http://www.esri.com>

Received March 2013; accepted in revised form June 2013