INTEGRATION OF SPATIAL DATA REPRESENTING BUILDINGS BY DETERMINING THE DEGREE OF SIMILARITY

Ing. Renata Ďuračiová, PhD., Ing. Magdaléna Igondová

ABSTRACT

Spatial data integration is the process of combining heterogeneous data to make them compatible, maintain data accuracy and actuality, minimize redundancy, and avoid data conflicts. Identification of spatial objects is one of the most important tasks in spatial data integration. In this paper, we propose to use determination of the similarity measures for assessment of spatial object identification. The result is an implementation of the proposed procedure into the geographic information system. In the case study we update the ZB*GIS*® database by the data from the real estate cadastre system. The innovation of the presented method is that the proposed process and the new implemented software tools are based on determination of the similarity measures such as the Dice similarity index and the Jaccard similarity coefficient with its modification as the Tanimoto similarity coefficient.

Key words: *spatial data integration, identification, similarity measure, GIS, buildings*

1 INTRODUCTION

When analyzing and using spatial data from different data sources, it is often necessary to determine the mutual identity of objects that are stored with different geometrical representation in different data sources. For example, it is unlikely that a building or a road would be represented in two different data sources by the same polygon or polyline. Identification of spatial objects in geographic information system (GIS) is necessary in process of spatial data integration and spatial analyses (Flowerdew, 1991; Shekhar and Xiong, 2008). In this paper, we propose to determine the identity or similarity of two geometrical representation of spatial polygon objects by calculating the degree of their similarity. We use the similarity measures such as the Jaccard coefficient and the Dice similarity index. The result is a proposal for a procedure to easily determine the likely identity of objects or, on the contrary, to identify new spatial objects in external data sources. In our study, we propose the new process of identification of buildings stored in the layer Buildings in the ZB*GIS*® database (the fundamental database for GIS in Slovakia) and spatial objects representing buildings in the real estate cadaster system. Calculation of degrees of similarity we implemented into the ArcGIS software environment by creation of the new tools *Similarity of KN buildings and ZBGIS buildings* and *Selection of buildings*. Using these tools, it is possible, for example, to select new or significantly modified buildings from the cadastral database, by which the layer Building of the ZB*GIS*® database can be quickly and efficiently updated.

2 MATERIALS AND METHODS

In GIS, we are working with digital representation of real objects. One object of reality can be represented in GIS in various ways, which implies that its representations are not the same in general. In assessing whether it is a representation of the same (unmodified) spatial object we propose to use the determination of their mutual similarity.

2.1 Similarity of geometrical representations of spatial objects and its use in spatial data integration

The similarity measure is used to express the degree of similarity between two objects. In GIS, similarity of spatial objects is used, for example, in clustering. The goal of clustering is to find or create sets of objects whose elements are the most similar to each other, but the elements of two different sets should be as different as possible. To determine the degree of similarity, various metrics of distance (Euklidean, Manhattan, Minkowski, Mahalanobis, etc.), correlation coefficients (e.g. Pearson or Spearman), or association rates (Sokal-Michener, Russell-Rao, Jaccard, etc.) are used.

In this work, when determining the similarity of two spatial objects, we apply the concept of similarity between two sets A and B (we consider the geometrical representation of the object as a set of points). Two sets A and B are the same if $A = A \cap B = B$, which is equivalent to: $A \subseteq B \wedge B \subseteq A$ (Bandemer, 2006). The similarity of the sets can be then described in simple way as follows: two sets are similar, if they are approximately the same. Interpretation of the vague term "approximately the same" is based on the following concept: the set of objects outside the intersection $A \cap B$ is small compared to the union $A \cup B$ (Bandemer, 2006), which corresponds to the Jaccard similarity coefficient (Jaccard, 1901):

$$Sim_{Jaccard}(A,B) = \frac{|A \cap B|}{|A \cup B|}. \tag{1}$$

To calculate the degree of similarity of two spatial polygon objects using the formula (1), we replace the cardinality of the set with the area of the polygon.

The second suitable method of calculation of the degree of similarity is to use the Dice similarity index (Dice, 1945; Schubert, 2013):

$$Sim_{Dice}(A,B) = \frac{2|A \cap B|}{|A| + |B|} = \frac{2|A \cap B|}{|A| / |A \cap B| + |B| / |A \cap B|} = \frac{2|A \cap B|}{|A \cup B| + |A \cap B|} \tag{2}$$

Both degrees of similarity take values from the interval $\langle 0,1 \rangle$. If the two objects or their geometrical representations are identical, their mutual similarity is 1 and if the spatial polygon objects have no intersection, the value of their mutual similarity is 0. Values close to 1 then point to the probable identity of the object represented in two different data sources. On the contrary, values close to 0 allow to select new or significantly modified objects, which one of the data sources does not contain. Therefore, this principle can be used to integrate spatial data from different data sources.

2.2 Implementation of spatial data integration using similarity measures in the ZB*GIS*® database update

In our case study, the calculation of the similarity measures we used for updating the layer Buildings of the ZB*GIS*® database (the ZB*GIS*® buildings) by the spatial data representing buildings in the real estate cadastre system (the KN buildings - the KN abbreviation expresses the real estate cadastre in Slovakia). For determining the similarity of the ZB*GIS*® buildings and the KN buildings based on the formulas (1) and (2), four following layers are required to calculate the similarity measure of two spatial objects (graphical representations of the real buildings):

- A - the layer of the ZB*GIS*® buildings with their attributes,
- B - the layer of the KN buildings,

- $A \cap B$ - the intersection of the ZB*GIS*® buildings and the KN buildings,
- $A \cup B$ - the union of the ZB*GIS*® buildings and the KN buildings.

Both layers (the intersection and the union) can be created in a common GIS software environment. The whole calculation procedure for determining the Jaccard similarity coefficient of two spatial polygon objects is presented in the paper (Ďuračiová, 2014).

However, the computation of the Jaccard similarity measure is problematic if the union operator is used for polygons that have a common line. For example, as seen in Figure 1, two buildings (9 and 6) form one building (13) after unification. Consequently, we would not determine the similarity between the objects 6-A and 9-B but between the objects 13-A and 13-B.

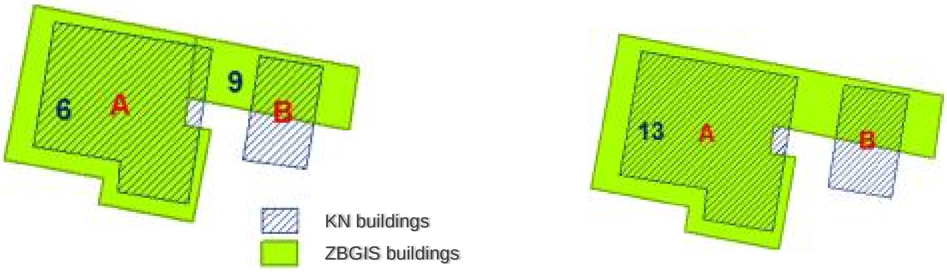


Fig. 1 Negative consequence of the use of the union operator in determining the degree of similarity

To solve this problem, in this paper we propose to express the Jaccard similarity coefficient by the Tanimoto similarity coefficient (without using the union operator):

$$Sim_{Tanimoto}(A,B) = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \tag{3}$$

Note, that the the Tanimoto similarity coefficient is another expression of the Jaccard similarity coefficient, since $|A| + |B| - |A \cap B| = |A \cup B|$.

Consequently, we can replace the original procedure with a much simpler one. For example, the following procedure can be applied into the ArcGIS software environment:

1. Create (or select) the KN layer (the layer that contain the geometrical representation of buildings in the real estate cadastre system),
2. Add the Area_ZBGIS attribute to the feature class ZBGIS buildings (using the *Add field* tool) and calculate the area of polygons (buildings) using the *Calculate Geometry* tool,
3. Add the Area_KN attribute to the feature class KN buildings and calculate the area of each polygon,
4. Create the INTERSECTION feature class (the intersection of the ZBGIS Buildings and the KN buildings),
5. Add the ID_INTERSECTION attribute and the Area_INTERSECTION attribute to the INTERSECTION feature class,
6. Join the feature classes KN buildings, ZBGIS buildings, and INTERSECTION into the JOIN layer (using the *Spatial Join* tool),
7. Create the DICE attribute (the Dice similarity index) and the JACCARD attribute (the Jaccard similarity coefficient) and calculate them using the formulas (1) - (3) and the *Field Calculator* tool.

3 RESULTS

Based on the proposed procedure for calculating the similarity measures, we created the *Similarity of KN buildings and ZBGIS buildings* tool. To implement the new tool, we used the ArcGIS software, specifically the *ModelBuilder* environment, which allows to create custom tools using existing tools from the *ArcToolbox* environment (including a number of spatial analytic functions).

The sample input layers used for testing the *Similarity of KN buildings and ZBGIS buildings* tool are shown in Figure 2. We used data from the cadastral territory Kočovce (district Nové Mesto nad Váhom, Slovakia). The example of the output layer created using the Dice similarity index is demonstrated in Figure 3.

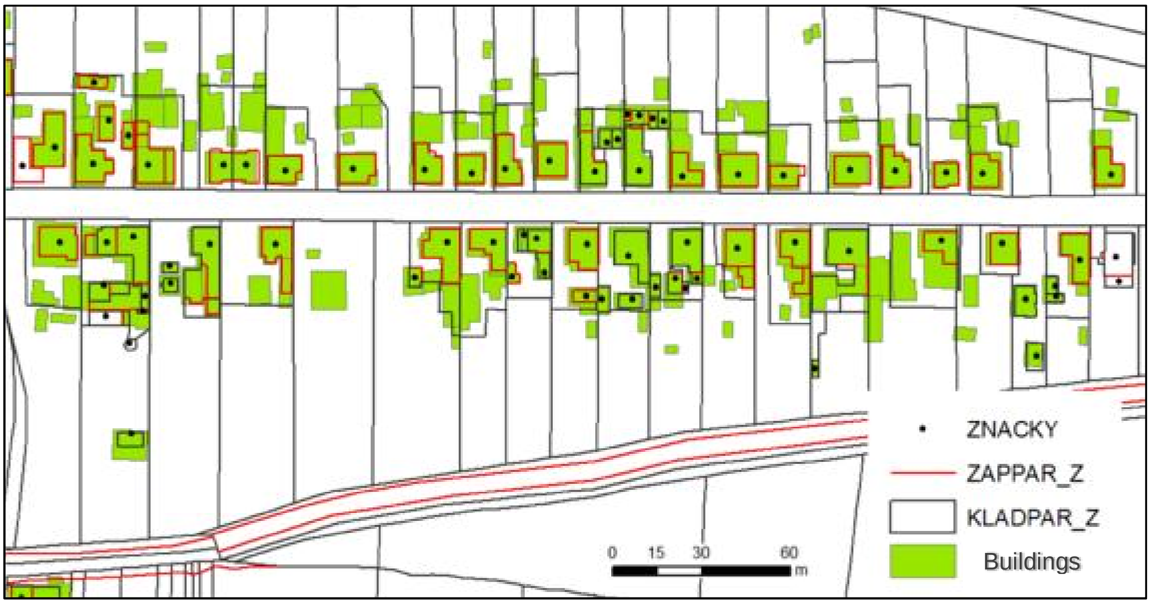


Fig. 2 The ZBGIS buildings (the layer Buildings) and the KN buildings (created based on the layers ZAPPAR_Z, KLADPAR_Z, and ZMNACKY - the names of the layers are derived from the official names used in the real estate cadastre system in Slovakia)

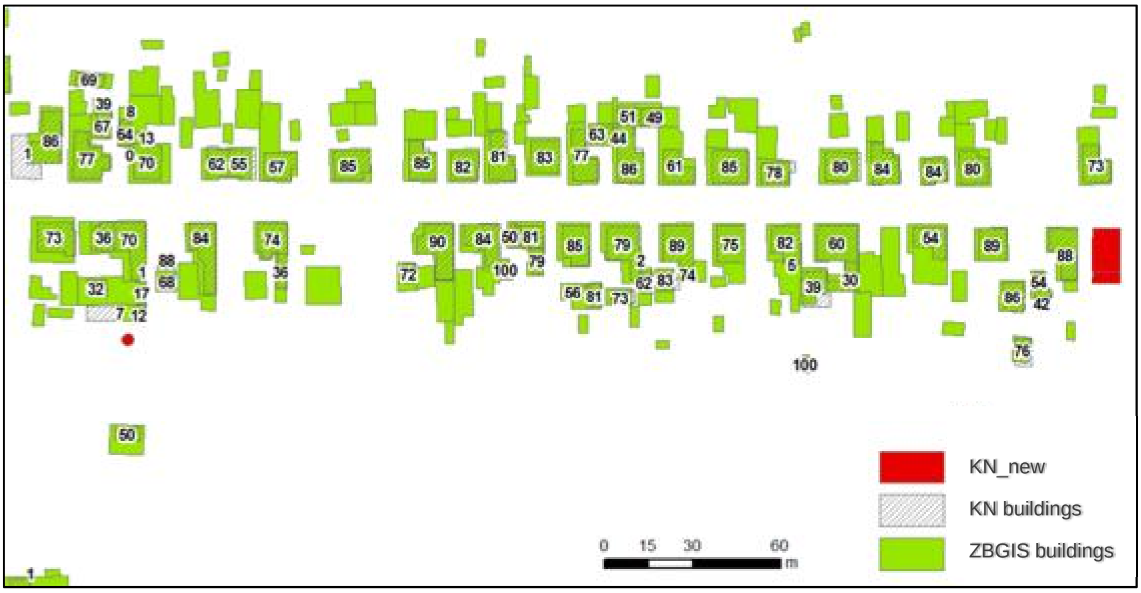


Fig. 3 The output layer (created based on the Dice similarity index): the KN_new layer contains the new buildings (the KN buildings that are missing in the ZBGIS buildings)

The problem situation can occur in the case shown in Figure 4 (left). It is clear that if the KN building (10624) intersects two (22877 and 22058) or more ZBGIS® buildings, the similarity measures for the KN building are calculated for each intersected ZBGIS® building separately (82 % and 7 %) (Fig. 4 (left). This means that if we assume that new buildings in process of spatial database updating are, for instance, the buildings with degrees of similarity less than or equal to 60 %, all buildings that have at least one degree of similarity less than or equal to 60 % will be selected (Fig. 5 (left)). This is the reason why we also created the second tool called *Selection of Buildings*.

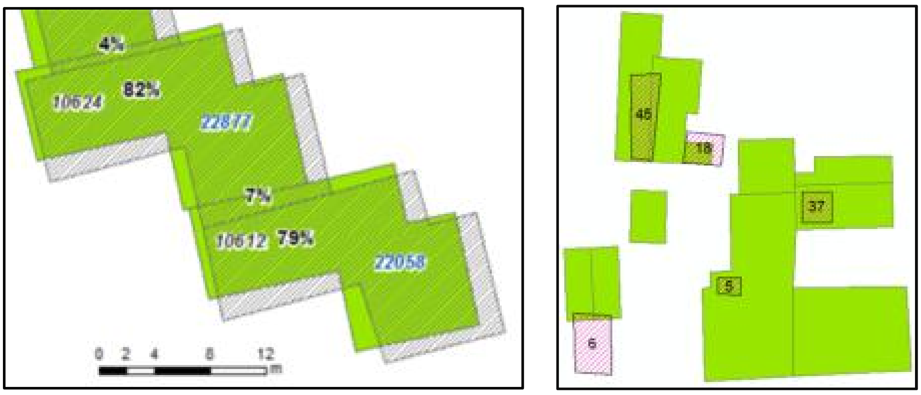
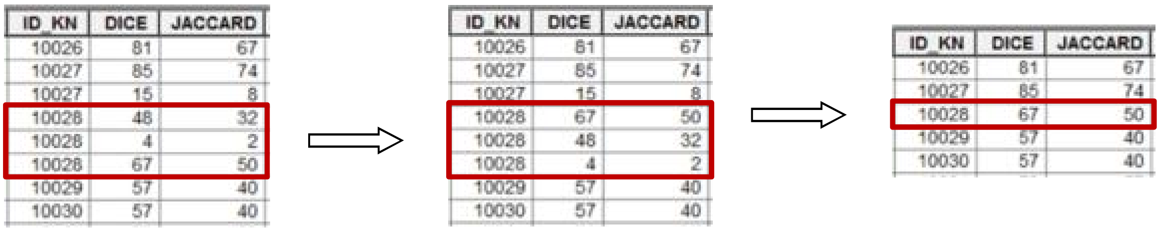


Fig. 4 The KN buildings that intersect two or more ZBGIS buildings (left) and the KN buildings that are completely included in ZBGIS buildings (right)

The *Selection of buildings* tool is designed to select new KN buildings based on degree of similarity determined using the *Similarity of KN buildings and ZBGIS buildings* tool. It is used to select objects with multiple values of similarity and it selects only the buildings with the highest value of the similarity measure. As a sample, the original attribute table (left), the attribute table after using the *Sort* tool (in the middle), and the result obtain using the *Delete Identical* tool (right) are shown in Figure 5. To implement the new tools, the ArcGIS *ModelBuilder* software environment we used.



ID_KN	DICE	JACCARD
10026	81	67
10027	85	74
10027	15	8
10028	48	32
10028	4	2
10028	67	50
10029	57	40
10030	57	40

ID_KN	DICE	JACCARD
10026	81	67
10027	85	74
10027	15	8
10028	67	50
10028	48	32
10028	4	2
10029	57	40
10030	57	40

ID_KN	DICE	JACCARD
10026	81	67
10027	85	74
10028	67	50
10029	57	40
10030	57	40

Fig. 5 The sample of modification of an attribute table using the proposed software tools

Subsequently, we selected new buildings based on the selected interval of the similarity measure. Using the *Select* tool in the *Expression line* is possible to specify a selection condition (e.g. "select all buildings where similarity (determined by the Dice similarity index or the Jaccard similarity coefficient) is less than or equal to 60 %").

After use of the *Building Selection* tool, all buildings where the degree of similarity is less or equal to 60 % are selected. Figure 4 (right) shows that the buildings (37, 5, 45) completely covered by the ZBGIS® buildings are also included in the selection. If we want to remove these buildings from the selection, it is necessary to calculate the area of the KN buildings and their intersection with the ZBGIS® buildings. Then it is possible to calculate how many percent of the KN building is included into the ZBGIS® building. Therefore, we determined a degree of inclusion of the building *B* in the building *A*:

$$Incl(A,B) = \frac{|A \cap B|}{|B|} \qquad (4)$$

The calculation of the degree of inclusion we implemented into the *Selection of buildings* tool. Details of the implementation of both tools are described in the work (Igondová, 2016), and the tools are available at: 147.175.19.15/similarity.

4 CONSLUSIONS

In this work, we deal with assessing the possibility of integration of heterogeneous spatial data sources using calculation of the degree of similarity or inclusion. For this purpose, we used the similarity measures that numerically express the similarity of two spatial objects either in the range 0 to 1 or in the percentage. We created two software tools, which are especially useful for recording the spatial objects represented new spatial polygon objects (e.g. buildings). We used the tools described in this paper for updating the ZB*GIS*® database by spatial data representing buildings in the real estate cadastre system. We tested the functionality of the tools on the spatial databases of the cadastral district of Kočovce (district Nové Mesto nad Váhom) and on two versions of the ZB*GIS*® database of the cadastral district of Dúbravka (Bratislava) (Igondová, 2016). Using the newly created tools *Similarity of KN buildings and ZBGIS buildings* and *Selection of buildings* it is possible to get valuable information about spatial objects, especially new buildings, and so efficiently update the ZB*GIS*® database using data from an external source.

ACKNOWLEDGEMENTS

This work was supported by the Grants No. 1/0682/16 and No. 1/0954/15 of the Grant Agency of Slovak Republic VEGA.

Literature

[1] BANDEMER, H.: Mathematics of uncertainty: Ideas, methods, application problems. Berlin: Springer Verlag, 2006, 190 p., ISBN 978-3-540-28457-4.

[2] DICE, L. R.: Measures of the Amount of Ecologic Association Between Species. Ecology, 26(3), 1945, pp. 297-302.

[3] ĎURAČIOVÁ, R.: Identifikácia rôznych reprezentácií priestorových objektov v geografických informačných systémoch na základe určenia mier podobnosti / Identification of spatial objects in geographic information systems based on determination of their similarity measure. In: Aktivity v kartografii venované Jánovi Pravdovi 2014: Zborník referátov zo seminára, Bratislava, SR, 23.10.2014. 1. vyd. Bratislava: Kartografická spoločnosť SR, 2014, pp. 17-27. ISBN 978-80-89060-23-8 (in Slovak).

[4] FLOWERDEW, R.: Spatial Data Integration, Geographical information systems, 1991 [online] Available at: http://www.msu.ac.zw/elearning/material/1344175939spatial%20data%20integration.pdf

[5] IGONDOVÁ, M.: Integrácia priestorových dát na základe určenia mier podobnosti / Spatial Data Integration Based on Determining Similarity Measures, Diploma thesis. Bratislava: Faculty of Civil Engineering, Slovak University of Tehnology in Bratislava, 2016, 48 p. (in Slovak).

[6] JACCARD, P.: Étude comparative de la distribution orale dans une portion des Alpes et des Jura. Bulletin de la Soci_et_e Vaudoise des Sciences Naturelles, 37, 1901, pp. 547-579.

[7] SHEKHAR, S., XIONG, H.: Encyclopedia of GIS. New York: Springer, 2008, 1370 p. ISBN 978-0-378-30858-6

[8] SCHUBERT, A., TELCS, A.: A note on the Jaccardized Czekanowski similarity index, 2013 [online] Available at: http://www.cs.bme.hu/~telcs/PUBS/note%20on%20jcz.pdf