Study of the method of spatial identification of polygonal features in the integration of geospatial datasets

*Y. Havryliuk, A. Lyashchenko (Kyiv National University of Construction and Architecture)

**SUMMARY**

Integration of geospatial datasets is complex and time-consuming process. Integration is especially difficult if the geospatial datasets do not share common feature identification attributes. In this study, we explored the software implementation of the method of spatial identification polygonal features of geospatial datasets by detecting matching pairs of features from two different datasets using measures of their spatial overlap and the similarity of their morphometric characteristics (perimeter, area, width-to-length ratio, blockiness, number of vertices). The polygonal feature identification program is implemented in PL/pgSQL for a spatial database in PostgreSQL/PostGIS. The results of the experiment on the example of the 13832 polygonal building models for the town of Bila Tserkva using an OSM dataset and a digital topographic plan of 1:2000 scale showed the ability to perform spatial identification of about 85% of buildings with little or no user intervention.
Introduction

The rapid development of Geographic Information Systems (GIS) and geoinformation web services is driving an ongoing growth of available sources for geospatial data. Many data providers offer different points of view on the same phenomenon: geospatial features. We can gain additional benefits and new knowledge by integrating data from different sources. However, this integration requires solving the problem how to match features of two vector geospatial datasets by comparison of their geometric properties. This process named geospatial data matching (Xavier et al., 2016, Gösseln & Sester, 2004, Safra et al., 2006).

We define spatial identification of features as the process of selecting pairs of objects from two of geospatial datasets by evaluating the match of features geometric properties. Spatial feature identification is a key tool for selecting candidate features for the subsequent integration of the two geospatial datasets. Tools of features spatial identification are important in solving problems in many applications, such as: updating topographic datasets, spatial matching of geosciences datasets and topographic data objects geometric properties, data quality assessment, and others.

The tools of spatial identification of objects are specific as for various types of spatial localization of objects (point, polyline, polygon, so for the purposes of applied use of these tools. The authors of the article (Xavier et al., 2016) performed a fundamental systematization of 30 years research in measures and methods of geospatial data sets reconciliation.

Many authors have explored and proposed different approaches to the polygonal features spatial identification for the reconciliation of the geospatial datasets (Ruiz-Lendínez et al., 2017, Jiyoung Kim & Kiyun Yu, 2015, Jung Ok Kim et al., 2010, Yong Hut et al., 2011). Note that, despite a large amount of research in this area, GIS software still does not offer good tools for automating the process of spatial identification of objects, with minimal involvement of a human GIS operator.

In this abstract, we present the results of the development of an applied SQL function for spatial identification of polygonal objects of geospatial datasets in the PostgreSQL object-relational database system with the PostGIS spatial extension. This SQL function designed for use in integration of geospatial datasets in a GIS of urban-planning cadastre (Lyashchenko & Cherin, 2019).

Method and Theory

To assess the matching of two polygons we used the Nikodym metric (Gryunbaum, 1971). On the sets $A, B \subseteq \mathbb{R}^2$, the absolute Nikodym metric can be defined as:

$$\rho(A, B) = S(A \cup B) - S(A \cap B),$$

where $S(A \cap B)$ is the area of the intersection of polygons $A$ and $B$, $S(A \cup B)$ is the area of the union of polygons $A$ and $B$. The relative Nikodym metric is defined as:

$$\mu(A, B) = \frac{S(A \cup B) - S(A \cap B)}{S(A \cup B)}$$

Since

$$S(A \cup B) = S(A) + S(B) - S(A \cap B),$$

then in order to calculate the distance between the sets of polygons $A$ and $B$ points according to the Nikodym metric, it is enough to know the area of polygon $A$, the area of polygon $B$ and the area of their intersection $S(A \cap B)$. The estimate of the matching of polygons $A$ and $B$, which intersect, can be defined as $M(A, B) = 1 - \mu(A, B)$. Matching evaluation for the two polygonal objects (Figure 1) using the Nikodym metric has the following values: a) $M(A, B) = 1$; b) $M(A, B) = 0.499$; c) $M(A, B) = 0.143$. In the general case, this evaluation changes in the range $1 - 0$, where $1$ corresponds to the complete coincidence of two polygons, and $0$ – to non-coincidence. If two polygonal features have an evaluation of $M(A, B) > 0.4$, then they can be identified as similar features. To make the final decision...
on the spatial identification of two polygonal features, we use additional morphometric characteristics of polygons (area, perimeter, size as ratio of total width to length, blockiness, number of vertices), as well as auxiliary estimates of the similarity measure of polygons by these characteristics. Each auxiliary estimate $C_i(A,B)$ of the similarity of two polygons by their morphometric characteristics $P_i(A)$ and $P_i(B)$ is calculated as:

$$C_i(A,B) = \frac{\min ((P_i(A), P_i(B))}{\max ((P_i(A), P_i(B))}.$$

Thus, each of the auxiliary estimates of the similarity of two polygons varies in the range 0 - 1.

Figure 1 The example of superimposing the contours of a building from two data sets: a) the contours of the building are almost identical; b) the contours of the building are shifted by 15 m; c) the contours of the building are shifted by 30 m.

As a hypothesis of coincidence of two polygonal features from different geospatial datasets we use following matching rule (MR): if two polygonal features have an estimate $M(A,B) > 0.3$ and each auxiliary estimate of the similarity measure $C(A,B) > 0.6$, then two polygonal features are spatially identified as coincident.

Implementation and experiment

In our experiment, we simulated a typical situation where, due to the lack of vector data for a city, local government departments use geospatial data from available open sources, for example, Open Street Map (OSM). Then, after the official vector data of the city topographic plan is created, there is a need to integrate OSM datasets and official vector data. The purpose of this integration is to supplement the feature attributes of the official datasets with the attributes of dataset features created using OSM data. We used two geospatial datasets with polygonal building models for the town of Bila Tserkva (Figure 2). OSM data was used as a target dataset for spatial identification of buildings using vector data of digital topographic plan of 1:2000 scale as a reference dataset.

Figure 2 Test datasets on the example of Belaya Tserkov: a) OSM dataset as a target set for spatial identification; b) a dataset of the topographic plan M 2000 as a reference set

For the Geospatial Database (GDB) in PostgreSQL/PostGIS, we have developed two PL/pgSQL functions called MatchGeomCheck and DefineFeature_ID. The function MatchGeomCheck (aGeom
The DefineFeature_ID ( ) : text function is designed for spatial identification of polygonal features of the target dataset, which is stored in the TargetSet table of the GDB, based on an analysis of their similarity with the objects of the reference dataset, which is stored in the RefSet table.

The DefineFeature_ID function implements the following algorithm: 1) for each object from the TargetSet dataset intersecting with the objects of the RefSet dataset, using the basic function MatchGeomCheck (a_geom, b_geom), the similarity of the TargetSet features with the RefSet dataset features is estimated; 2) if the similarity estimates of an object from the TargetSet dataset correspond to the MR rule, then a record (TS_ID, RS_ID, M (A, B), {C_i (A, B)}) is inserted into the TargetSet_ID results table, where: TS_ID is the object identifier in TargetSet table; RS_ID – identifier of the matching object from the RefSet table; M (A, B), {C_i (A, B)} – a set of object similarity estimates calculated using the MatchGeomCheck function. 3) calculation of summary statistics of the process of spatial identification of the TargetSet dataset, namely: the total number of processed objects, the number of identified objects according to the MR rule, the number of objects with similarity scores 0 < M(A,B) < 0.4 and M(A,B ) = 0.

The results of testing the developed functions on datasets for the city of Belaya Tserkov (Figure 2) are presented in Figure 3 and in the table of generalized statistics of the process of spatial identification of the 13832 buildings from the OSM dataset using the dataset of the M 2000 topographical plan.

**Figure 3 Result of spatial identification of the building polygons of the OSM dataset (red outlines) using the reference dataset of the M 2000 toographic plan**

**Table 1 The statistics of the process of spatial identification of the buildings of the OSM dataset**

<table>
<thead>
<tr>
<th>The similarity estimate of features M (A,B)</th>
<th>Number of features</th>
<th>Average area of feature, m²</th>
<th>Average values of auxiliary estimates of the similarity of objects C_i (A, B) as the ratio of morphometric characteristics of objects:</th>
</tr>
</thead>
<tbody>
<tr>
<td>range</td>
<td>average</td>
<td></td>
<td>areas</td>
</tr>
<tr>
<td>M = 0</td>
<td>0</td>
<td>763</td>
<td>271.70</td>
</tr>
<tr>
<td>0 &lt; M &lt; 0.3</td>
<td>0.21</td>
<td>1405</td>
<td>328.16</td>
</tr>
<tr>
<td>0.3 &lt;= M &lt; 0.5</td>
<td>0.41</td>
<td>4567</td>
<td>195.45</td>
</tr>
<tr>
<td>0.5 &lt;= M &lt; 0.75</td>
<td>0.61</td>
<td>6041</td>
<td>282.37</td>
</tr>
<tr>
<td>N &gt;=0.75</td>
<td>0.82</td>
<td>1056</td>
<td>1758.31</td>
</tr>
</tbody>
</table>
Conclusions

In this study, we explored our software implementation of the method of spatial identification polygons features of geospatial datasets by detecting matching pairs of features from two different datasets using measures of their spatial overlap and the similarity of their morphometric characteristics. Testing the developed PL/pgSQL functions on real geospatial datasets of close large scales showed the ability to perform spatial identification of about 85% of buildings with little or no user intervention. Further improvement of the proposed tools for spatial identification of polygonal features is possible by supplementing them with tools for assessing the semantic similarity of features of two geospatial datasets using the values of attributes of selected pairs of features, as well as tools for analyzing multi-overlap of features in datasets with different scales of representation.

Reference


