ENSURING THE FUNDAMENTAL SPATIAL DATA QUALITY FOR SUCCESSFUL IMPLEMENTATION OF THE CZECH NSDI

Jindra Marvalová, Karel Janečka

Ing. Jindra Marvalová
University of West Bohemia
Univerzitni 22, 306 14 Pilsen, Czech Republic
Tel.: +4206028062831, Email: jindram@kma.zcu.cz

Ing. Karel Janečka, Ph.D.
University of West Bohemia
Univerzitni 22, 306 14 Pilsen, Czech Republic
Tel.: +420377632691, Email: kjanecka@kma.zcu.cz

Abstract
In October 2014, the Czech government approved the conception of The Strategy for the Development of the Infrastructure for Spatial Information in the Czech Republic to 2020 (GeoInfoStrategy), which serves as a basis for the National Spatial Data Infrastructure (NSDI). From the fundamental spatial data point of view, the Register of Territorial Identification, Addresses and Real Estates (RTIARE) can be considered as a cornerstone of the Czech NSDI. The data registered in RTIARE serve as reference data (it means actual, valid and without further verification) for other information systems and decision-making processes in public administration. Therefore, it is crucial to ensure the completeness and accuracy of the data. However, initially RTIARE database was filled by importing data from several data sources that were inconsistent in some cases. By this way some incorrect data got into the register. Administrator of RTIARE, Czech office for Surveying, Mapping and Cadastre (COSMC), solves this problem, it periodically controls and corrects data and its quality improves but without using a geometry part of geographical data. To improve the quality of controls the geometry part has to be used. This paper describes the solution of selected controls, focused on buildings registered in RTIARE based on using both attribute and geometry parts of RTIARE data. A total of four controls were carried out: searching buildings which have identical (near) definition points, searching buildings which have suspicious numbers (house registration), search buildings which have a common bond to the same parcel in cadastre and searching buildings which have a definition point of a building and a definition point of a parcel far more than the maximal distance. Some of the controls were solved by the application of computational geometry algorithms, whereas in some cases the new algorithms had to be proposed. All the controls were tested and successfully implemented using Oracle Spatial technology and its procedural language PL/SQL. The COSMC has already adopted the controls in the RTIARE production database.

Keywords*: NSDI, GeoInfoStrategy, RTIARE, data control*

INTRODUCTION

On 14th November 2012, the Czech government decided to prepare a conception of NSDI „The Strategy for the Development of the Infrastructure for Spatial Information in the Czech Republic to 2020" (GeoInfoStrategy). The conception of the GeoInfoStrategy was finally approved by the governmental resolution of 8th October 2014 (GeoInfoStrategy, 2014). The GeoInfoStrategy reflects the lessons learnt from the NSDIs implementations in other countries. An approach of the GeoInfoStrategy in building the national spatial data infrastructure is similar to the INSPIRE (INSPIRE, 2007) idea. INSPIRE, as one of the most important European activity in the field of spatial information, is the key European foundation for the GeoInfoStrategy.

The GeoInfoStrategy defines the principles and strategic aims for effective use of the spatial information in public administration. The GeoInfoStrategy is a conceptual material that also has a close relation to other strategic documents of public administration and eGovernment. The vision of the GeoInfoStrategy is that in 2020 the Czech Republic is a knowledge society effectively using spatial information. To fulfil this vision, it is necessary that spatial information and services will be used in every aspect of public life. The services that will be created with relevant spatial data will support the competitiveness, safety, social cohesion and sustainable development. The accessibility of spatial

information and services enables the public sector to offer the modern and high quality public services (ČADA, JANEČKA, 2016).

The system of basic registers, launching in the Czech Republic from 2012, is the central information source for information systems of public authorities. The system consists of four main basic registers:

- Register of inhabitants,
- Register of persons (companies),
- Register of territorial identification, addresses and real estates,
- Register of rights and responsibilities of public authorities.

The Register of territorial identification, addresses and real estates (RTIARE) can be considered as a cornerstone of the Czech NSDI. RTIARE contains territorial elements and RTIARE contains territorial elements. It is absolutely necessary to ensure that RTIARE contains only valid data.

The quality of RTIARE data is crucial not only from the perspective of national data, but also from the transnational perspective, because along with the Information System of the Cadastre of Real Estate data, the RTIARE data are published according to INSPIRE Directive and thus it becomes a part of the European data infrastructure.

CONTROLS OF RTIARE DATA DONE BY COSMC

The Czech Office for Surveying, Mapping and Cadastre (COSMC) is the responsible body for doing controls of the RTIARE data. As stated in (JANEČKA & HEJDOVÁ, 2014) the geometry of the features has not been usually considered during these controls. The outputs of these controls are identified errors which should be repaired by operators.

For example, COSMC has done the following controls of the RTIARE data without using geometries, just using attribute data:

- a control of address points without a definition point,
- a control of unrelated address points on the building objects with entrances,
- a control of streets without related address points,
- a control of streets,
- a control of building objects without relation to the parts of municipalities.

To make the controls more precise considering the geometry is crucial. (JANEČKA & HEJDOVÁ 2014) describe the geometry based controls of address points and streets.

CONTROLS OF BUILDINGS REGISTERED IN RTIARE

Methodology

Newly 4 controls are solved:

searching buildings which have identical (near) definition points,

searching buildings which have suspicious numbers (house, registration),

searching buildings which are linked to the same parcel in cadastre and

searching buildings which have a definition point of a building and a definition point of a parcel far more than the maximal distance.

The work methodology is as follows. First, problem situations that should be found are identified for each control. If possible, the problem is converted to the computational geometry problem. Available algorithms that can be used to solve the problem are searched and implemented or own solutions are suggested. The results of each solution method are compared in terms of results and time consumption and the most appropriate solution method for use in a RTIARE production environment is selected.

All algorithms were implemented and tested using Oracle Spatial technology and its procedural language PL/SQL, because RTIARE data are managed in Oracle database.

Two tests datasets for testing of algorithms were used. Test data are original RTIARE data downloaded in VFR[1] (specialized exchange format) using Public Remote Access[2] application. The first dataset (Pilsen dataset) contains the data of Pilsen municipality. The amount of the data is about 1% of the total RTIARE registry data. The second dataset (Regional cities dataset) contains the data of all regional cities in the Czech Republic. The amount of the data is about 10% of the total registry data. The Pilsen dataset is used for algorithms implementation testing. The most appropriate selected solutions are also tested on the Regional cities dataset.

For the data import and primary data processing the process described in (JANEČKA & HEJDOVÁ 2014) was used.

Searching buildings which have identical (near) definition points

The aim of this control is to find buildings that are probably duplicated in RTIARE. It means, that there are two or more records in the database for one real building. Therefore it is necessary to find duplicate definition points of buildings, but not only those points that have the same coordinates, but also the definition points of different buildings, that are too close so that it could be considered as identical. There is an example of searched situations in the figure 1.

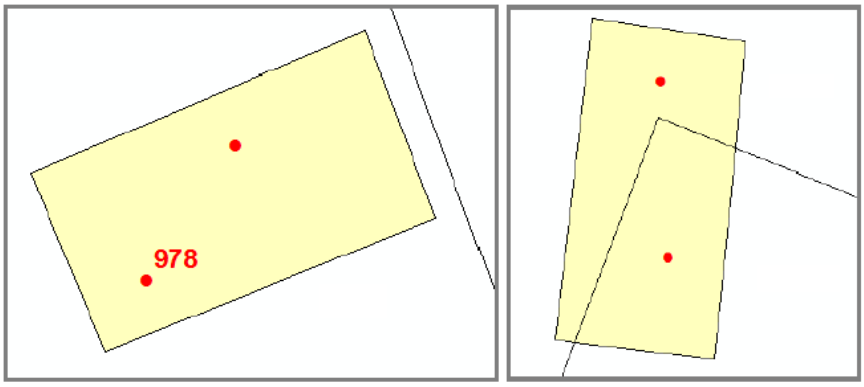

Figure 1. Near definition points of buildings. Definition points are red, building polygons are yellow.

The search of identical or close definition points can be converted to the geometric problem of Closest Point Problem (CPP) that is a subset of Near Neighbor Search (NNS). The CPP is described in (SMID, 1995), the NNS problem is described (Zhang, 2004). There were proposed many algorithms to address this problem, some of them are described e. g. in (CHÁVEZ, 2005), (KAMOUSI, 2014), (BRIN, 1995), (ARYA, 2008).

Three of the available algorithms were selected, implemented and tested. It was the brute force method (described in (LV, 2007)), spatial index R-tree (described for example in (ROUSSOPOULOS, 1995)) and sweep-line algorithm (described in (SMID, 1995)).

The R-tree index was selected due to the possibility of easy implementation, this index is a part of Oracle system and it can be assumed that its use will be effective. The sweep line algorithm was selected due to its popularity. The brute force algorithm implementation illustrates comparison, how the algorithm choice affects the time requirements of the solved problem.

Individual algorithms were implemented in the form of PL/SQL procedure with the minimal distance (i.e. the distance into which two points are considered to be the same) input parameter. As the most appropriate minimum distance parameter appears the distance of 2-3 meters. By the use of the higher minimum distance parameter there is a number of close buildings definition points pairs found, but in the fact these buildings are not duplicated (most often garages or other small objects).

[1] http://www.cuzk.cz/Uvod/Produkty-a-sluzby/RUIAN/2-Poskytovani-udaju-RUIAN-ISUI-VDP/Vymenny-format-RUIAN/Vymenny-format-RUIAN-(VFR).aspx

[2] http://vdp.cuzk.cz/

The time consumption of individual tested algorithms for selected minimum distance parameter on the test data is shown in table 1.

Table 1. Time consumption of tested algorithms on test data.

Algorithm	Minimum distance 3m		Minimum distance 5m	
	Time [s]	Time [min]	Time [s]	Time [min]
Brute force	59860	998	59875	998
R-tree	128	2	959	16
Sweep line	1098	18	1585	26

It is evident that in terms of the time consumption the most appropriate solution is the R-tree index algorithm. It was implemented using standard Oracle functions *SDO_NN* and *SDO_NN_DISTANCE* (see (ORACLE)). The created procedure was applied to the test datasets and the number of found close buildings definition points pairs for the minimum distance of 2, 4 and 6 meters is shown in table 2.

Table 2. Count of close buildings definition points pairs for the minimum distance of 2, 4 and 6 meters.

Data	2 m	4 m	6 m
Pilsen	355	7 084	11 021
Regional cities	6 355	77 946	118 848

The problem is that only according to the definition points distance it could not be clearly decided whether or not the buildings are duplicated. It is necessary to control other building attributes (such as the house number or registration number, belonging to the municipality and cadastre unit and other) or cadastral map insight in order to certainly determination, which close buildings definition points pairs are duplicated. In some cases the definition points closeness can be caused by their bad placement on the edge of the building polygon. According to the decree (Basic Registry Act, 2012) the definition points should be placed in the centroid of the building polygon.

Searching buildings which have suspicious numbers (house registration)

The aim of this control is to find buildings with suspicious house numbers or registration numbers. These numbers are buildings identifiers within the village. It means that house numbers and registration numbers are unique in the village and they do not repeat.

Numbers in the village should create continuous series of numbers that can be interrupted only in case that the building, which had assigned the number, was canceled. For this reason, large discontinuation of the continuous series of numbers are suspicious.

There are two approaches to address this problem. Specific numbers, in which neighborhood is on both sides (or only on one side if the number is at the beginning or the end of the number series) an excessively large discontinuation of the continuous series. The second approach is, that not the numbers but the discontinuations of the continuous series are suspicious.

To address this control some pre-processing is needed, it is in detail described in (MARVALOVÁ & JANEČKA, 2015). Two procedures were created to solve this problem. One of the looks for specific suspicious house and registration numbers and the second looks for suspicious discontinuation of the continuous series of numbers. Both procedures have input parameter discontinuation that in the first case determines the size of discontinuation around the numbers on both sides to consider the number as suspicious. In the second case the input parameter determines the size of discontinuation to consider it as suspicious.

The time consumption of the both procedures is about 25 seconds for the Region cities dataset. The number of suspicious house numbers and registration numbers for selected discontinuations in the Pilsen dataset is shown in table 3 and in the region cities dataset is shown in table 4.

Table 3. Count of suspicious numbers in the test datasets.

Discontinuation	Suspicious house numbers			Suspicious registration numbers		
	15	30	50	15	30	50
Pilsen	9	4	2	214	106	70
Region cities	137	102	87	929	583	417

Table 4. Count of suspicious discontinuations of numbers series in test data.

Discontinuation	Suspicious discontinuations of house numbers series			Suspicious discontinuations of registration numbers series		
	15	30	50	15	30	50
Pilsen	21	11	5	453	251	155
Region cities	406	231	170	1955	1288	940

This is determined by the fact, that among other temporary buildings are numbered by registration numbers. These buildings are canceled more often than other buildings and that is why registration numbers series contain more discontinuations.

Searching buildings which are linked to the same parcel in cadastre

The aim of this control is to search the buildings that are linked to the same parcel in cadastre. These are buildings that are probably duplicated in RTIARE or buildings that have an incorrect link to the parcel. Two procedures that use two methods were developed and tested to solve this problem:

1. Using the PL/SQL cursor that loads the building identifier and the linked parcel of the building. There are searched buildings for each row of the cursor hat have the link to the same parcel.
2. Using the JOIN operator that links the building table to itself using parcel identifier.

There was found 549 buildings linked to 272 parcels in Pilsen dataset. The time consumption on the Pilsen dataset is shown in table 5.

Table 5. Time consumption of PL/SQL cursor and JOIN operator on the Pilsen dataset.

Method	Time [s]
PL/SQL cursor	243,28
JOIN operator	0,06

Considering the significantly lower time consumption the use of the JOIN operator is more appropriate from the RTIARE production environment. Furthermore, this procedure was tested on the Region cities dataset. There 14743 buildings were found linked to 4856 parcels in 82 seconds. 4814 buildings in Region cities dataset are linked to the parcel without identifier. It means that they have no link to the parcel.

Searching buildings which have a remote definition point of a building and a definition point of a parcel

The aim of this control is finding buildings that have their definition point too far from the definition point of the linked parcel, it means incorrect buildings whose definition point does not lie on the parcel, on which the building stays, or buildings with the incorrect placed definition point. There is an example of searched situations in the figure 2.

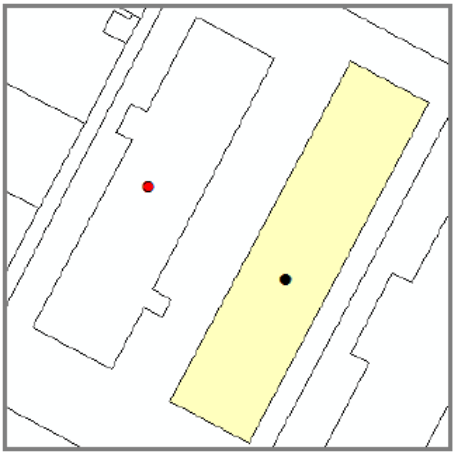

Figure 2. Remote building definition (red point) point and parcel definition point (black point)-incorrect placed building definition point. The building is highlighted in yellow.

Problem occurs with buildings that stand on more parcels. These are often residential buildings with more address places. In RTIARE, all buildings are linked only to one parcel, to the identifying parcel. It often happens in this cases that the building definition point (correctly placed in the building centroid) and the parcel definition point (correctly placed in the parcel centroid) are too far.

It is difficult to determine the maximum possible distance of the building and parcel definition point. This distance *m* depends on the parcel size and it is determined by the parcels acreage size.

Three methods were designed to the solution of this problem that deal in different ways with the determination of the maximum possible definition points distance and with the problem of the buildings that stay on more parcels. All methods are described in detail in (MARVALOVÁ & JANEČKA 2015). A brief description of the methods is:

Method 1: There is determined a number of intervals into that parcels are divided using Sturgess rule. Parcels are divided into the intervals using the natural breaks algorithm. The maximum definition points distance for each interval is determined.

Method 2: There is determined the maximum distance of the building definition point and parcel definition point for each parcel individually. This distance value depends again on the parcel acreage.

Method 3: A geometry intersection is used to find suspicious object instead of the definition points distance. It holds that the building definition point must be inside the building polygon and the linked parcel polygon. Conversely, for the buildings that lay on more parcels it holds that the parcel definition point must lay inside the building polygon. Unfortunately, there are buildings in RTIARE without polygon. In this cases method 2 is used.

The time consumption of all three methods on the Pilsen dataset and the number of found suspicious objects is shown in table 6.

Table 6. The time consumption and number of found suspicious object.

Method	Time [s]	Count of suspicious objects
Method 1	15,4	187
Method 2	15,3	164
Method 3	42,4	322

In order to compare method 1 and method 2 it would be necessary to check all found suspicious objects. Method 3, where the geometry intersection was used to resolve the control, gives the most confident results. Due to the results quality, method 3 appears to be the most appropriate solution for use in the RTIARE production environment despite the higher time demands.

Due to the results quality, method 3 appears to be the most appropriate solution for use in the RTIARE production environment despite the higher time demands. The PL/SQL procedure based on method 3 was applied on the Regional cities dataset. There were found 4917 suspicious objects in this data in 17 minutes.

CONCLUSION

Due to the wide use of RTIARE data the data quality is important. For this reason, the data are regularly controlled and new data controls are created. Due to the data variability and repeating of the controls the effectiveness of used algorithms is important. Newly, there were 4 controls created. The controls search for the objects that are suspicious of errors, because it is not possible to denote the object as incorrect only on the basis of the control results. The evaluation of other attributes and objects properties is necessary to denote the object as incorrect.

On the control results it is possible to create a list of objects that are suspected of errors. The list is a basis for a differential report creation. This differential report is passed to the RTIARE editors who perform the corrections of incorrect objects.

ACKNOWLEDGEMENT

The first author was supported by the Project SGS-2016-004 Application of Mathematics and Informatics in Geomatics III and the second author was supported by the project LO1506 of the Czech Ministry of Education, Youth and Sports.

REFERENCES

ARYA, Sunil, et al. An optimal algorithm for approximate nearest neighbor searching fixed dimensions. Journal of the ACM (JACM), 1998, 45.6: 891-923.

BRIN, Sergey. Near Neighbor Search in Large Metric Spaces. In*: 21th International Conference on Very Large Data Bases (VLDB 1995)*. Zurich, 1995.

ČADA, V.; JANEČKA, K. (2016) The Strategy for the Development of the Infrastructure for Spatial Information in the Czech Republic. *ISPRS International Journal of Geo-Information*, *5*(3), 33; doi:10.3390/ijgi5030033.

European Parliament. Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007: Establishing an Infrastructure for Spatial Information in the European Community (INSPIRE). Off. J. Eur. Union 2007, 30, 270–283.

CHÁVEZ, Edgar; FIGUEROA, Karina; NAVARRO, Gonzalo. Proximity searching in high dimensional spaces with a proximity preserving order. In: *MICAI 2005: Advances in Artificial Intelligence*. Springer Berlin Heidelberg, 2005. p. 405-414.

JANEČKA, K.; HEJDOVÁ, J. Validation of Data of the Basic Register of Territory Identification, Addresses and Real Estates. In: *Proceedings of the 5th International Conference on Cartography & GIS*, e-Proceedings, publisher: Bulgarian Cartographic Association, 2014, Riviera, Bulgaria, ISSN 1314-0604.

KAMOUSI, Pegah; CHAN, Timothy M.; SURI, Subhash. Closest pair and the post office problem for stochastic points. *Computational Geometry*, 2014, 47.2: 214-223.

LV, Qin, et al. Multi-probe LSH: efficient indexing for high-dimensional similarity search. In: *Proceedings of the 33rd international conference on Very large data bases*. VLDB Endowment, 2007. p. 950-961.

MARVALOVÁ, J., JANEČKA, K. Validation of Information and Relationships for Building Objects Registered in the Register of Territorial Identification, Addresses and Real Estates. Master's Thesis, University of West Bohemia, Faculty of Applied Science, Plzeň, Czech Republic. (In Czech).

ORACLE CORPORATION. Oracle Spatial User's Guide and Reference [online]. [cit. 2015-12-08]. Available online: https://docs.oracle.com/en/.

Basic Registry Act, No 111/2009 Coll; State Administration of Land Surveying and Cadastre: Prague, Czech Republic, 2012. (In Czech).

ROUSSOPOULOS, Nick; KELLEY, Stephen; VINCENT, Frédéric. Nearest neighbor queries. In: *ACM sigmod record*. ACM, 1995. p. 71-79.

SMID, Michiel. *Closest point problems in computational geometry*. Max-Planck-Institut für Informatik, 1995.

The Strategy for the Development of the Infrastructure for Spatial Information in the Czech Republic to 2020; Ministry of the Interior of the Czech Republic: Prague, Czech Republic, 2014. (In Czech)

ZHANG, Jun, et al. All-nearest-neighbors queries in spatial databases. In:*Scientific and Statis*tical Database Management, 2004. Proceedings. 16th International Conference on. IEEE, 2004. p. 297-306.