Dobesova, Z.(2012): The Geocode Correction in the Database as a Base for Spatial Analysis.
Advances in Computer Science, Proceedings of the 6th WSEAS European Computing Conference ECC 12 Prague, WSEAS Press, 511-516s.ISBN 978-1-61804-126-5

The Geocode Correction in the Database as a Base for Spatial Analysis

ZDENA DOBESOVA
Department of Geoinformatics
Palacký University in Olomouc
Tr. Svobody 26, 771 46 Olomouc
CZECH REPUBLIC
zdena.dobesova@upol.cz, http://www.geoinformatics.upol.cz/epracovnici-det.php?menu=zddo

Abstract: This article describes data correction in epidemiological database EPIDAT. The correction is aimed to correct all information about address of patients, place of infection and the get ill place of patients. Address and a location were compared against the valid Czech Territorial Identification Address Register UIR-ADR. Some records have been automatically repaired by SQL queries. There are also some irreparable records in EPIDAT. Trial database were taken for Olomouc Region for 10 selected diagnoses for period 2004-2008. Number of records was 24 thousand in trial database. The suggested universal steps of geocode repairing can be repeated to records about other 43 diagnoses or to newly recorded data from 2009 up to the present time. Moreover, the steps of semiautomatic correction can be applied to other 13 separate EPIDAT databases from each Regional Hygiene Station in the Czech Republic. The attribute correction is crucial for further spatial analysis of epidemiological data from the point of spreading diseases and spatial correlations.

Key-Words: Database, SQL, geocoding, epidemiological data.

1 Introduction

Nowadays a lot of data exists in databases. The quality of data limits the subsequent usage for the spatial health analysis. The results of analysis are not reliable if errors exist in the database. Research studies about the cancer spreading in USA pointed out the importance of the correct spatial information in cancer database. Geocoding errors generally had an adverse effect on statistical analyses of cancer data [1]. Verifying and evaluation of the accuracy of geocoded databases was recommended before public health studies in American Journal of Public Health [2]. Some attention can be found in the literature in the detailed house geocoding before epidemiological studies connected with traffic on the roads [3]. Address matching and disease distributions allows to be studied relative to ecologically associative environmental and socioeconomic status factors [4].

Department of Geoinformatics at Palacký University cooperates with Regional Hygiene Station in Olomouc. EPIDAT database was provided at the university. The university planned to realize some spatial analysis to identify spreading diseases, time evolution of diseases and identify spatio-demography correlation in Olomouc Region. The attributes about diagnose and time information were considered to be correct. The information about localization of the ill person we found sometimes inaccurate. Geocoding, as the correction of spatial information before follow-up spatial processing, was appeared necessary. Next sections describe content of EPIDAT database, Czech Territorial Identification Address Register UIR-ADR and mainly steps of correction of the geocoded attributes by SQL queries.

2 Spatial information in EPIDAT database

EPIDAT database stores information about the patients and their infectious diseases. Beside diseases, there are addresses of patients. It is necessary verify all data about localization connected with recorded phenomena for the subsequent correct assessment of the spatial distribution in Olomouc district. Both the address of patient and places of patient's infection and sicken were filled manually by an operator without verification to state address register.

2.1 Theory of geocoding

There are two methods how to determine spatial localization of features or phenomena on the globe from the point of geoinformatical theory. The first method is georeferencing. Feature is directly localized by the numeric coordinates X and Y in the

specific coordinate system. The second method is geocoding. Feature is indirectly localized by geocodes [5]. The geocodes are very often addresses, names of districts, ZIP codes, numbers of parcels etc. Geocode can relate features to named areas (district, cadastral area), to line (street, bus line) and point (house number, name of the bus station). Georeferencing is also called a positional geocoding. Second type, in which the assignment of a name or code for an area, line and point are used, is called nominal geocoding in the literature [5, 6]. The second way of localization - geocoding - is used in EPIDAT database for storing places of events connected with the infectious diseases.

2.2 EPIDAT database

EPIDAT database is used to ensure the mandatory reporting, recording and analysis of infectious diseases in the Czech Republic. EPIDAT is used nation-widely by Public Health Service of the Czech Republic from 1st January 1993. The program EPIDAT builds on ISPO (Information System for Communicable Diseases) from the years 1982-1992. Reports of infectious diseases are the basis for local, regional, national and international control of infectious diseases. Its legal basis is mandatory legislation: Act No. 258/2000 Coll. Data storage is used to secure exchange of actual data set on the prevalence of infections among the departments of Public Health Service of CR, Ministry of Health of CR and Public Health Institute in Prague.

Basic outputs of the EPIDAT are published in the journal Epidemiology and Microbiology Reports, which can be downloaded from the Internet. A monthly period running state of selected infections reported in the CR can be monitored on the web-site of the Public Health Institute.

Total of 53 diagnoses of infectious diseases are monitored into the EPIDAT data-base. The database is continually filled with data by employees of Regional Hygiene Stations. They are rewriting the medical records of reported infections from medical facilities. Each record contains 50 attributes. In terms of geocoding, the most important attributes are “STREET”, “TOWN” and “DISTRICT” defining the patient's residence, next “PLACE_OF_INFECTION” (where the patient was infected) and “PLACE_OF_SICKEN” (the place where the patient became ill, often place of clinic or doctor's office).

Program for data input was not updated since 1993, and the development of the new version has been halted. The program interface is outdated. There are no tools to prevent erroneous data, especially in term of issues of geospatial location of records.

Selected records from the EPIDAT database were used for this work. The provided data set from EPIDAT database contains only 10 diagnoses of infectious diseases from 53 diagnoses. They are (in parentheses is used the abbreviation): salmonellosis (A02), viral intestinal infections (A08), Lyme disease (A69.2), tick-borne encephalitis (A84.1), unspecified viral encephalitis (A86), viral meningitis (A87.9), varicella (B01), hepatitis A (B15), acute hepatitis B (B16), mumps (B26). All diagnoses data were available for the years 2004-2008.

Data was provided by the Regional Hygiene Station for Olomouc Region. The identification of land units were connected only with municipalities of Olomouc Region. Data was provided without personal, patient's information. The name, surname, identity number and full address (identification numbers of houses) of the patients are available only in the original EPIDAT database. The trial data set contained a total of 23,999 records for Olomouc region for 10 diagnoses.

2.3 Database UIR-ADR

The Ministry of Labor and Social Affairs of the Czech Republic guarantees one of special address register - UIR–ADR database (Territorial Identification Address Register). This register contains registers of regions, districts, NUTS4 districts, municipalities, towns, town districts, post offices, streets, public grounds and numbers of houses in the Czech Republic [8]. Register UIR-ADR is updated and disseminated to users for free. Latest version of database structure is 4.2. This register was taken as the national standard for geocoding of records in EPIDAT database.

Two tables based on UIR-ADR database were used for correction of geocoding in EPIDAT database. There are Town and Street correction tables. The two-letters code as district abbreviation was added to these tables by SQL query from UIR-ADR data-base. The third table CadastralArea was taken from the ISKN by Czech Office for Surveying, Mapping and Cadaster for correction of geocodes [9]. The structures of tables are in Fig. 1. Number of records is Town-62,262, Street-72,532 for the Czech Republic. Number of record for Cadastral Area is 13,028 in Olomouc Region. Important information is that one or more cadastral areas belong to one town. The information about a name of cadaster appears in a postal address sometimes as a place of a post office. This situation

product some mistakes and ambiguous in EPIDAT database. The structure of tables was expanded by the duplicate attribute ID_District and DistrictAbbreviation for different steps of correction. The values of TownName attribute in the Town table were converted to capital letters because EPIDAT database contains the names of town only in capital letters without diacritics. All three tables are called for this project UirAdr Correction Tables.

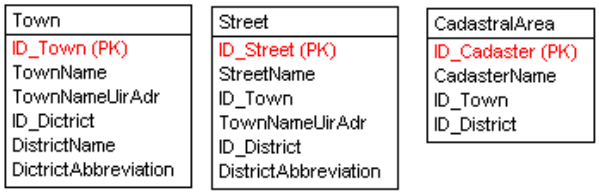

Fig.1 Structure of the Correction Tables based on UIR-ADR database and Cadaster

3 Process of the data correction

Manual correction of values is impossible for 24,000 records. We tried to find types of mistakes and ways how correct them. The first inspection discovered error values in the name of towns. Abbreviations and typing errors appeared in the town names. The name of town was used in three attributes in EPIDAT database: TOWN in the permanent address, PLACE_OF_INFECTION and PLACE_OF_SICKEN.

The second type of mistakes appeared from these administrative situations. Names of streets do not exist in small villages. Another problem is a situation that small villages belong to bigger villages, and some suburb belongs to a large town. Operator putted a name of a small village or a suburb to the address incorrectly. The address contained incorrect information.

The third type of mistakes was a non-corresponding information in combination of TOWN, DISTRICT and STREET. E.g. the town belongs to another district, or the street does not exist in this town but exists in another town. The solution of these mistakes was the worse.

Seven attributes were finally corrected: TOWN, STREET, abbreviation of name of DISTRICT, PLACE_OF_INFECTION, PLACE_OF_SICKEN, DISTRICT_OF_INFECTION, DISTRICT_OF_SICKEN.

Firstly the attribute TOWN (part of permanent address) was corrected separately from other values of attributes. District and street correction are possible only with simultaneous verifying of the name of the town. The new attribute CADASTER was added to every record as product simultaneous correction of town, street and district.

The stages of correction were two:

- Correction of the typing error in every of five geocoding attribute separately.
- Correction of non-corresponding values in the same record.

Some records were corrected repetitively in differ attributes. It was needful to display old values and new values for repetitive correction for the comparison purpose. It was solved by keeping the original old values and adding new attributes with correct values. New seven attributes were added to the structure of EPIDAT database. The prefix CORRECT was added to all names of attributes (attribute for the correct district was named CORRECT_TOWN). New seven attributes for messages about correction for each mentioned attribute were added to the database structure. Furthermore, one new attribute was added for cadastral area (CA).

3.1 Correction of abbreviations and typing errors

The process of correction consists of several steps. The first is the collection of distinct error for every attribute and creation of special two columns Error Table. The second step was the manual filling of "Error Table" with correct values from UirAdr Correction Table based on register UIR-ADR (mentioned in section 2.3). The third step was automatic update all wrong records in EPIDAT database according new attributes in Error Table with filling CORRECT attributes.

The process of correction is a combination of automatic and manual corrections. The manual stage is necessary because setting the correct values are sometimes difficult and depends on operator geographic knowledge especially in combination of more errors in one record (wrong street is mixed with the name of local part of town and the wrong name of district). The combination of automatic identification of errors, manual filling of “Error Tables” and automatic update of records in EPIDAT brings satisfactory results in the data correction.

Names of towns were filled from the simple internal register or filled manually by an operator in original EPIDAT program at Regional Hygiene Station. This register contains only town in capitals letters and without diacritics. Multiword names contain abbreviations in the internal register (e.g. BYSTRICE P.HOSTYN., Bystřice pod Hostýnem is correct). Typing errors were also frequent.

TOWN	CORRECT_TOWN
BRODEK U PROSTEJ.	BRODEK U PROSTEJOVA
BYSTRICE P.HOSTYN	BYSTRICE POD HOSTYNEM
CELECHOVICE NA H.	CELECHOVICE NA HANE
DOMASOV N.BYSTRIC	DOMASOV NAD BYSTRICI
DOMASOV U STERNB.	DOMASOV U STERNBERKA
DVUR KRALOVE N.L.	DVUR KRALOVE NAD LABEM
HRADEC N.MORAVICI	HRADEC NAD MORAVICI
HUSTOPECE N.BECV.	HUSTOPECE NAD BECVOU
KOVALOVICE-OSICA.	HUSLENKY
LOUCNA	HUSTENOVICE
MEROVICE N.HANOU	HUSTOPECE
MESTO LIBAVA	HUSTOPECE NAD BECVOU
MILOTICE N.BECVOU	HUTISKO-SOLANEC
MIMO UZEMI CR	HUZOVA
NOVE MESTO NA MOR	HVEZDLICE
OLSANY U PROSTEJ.	HVEZDONICE

Fig.2 Example of Error Table for correction of abbreviations in town name

The same repetitive mistakes appeared. Only distinct mistakes-the names of towns were added to "Error Table". The number of distinct records in "Error Table" of towns was 21. The total number of repaired records in EPIDAT database according Error Table for the town was 323.

3.2 Correction of streets and districts

Very problematic was the correction of attribute STREET. The names of streets are missing very commonly in small villages, in the Czech Republic. The name of cadaster is incorrectly used in a postal address (the name of a small village) instead of the name of town with a post office. Under this influence, operator filled to EPIDAT database to the attribute STREET the name of village (cadastral name) or another wrong local name. The problem was solved by adding extra attribute CA (cadastral area).

Correction of streets and districts was followed correction of towns. The steps of correction of the street and districts were the same as correction of town. Another Error Table of the street was automatically filled by distinct errors from EPIDAT. Right information is manually added according UirAdr Correction Table – Street. Finally, it is automatically updated EPIDAT database according this Error Table. Interesting situation is shown in Fig.3. The same wrong street "Bezručova" exists in town HORKA NA MORAVOU as correct form "P. Bezruče" and in UNIČOV as "Bezručovo nám.". Report about every correction was written to last column NOTE for every record by a person who corrected values (Fig.3). Separate notes were written for correction of town, street and district.

STREET	CORRECT_TOWN	CORRECT_DISTRICT	CORRECT_STREET	NOTE
Bezručova	LIPNIK NAD BECVOU	PR	P. Bezruče	street correction
Bezručova	PROSTEJOV	PV	Bezručovo nám.	street correction
Bezručova	UNICOV	OC	Bezručovo nám.	street correction
Bílá Lhota	BILA LHOTA	OC		street correction
Bilsko	CHOLINA	OC		street correction
Bilsko	BILSKO	OC		town correction
Bilsko	CHOLINA	OC		town correction
Biskupice	CHOLINA	OC		town,district correction

Fig.3 Example of Error Table for the street correction with a note about corrections

The number of distinct records in Error Table of the street was 2,675. The total number of repaired records in EPIDAT database according Error Table for the STREET was 8,403. Null value at STREET attribute value has 6,529 records. Not all records were repaired. About 409 records remain uncorrected in STREET attribute. The number of distinct records in Error Table for districts was 191. It was responded to 2,757 repaired records in EPIDAT for permanent address of patients. About 60 records are wrong from the point of permanent address of patient at town and district level. Furthermore, 115 records remained uncorrected at attributes PLACE_OF_INFECTION and DISTRICT_OF_ INFECTION [10].

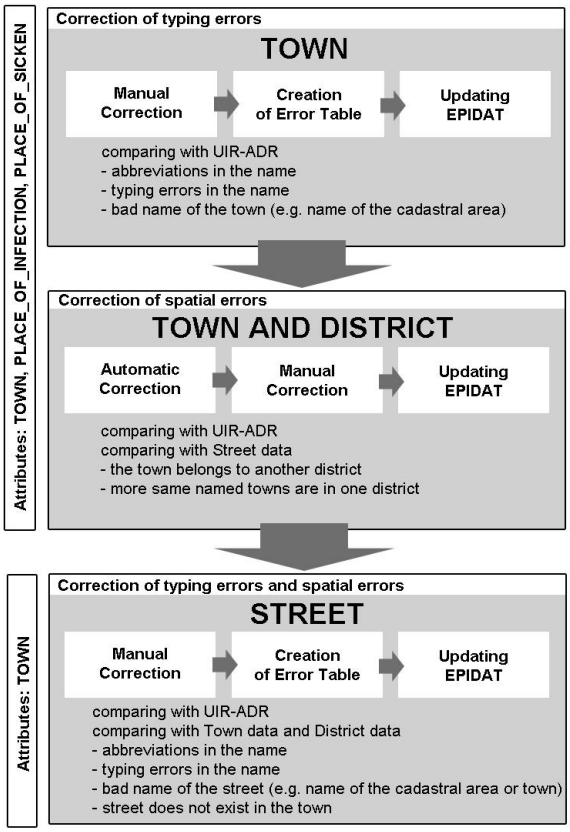

Fig.4 Data flow in the correction process

3.3 EPIDAT database correction wizard

The simple wizard was designed for correction and navigation in the steps of the EPIDAT database correction (Fig.5). All database processing and interface of wizard were made in the database system Microsoft Access 2003. The correct database structure and correct setting of cardinality of relation were made according conceptual model. Conceptual database model is the important stage in database design process [11].

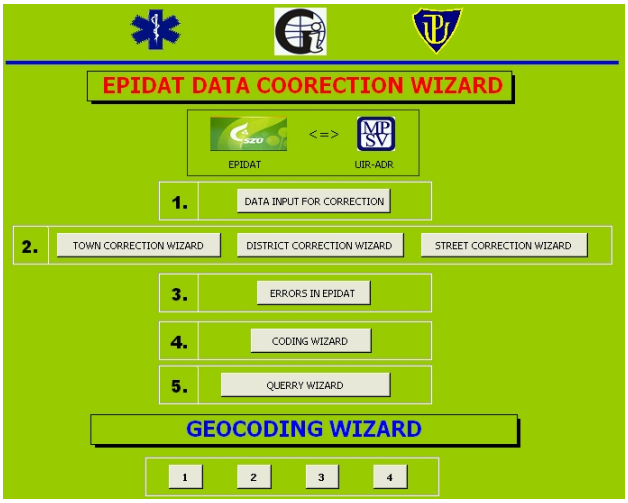

Fig.5 Interface of EPIDAT data correction wizard

All corrections of records were based on SQL queries: SELECT, INSERT and UPDATE [12, 13, 14]. The group of six SELECT queries identified errors in EPIDAT table. These SELECT queries joined the EPIDAT table with the UirAdr Correction Table by LEFT OUTER JOIN. Corrected records match each other. Differences in the left outer joined records were selected by WHERE condition IS NOT NULL and IS NULL.

Example of SELECT query for discovery of the wrong distinct records in Town-Name attribute is:

```
SELECT DISTINCT Epidat.Town
  FROM Epidat LEFT JOIN Town ON
  Epidat.Town = Town.TownName WHERE
  (((Epidat.Town) Is Not Null) AND
  ((Town.TownName) Is Null));
```

The filling of six Error Tables was realized by six INSERT queries. They were based on previous SELECT queries. These SELECT and INSERT queries served for the first automatics identification of error records. In the second step of correction, the manual filling of Error Tables with correct attributes was also based on SELECT query as the source for item list at the second column (Fig.1). Finally, queries UPDATE automatically updated the new values of seven new geocode attributes and also seven NOTE attributes about correction to EPIDAT table from Error Tables. More than 40 queries were used for all stage of correction. Records in Error Tables were possible to delete by DELETE queries.

4 Conclusion

The quality of spatial information in EPIDAT database is limited. The detailed survey of geocode attribute in trial EPIDAT database for the Olomouc region brings results that are shown in Table 1. Total number of processed records was 23,999 for 10 diagnoses. Percentage of correct records is more than 90% for the attributes Town, Place of infection and Place of sicken. The worse situation is in street attribute - 35% wrong records caused by lot of typing errors.

Table 1. Number of correct, empty, wrong and repair records in EPIDAT database

Attribute	Correct records [%]	Empty records [%]	Wrong records [%]	Repair records [%]
Town	92,73	0,00	7,27	7,27
Street	37,78	27,21	35,01	33
Place of infection	91,27	0,49	8,24	7,9
Place of the sicken	91,85	0,03	8,12	7,6

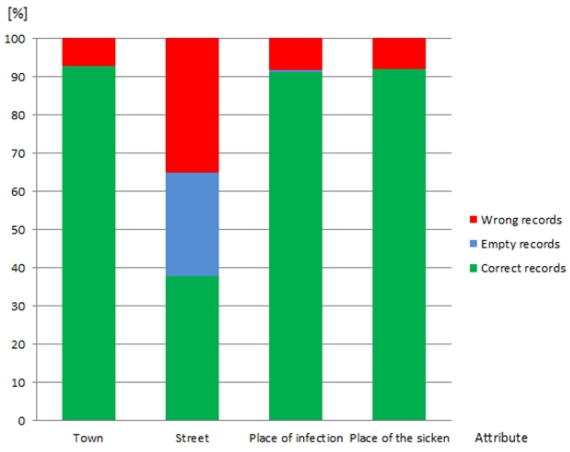

Fig.6 Bar chart of structure of records for four location attribute (Town, Street, Place of infection, Place of the sicken)

The project realized at Department of Geoinformatics found the way how to correct wrong records with national address register UIR-ADR and cadastral register. The correction process is semiautomatic process based on SQL queries that automatically identify errors and finally automatically update EPIDAT database. Manual setting of the correct value by the operator is needful in Error Tables. Strong regional geographic knowledge is also useful especially in combination of more errors in one record. The Error Tables can be used repetitively for new data set of all diagnoses or newer data set. The correction of EPIDAT database for another region or district in the Czech Republic than Olomouc Region can be realized in the same way.

The useful result is also recommendation for improving the new version of EDIDAT program. It will be necessary to adopt Territorial Identification Address Register UIR-ADR into EPIDAT program in a new version. Strong recommendation is to stop the manual filling of values in street attribute. The best way is to fill all geocode attributes from Territorial Identification Address Register UIR-ADR. Comparison with the house address can also increase the quality level of spatial information.

The way how to correct wrong records, namely program wizard, can be used for another data set from EPIDAT. It can be considered data set for all 53 diagnoses for Olomouc region with utilization of the same Error Tables. The repairing of newer data set form 2009 – 2011 can be realized by the same way. The Czech Republic is divided into 14 districts. Every district has own Regional Hygiene Station. Steps of correction can be also used for correction of EPIDAT database at another Regional Hygiene Station. The filling of Error Tables is necessary in case another region than Olomouc region. The final assets of research are two. The first asset is the recommendation for the new version of EPIDAT program. The second asset is the wizard for semiautomatic correction of all old epidemiological data in geocode attributes.

References:

[1] Rushton, G., Armstrong, M.P., Gittler J., Greene B.R., Pavlik C.E., West M.M., Zimmerman, D.L.: Geocoding in Cancer Research: A Review. *American Journal of Preventive Medicine*. Volume 30, Issue 2, Supplement 1, 2006, pp.16-24.

[2] Krieger, N., Waterman, P., Lemieux, K. Zierler, S., Hogan J.W.: On the wrong side of the tracts? Evaluating the accuracy of geocoding in public health research, *Am J Public Health* 91, 2001, pp. 1114–1116.

[3] Zandbergen P.A.: Influence of geocoding quality on environmental exposure assessment of children living near high traffic roads, *BMC Public Health* 7, 37, 2007.

[4] Croner, C.M., Sperling, J., Broome, F.R.: Geographic information systems (GIS): new perspectives in understanding human health and environmental relationships, *Statistics in Medicine* 15, 1996, pp. 1961–1977.

[5] ESRI, Inc. ArcGIS Desktop Help 9.3.

[6] Geokódování a reverzní geokódování (Geocoding and reverse geocoding). *Geobusiness* No.1. Springwinter, Praha, 2009, pp. 38-39.

[7] Cromley, E.K., McLafferty, S.L.: *GIS and public health*, The Guilford Press, New York, 2002.

[8] Ministry of Labour and Social Affairs: UIR-ADR Územně identifikační registr adres, 2010, Available at: http://forms.mpsv.cz/uir/

[9] Czech Office for Surveying, Mapping and Cadastre. (2010) [cit. 2011-02-26]. Available at: <http://www.cuzk.cz>

[10] Havlík, M.: *Průvodce geokódováním zdravotnických dat databáze EPIDAT* (Manual of health data geocoding in EPIDAT database), bachelor thesis, Department of Geoinformatics, Palacký University, Olomouc, 2010.

[11] Dobesova, Z.: Database modelling in Cartography for the "Atlas of Election". *Geodesy and Cartography* 38(1). Vilnius Gediminas Technical University (VGTU) Press Technika, Litva, Taylor & Francis Group, pp. 20-26. ISSN 2029-6991 (print), ISSN 2029-7009 (on line), DOI:10.3846/20296991.2012.

[12] Dobesova, Z.: *Databázové systémy v GIS* (Database systems in GIS). Publishing house of Palacký University, Olomouc, 2004.

[13] Pokorný, J.: *Dotazovací jazyky* (Query languages), Karolinum, Carles University, Prague, 2002.

[14] Šimůnek, M.: SQL, Grada Publishing, Praha, 1999.