OLAP for heterogeneous socio-economic data – the challenge of integration, analysis and crime prevention: a Czech case study

Jiří Horák¹, Igor Ivan¹, Bronislava Horáková²
Jiri.horak@vsb.cz; igor.ivan@vsb.cz; bronislava.horakova@intergraph.com

¹VSB-Technical University of Ostrava, ²Intergraph CS, Czech Republic

Keywords: OLAP, grid, harmonisation, integration, crime

Abstract:
On-line analytical processing (OLAP) represents an approach to provide quick overview of consolidated data (harmonized and aggregated) stored in multidimensional data structures. Unlike traditional applications in business intelligence the paper demonstrates the design and development of OLAP for social environment, especially for crime and human factors in the territory. The multidimensional data structure is proposed and tested to geographically, temporally and thematically combine data from different sources. They include population data (NSI – CZSO), Registry of land identification, addresses and properties (NMCA - Czech Office for Surveying, Mapping and Cadaster), various crime registers (Police of the Czech Republic, city police), register of schools (Min. of Education, Youth and Sports), register of health service providers (Min. of Health), register of unemployed (Labour office), register of lottery game devices (Min. of Finance), and others. Obviously the data sources differ from technical point of view (formats, accesses), legal and ethical aspects (license policy, sensitivity), and maintenance (system of updating). Some authors call such data as a big spatial data – due to its volume, velocity (timeliness) and high variety.

The system enables to harmonize, aggregate, anonymize and integrate data among different topics due to several shared dimensions, especially geographical and temporal dimensions. Thus it is possible to easily convert data into environmental factors or social indicators using the same spatial and temporal units. Even more this enables to express the intensity of crime based not only on population but also based on other relevant targets like the number of dwellings (for burglaries). The examples of such
analysis are provided. The analytical capabilities are demonstrated on the mapping of the relationships between grid aggregated data of burglaries in 2014 and building profiles.

The data aggregation (combined with several other methods) is intended as one of the method to assure privacy protection.

The proposed elementary spatial cell is 100x100m, arbitrarily aggregated. The temporal dimension uses elementary one day unit. Separately the dimension for day-cycle hours is organized. Other dimensions are age, sex, education, classification of crime, classification of facilities, etc. The core of the database represents following five fact tables - Population, Buildings, Crime, Unemployment, Facilities. Each of them is surrounded by their dimension tables; shared dimension tables link fact tables together.

The OLAP principles facilitate to design GUI for end users and provide them easy tools for analysing, exploring and understanding multifactors acting in the territory and to deeply investigate relationships among social and economic indicators. The pilot client solution is currently under development.

Such OLAP systems may be useful for monitoring and early warning of changed social conditions, for complex analysis of factors and impacts of changes, and finally also for strategic crime prevention.

Introduction

The volume of data created in socioeconomic sphere rapidly increases. Information systems of public sector store and manage enormous data from various operational activities in public sector, mainly for evidence and control purposes of different governmental bodies. Such data can be effectively utilised for secondary purposes, different in nature, analytical processing and aims.

Big spatial data is characterised by three main features: Volume beyond the limit of usual geoprocessing, Velocity higher than available by usual processes, and Variety, combining more diverse geodata sources than usual. The popular term denotes a situation when one or more of these key properties reach a state when traditional methods of geodata collection, storing, processing, controlling, analysing, modelling, validating and visualizing fail to provide effective solutions. The most important issue is how to exploit the big spatial data. Trainor (2014) point out the problem there are no currently acceptable processes or procedures for using Big Data to produce Official Statistics.

One of the possibilities how to deal with such big data represents the concept of multidimensional database and OLAP. Both of these parts are usually understand as a part of Business intelligence. Business intelligence (BI) mainly refers to computer-based techniques for identifying, extracting and analysing business data. This data is in functional applications typically at too fine level of granularity for C-level executives. They are therefore abstracted using ETL tools into a data warehouse (Badard et al., 2012).
On-line analytical processing (OLAP) represents an approach to provide an access, in an efficient and intuitive manner, to consolidated data (harmonized and aggregated) stored in multidimensional data structures.

OLAP is the core (computational engine) of business intelligence. All data pre-processing is designed to improve OLAP, and reporting is only as good as the OLAP's computation. It supports efficient data processing such as aggregation, statistical analysis, and computational modelling (Badard et al., 2012).

OLAP can support aggregation of cumulative data (e.g., sums, averages, minimums, maximums) or numeric data (such as sales, number of employees, number of products produced, inventories) presented in a way that allows (Loshin, 2012):

- Drill-down (success in hierarchy down, towards more details),
- Roll-Up (success in hierarchy up, obtaining more aggregated data)
- Drill-Across (link several fact tables with the same granularity)
- Slice-and-Dice (splitting data)
- Pivot (exchange of dimension in designed view)

The spatial extension of OLAP is called SOLAP. A SOLAP tool can be defined as a visual platform built especially to support rapid and easy spatio-temporal analysis and exploration of data following a multidimensional approach comprised of aggregation levels available in cartographic displays as well as in tabular and diagram displays (Bédard Y. 1999, Rivest et al., 2001).

The multidimensional database is usually implemented as a Data Warehouse (DWH). A data warehouse is a subject-oriented, integrated, time-variant and non-volatile collection of data in support of management's decision making process (Bill Inmon in Humphries et al., 2002). “Subject-Oriented” means oriented to a particular subject, according which data is classified in data warehouse. The subject can represent an employee, a product, or a territorial unit. A data warehouse integrates data from multiple data sources. A request for integration includes i.e. unification of terminology or used units. Stored data should be consistent and reliable. Time-variant can be understood as a storage of series of snapshots, where each of them represents some time period.

The multidimensional database (data warehouses) can be based on relational database models with appropriate multidimensional structure (ROLAP) or special multidimensional databases where multidimensional structure is naturally supported (usually multidimensional arrays are applied) and specialised multidimensional DBMS is available (Pokorný 2004).

In ROLAP the **dimensional modelling** represents information in a way that is more suitable for high-performance access. The basic unit of representation is a single multikeyed entry in a slender fact table, with each key exploiting the relational model to refer to the different dimensions associated with those facts (Loshin 2012).

Dimensions represent selected aspects (features) used for organisation and aggregation of data (facts). Analyses of aggregated data are provided according to dimensions. Dimensions usually contain hierarchical structure. Each element of hierarchy can be applied for grouping facts.
Facts represent aggregated values interesting for decision making. Usually they are summarized data – numerical, measurable and repeatedly obtained which generate M:N relationships between dimensions. Two properties are recommended to check (Pokorný 2004):

- Granularity – specifies the level of detail for facts. It depends on the level of detail for dimensions. High granularity means data storage in low level of aggregation, thus with high details.
- Additivity - specifies if it is possible to summarize data according to dimensions.

Unlike traditional applications in business intelligence the paper demonstrates the design and development of DWH and OLAP for social environment, especially for crime and human factors in the territory. The multidimensional data structure is proposed and tested to geographically, temporally and thematically combine data from different sources.

**Design multidimensional database**

We have focused to integrate following data sources:

- population data (NSI – CZSO). Currently, population grid 1 km² is provided by CZSO from results of Census 2011, but it contains only number of inhabitants. Alternatively, we have processed locally available data from municipal information systems which give us the opportunity to obtain fresh data (i.e. each month) and to generate detail sex and age distribution in each cell. The example was given in Horák (2012) (fig. 1). Of course, spectrum of available information about inhabitants is quite short and cannot be compared with richness of census data. Unfortunately, the new legislation limits the usage of such municipal sources. The central population register exists since 2012, but the access to this live system is strongly restricted and no anonymised aggregated outputs from this system are available yet.

![Fig. 1 Age homogeneous (left) and variable (right) locality (Ostrava, 9/2009) (Horák, 2012)](image-url)
Registry of land identification, addresses and properties (NMCA - Czech Office for Surveying, Mapping and Cadaster). RLIAP is the only public available “basic register” in our country. It includes the evidence of buildings of different types classified according to usage, construction types, number of floors, heating types etc.

crime register (Police of the Czech Republic). Police of the Czech Republic (PCR) uses tens of various information systems where centralisation and integration processes are still in progress. The statistics of crime (using ESKK system) are provided according to the organisation scheme of PCR, starting from individual police unit which usually mean one unit per municipality or several units in cities. There are two main problems – delimitation of the police units do not correspond to administrative units, and the statistics summarise all events reported in its evidence not taking into account the place of crime (register is based on announcement of crime), thus it contains also events outside the territory of the given police unit. The new processing uses data exports from centralised information system for event recording (ETR = Evidence of Criminal Proceedings) which contains detail information about each crime commission.

offence register (city police). The city policies use different information systems for registration of offences (and crime). The unification is still ad hoc, driven mainly by expansion of individual software solutions. There are big differences mainly in data harmonisation level which means the processing requires a strong focus on individual data structure. The central evidence is currently limited to traffic offences and several other types (environmental offences etc.). The new central register of offences (with limited evidence) will come into force since 2017.

register of schools (Min. of Education, Youth and Sports) offers contact information and classification of education level and type.

register of health service providers (Min. of Health) provides contact information and the capacity (number of beds)

register of unemployed (Labour office). The total number of registered unemployed should be accompanied by other valuable aggregates like low educated or long-term unemployed.

register of gambling devices (Min. of Finance)

Register of companies (CZSO, or others)

The data sources differ from technical point of view (formats, accesses), legal and ethical aspects (license policy, sensitivity), and maintenance (system of updating). Obviously, the crime register is protected by the most restricted access, followed by the offence registers and register of unemployed. It is assumed to harmonise and aggregate such data by internal processes of its provider. Other registers are public available and not restricted in access and utilization. In such cases harmonisation and aggregation processing would be provided by the integrator of the OLAP system. The updating procedures are variable, but it is recommended to extract majority of source once a month.

There are more issues in data extraction and integration (Horák, 2012) – selection date, consistency, confidentiality and scantiness. Registers are „live“ systems. The continuous process of data editing is still running and it includes not only adding new records, but also modification, deleting of some older records. The data selection of one date (i.e. data relevant to the end of the year) may differ when we repeat the selection after some time. Also consistency of data integrated from different sources may
fail. Differences in time, semantics, and methodology lead to inconsistencies when joining data from more registers (i.e. we can find addresses with number of unemployed persons higher than number of residents).

The data entering the multidimensional database has to be harmonised. It is the objective of ETL processes. The ETL pre-processing is focused on following issues:

1) control procedures (integrity constrains, check validity of time range, geographical range, validity of codes in the database, i.e. check referential integrity etc.

2) harmonisation – smart procedures to calculate referential time of event from time interval (the procedure has to take into account type of event, length, beginning and ending of the interval), harmonisation of addresses (harmonised name of municipality, municipal part, street, numbers etc.), classification of facilities, buildings etc. Harmonisation is usually accompanied by geocoding (Longley et al., 2005) i.e. transformation of textual description of placement into coordinates.

The quality of harmonisation is crucial for the quality of generated secondary data. Especially quality of geocoding has to be assured on the satisfactory level. The minimal limit of the success rate is 85 % which should not cause systematic spatial error (Ratcliffe, 2004 in Andersen, Malleson 2013). In the case of imprecise or vague description of location which causes failure in identification the smallest administrative/geometrical units, the system enables to store identification of higher hierarchical units fitted to fully cover the possible geographical range of event/object occurrence within the vague specification.

The multidimensional database contains aggregated and not isolated data. Thus the next step in data processing is aggregation.

The aim of aggregation is to suppress the influence of random effects for data with high resolution and improved identification of trends and relationships (i.e. dependencies, associations) (Harries, 1999, Chainey, Ratcliffe 2013). The aggregation is also an important tool for data anonymization.

Data anonymization can be provided by combination of following procedures:

- Data aggregation – namely territorial aggregation
- data filtering (i.e. threshold 25 persons)
- data scramble (Batista et al. 2011) or jittering
- random adding or subtracting small value (Kraak, Ormelling, 1996)
- data rounding (Quinn 1996)
- data projection (to restricted list of attributes)

The selection of appropriate procedures and their combination has to take into account namely classification of event, its frequency and the type of relevant territory.

The following recommendations for multidimensional database construction were given:

- Data warehouse should be separated from operational systems. Transformations from operational into warehousing systems are supported by ETL processes.
• Implementation can be both in relational data structure (relational OLAP) as well as in specialised database SW solutions
• Data are stored with high granularity, thus with low level of aggregation. It is due to the requirements for local analysis and interpretations.
• Additive facts are preferred. Nevertheless several types of facts like number of facilities are stored as a total number to the given date because they are used in analysis in this form, the temporal changes are slow and thus not interested.
• Implicit hierarchy in dimensional tables is preferred.

The core of the database represents following five fact tables - Population, Buildings, Crime, Unemployment, Facilities. The set is arbitrary extensible. Each of fact tables is surrounded by their dimension tables.

Fig. 2 Part of the multidimensional database scheme (fact tables with yellow labels, other shared dimension tables)
Two dimensions are shared among all fact tables – geographical and temporal dimensions. They enable to link fact tables together. The location is considered as a corner stone of the database. The common grid unit of 1km² is used as reference unit for various data also in many application in European Forum for Geostatistics (i.e. Poelman (2014) for harmonised indicators on urban public transport, Martin (2013) for adjustment of population). The location is expressed in each fact table by two geographical dimensions – the administrative dimension and geometrical dimension (square grid). The administrative dimension identifies the administrative units. The basic unit is the part of municipality which seems to be appropriate compromise between data availability, privacy protection and the required highest detail. The smaller units (i.e. basic territorial units, statistical units) are considered too small and limited in available various socio-economic data sources. The larger unit (municipality) is not sufficient for all cities. The elementary unit for geometrical dimension is 100x100m cell, arbitrarily aggregated. It is consistent with 4th level of the scale system for public authority where 10 km window and 100m grids for communes and urban districts are recommended (Bacler, 2014).

Each dimension enables to provide different analysis and easy approach to integrate data both referenced by administrative units as well as by coordinates. Of course, these two dimensions are not independent and cannot be both concurrently included in the data aggregation.

The temporal dimension uses elementary one day unit. Separately the dimension for day-cycle hours is organized. The reason for separation is the joined evidence of day and hour would generate enormous number of records in fact tables, practically eliminate aggregation (most records would contain only one event) and limit possibilities of easy solution for aggregation in case of specific temporal interval selection like “morning time”, “rush hours”, “late evening” etc.

Of course, such fine temporal resolution seems to be reasonless for common statistical surveys. Nevertheless we use (or would like to use) this shared dimension also for continuously monitored events like crime, offences, unemployed or i.e. even selected health cases.

Other dimensions are age, sex, education, classification of crime, classification of facilities, etc.

The dimensional table for age is organised into 5-years basic categories, but it includes also more wide categories (10-years, 20-years) and even half-opened intervals like “30 and more”. It enables to record our vague knowledge i.e. about offenders. Of course, each person has only one age class (the most accurate) not to summarize him/her more times.

The classification of crime (including offences) use the standard 3-level classification system commonly applied in Czech police information systems. Each event may have many classes of crime which facilitate appropriate selection of all type of cases even though the final count of crime events can exceed those published in official statistics.

The classification of facilities is focused on the purpose of the facility and the hierarchical structure enables to aggregate according to the detail (i.e. night club) or broader (i.e. entertainment) classification.
The type of location of crime events are described using “subobject” dimension. It is intended to recognise some type of internal/external spaces, construction or arrangements which may be related to the type or intensity of crime. Good examples are corridor, subway, cloakroom, underground parking etc.

The fact table about crimes

The geographical (administrative and geometrical) dimensions are tripled due to the need to organise (aggregate) data according to place of crime, to residence of offender and to residence of victim. In case of multiplicity of places of crime, offenders or victims, the geographical identification has to fulfil all places which means that i.e. in case of three offenders with residences inside the city but in different parts of city, the identification of appropriate municipality is provided but not any municipal part. Similar solutions are applied for other characteristics of offenders and victims like age, sex, habits. In case of different 5-years age categories, wider age category to union all individual basic age categories is to be applied (and identification of basic 5-years units are excluded). Habits are currently focused only on drugs and alcohol.

Facts about crimes are organised also according to the organisational unit of police, and type of facilities.

The fact table about population contains basic features like sex, age, education, nationality, family status, and economic activity, and the corresponding number of population. Such data is available from census (if it is aggregated to the grid). These features have been selected to provide appropriate denominator to intensity calculation (incl. sex and age specific intensities) (index of crime, rate of unemployment etc.) and to distinguish anomalies in socio-economic conditions (hot spots) which may initiate various socially-pathogen phenomena.

The fact table about facilities contains the number of facilities in each geographical and temporal unit according to the type of facility. In case of multiple dedications (purposes) of one physical building, more classes are recorded – this is one of the main differences between facts about facilities and facts about buildings. The data sources are various – register of schools (Min. of Education, Youth and Sports), register of health service providers (Min. of Health), register of gambling devices (Min. of Finance), database of time schedules for transport stops (CHAPS Ltd. as a national administrator), register of companies etc. The aggregation usually provides number of objects but if available also information about capacities are interested.

The fact table about health is intended to provide aggregated anonymised information about the selected health issues in the territory (and time). Especially those indicating local occurrence of socially-pathogen phenomena like number of drug users (incidence, prevalence), number of distributed injection sets in help centres, number of collected used injections (or grouting points), number of cases of selected infection diagnoses (hepatitis B or C, syphilis, scabies etc.), number of hospitalized and number of death for selected diagnoses and circumstances.
The fact table about registered unemployed enables to monitor the current local situation in unemployment and structural unemployment. The number of registered unemployed is distributed according to education, duration of current unemployment, required job, sex, and age. The combination with the population fact table provides the unemployment rate and group specific unemployment rate. The combination with other fact tables offers great opportunities to investigate spatial and temporal relationships to facilities like lottery game devices, pawnshops, or social centres, charity, social advisors; to crime and offences; or to health issues like drug distribution; or to dwelling characteristics (buildings). The recommendation for basic temporal step is 1 month.

The fact table about buildings provide information about the number of buildings according to type of building, age category, inhabitation, ownership, and usage. Complete set of such data is available from census. Nevertheless there are different data sources providing not complete but current data. The “Registry of land identification, addresses and properties” contains the list of buildings (and constructions) with several classifications where the most interesting is the “type of usage” and number of floors. It enables to distinguish family residential buildings, block of flats, recreational buildings, industrial, commercial, civil amenities buildings, etc.

The Czech Statistical Office builds the Register of Statistical Units and Buildings (RSO) which contains current information about address points and buildings. Such data can partly substitute the old data from Census.

The case study

The analytical capabilities are demonstrated for selected data using 1km grids. We focus to demonstrate Drill-Across analysis of OLAP. 1km grids (level 3.) is recommended for national regions by Bacler (2014). 1km grids provide better statistical results of cross-facts relationships than in the case of 100m grids. Data about crime is limited to Jan-Nov 2014.

Usually used index of burglaries can be compared with another relevant index such as density of burglaries mapping using number of flats (FTAB_buildings, CZSO RSO 2014) as a denominator (Fig.3 ) Data is filtered to grid cells containing more than 10 dwellings (flats).
The relationships between gridded data of burglaries to dwellings and number of residential buildings are better to explore at local level (in a city) than in a smaller scale because significant differences can be found. The comparison of 3 towns (CB Ceske Budejovice, KO Kolin, and OV Ostrava) shows very high density of buildings in part of CB (compare to OV and KO), and namely different dependencies. Two types of dependencies can be distinguished in KO – no burglaries in some parts regardless building density, and almost linear relationship between these variables showing expected “random” occurrence of burglaries. Approximately the same line can be seen in data of CB.
Fig. 4 Relationships between gridded data of burglaries to dwellings (VloupBydl) and number of residential buildings (BUDBYD) in 2014 for 3 towns (CB Ceske Budejovice, KO Kolin, and OV Ostrava)

The map of gambling machines density shows evenly distributed pattern in majority of the Czech republic except of several areas like the SW part (fig. 5). The number of gambling clubs per 100 inhabitants (fig. 6) highlights only several outlier locations with high concentrations situated close to the border with Austria and Germany and in the middle close to the main highway.
Fig. 5 Number of gambling machines per 1km²

Fig. 6 Number of gambling clubs per 100 inhabitants
Conclusion

The integration of socio-economic data from different sources can be effectively provided using multidimensional modelling and the OLAP concept. Harmonisation of data, aggregation to geographical administrative and geometrical units, and simultaneously to temporal and thematic units enable to release useful secondary information content from operational data generated in large volumes in the public sector. Different registers and evidences may provide valuable data and new snapshots of the current, especially socio-economic, conditions. Majority of these data sources are well-known; the challenge is in the new ways of processing, requirements for velocity and volumes which are true signs of Big Data.

The design of multidimensional database for heterogeneous socio-economic data has been introduced. Five main fact tables store information about Population, Buildings, Crime, Unemployment, and Facilities. The structure can be easily extended. The shared geographical and temporal dimensions enable to fast combine data from various thematic domains using the same referential spatio-temporal units. The proposed elementary unit has 100m x 100m and 1 day. The system enables to flexible select and aggregate information across different domain and discovers new possibilities to explore the socio-economic situation and support decision making in prevention strategies. It is documented by several examples of analysis.

Acknowledgements

Data is provided by the courtesy of the Czech Statistical Office, Police of the Czech Republic, Czech Office for Surveying, Mapping and Cadaster, Czech Ministry of Finance, Labour offices, Czech Ministry of Health and Municipal Police departments in Ostrava, Kolín and České Budějovice. The research is supported by the research of the Czech Ministry of Interior, project “Geoinformatics as a tool to support integrated activities of safety and emergency units”, No. MV-32046-58/VZ-2012.

References


