# OLAP A PRIESTOROVÉ DÁTA

Pavel BELAJ[1]

[1] Institut geoinformatiky, Hornicko-geologická fakulta, VŠB-TU Ostrava,
17.listopadu 15/2172, 708 33, Ostrava Poruba, Česká republika
[1]*pavel.belaj.st@vsb.cz*

**Abstrakt**

Pojem Business intelligence(BI) je v dnešnej dobe pomerne často frekventovaný. Najčastejšie sa využíva pre popis sady nástrojov, techník, a procedúr, ktoré umožňujú rýchlo a pružne získavať informácie potrebné pre strategické a dennodenné riadenie a rozhodovanie. Architektúra BI môže byť v rôznych implementáciách rôzna. BI sa najčastejšie skladá s transformačných nástrojov (ETL), dátových skladov (DWH), nástrojov (OLAP), nástrojov pre reportovanie a dolovanie dát.

S rozvojom priestorových databáz sa čoraz častejšie začínajú objavovať BI riešenia, ktoré dokážu využívať priestorové dáta napríklad v podobe reportov. Problémom priestorových dát je ich značná zložitosť, preto sú pre mnohé časti BI použiteľné iba v obmedzenej miere, alebo ich využitie nie je vôbec podporované.

Využívanie priestorových dát v prostredí BI je vcelku problematické. Pri práci s BI sa často využívajú pokročilé nástroje, ako napríklad Business Intelligence Development Studio, ktoré však nedokážu využiť potenciál priestorových dát uložených v priestorových databázach naplno a efektívne. Jedným z hlavných problémov prečo BI aplikácie nedokážu plnohodnotne využiť priestorové dáta, je aj neexistencia konceptu, ktorý by dostatočne integroval GIS funkcionalitu do BI sféry. Baltzer (2011) hovorí o možnosti integrovania súčastí BI a geografických informačných systémov do jednotnej sady nástrojov.

Bédard et. al. (2006) ako prvý navrhol svoju koncepciu ,ktorá integrovala jednotlivé komponenty BI a priestorových dát, pre ktoré sa ujala skratka SBI (Spatial business intelligence).

V mojom príspevku sa venujem hlavne možnostiam využívania OLAP v prostredí priestorových databáz. V príspevku sa snažím poukázať na problémy ktoré, súvisia s používaním priestorových dát, pre OLAP analýzy.

**Kľúčové slová: Business intelligence, OLAP, SOLAP, MDX, GIS**

**Abstract**

The term Business Intelligence (BI) is nowadays relatively frequently used. It is most commonly used to describe the set of tools, techniques and procedures, which allow quick and flexible obtaining of information necessary for strategic management control and decision making. BI architecture may vary with different implementations. BI is most often composed of transformation tools (ETL), data warehouse (DWH) tools (OLAP) tools for reporting and data mining tools.

Development of spatial databases enabled increasingly emerging BI solutions, which use spatial data for example in form of reports. The problem of spatial data is the large complexity. Thus many BI components have limited application, or their usage is not supported.

The usage of spatial data in the BI environment is quite problematic. Working with BI requires advanced tools such as Business Intelligence Development Studio, but they cannot completely and effectively utilize the potential of spatial data stored in spatial databases for advanced spatial analysis. One of the main problems why BI applications cannot fully exploit the spatial data is the absence of a concept that would adequately integrate GIS functionality into the BI sphere. Baltzer (2011) mentioned the possibility of integration of BI components and geographic information systems into one single toolkit.

Bédard et. al. (2006) as the first designed a concept, which integrated all BI components and spatial data. It has been called Spatial Business Intelligence (SBI).

The aim of this contribution is to depict the possibilities of OLAP application in the environment of spatial databases. In this paper I tried to highlight the problems that have arisen from the usage of spatial data for OLAP analysis.

**Keywords: Business intelligence, OLAP, SOLAP , MDX, GIS**

## 1 ÚVOD

OLAP môžeme chápať ako koncepciu pre multidimenzionálne spracovanie dát získaných hlavne z dátových skladov alebo Executive information system (EIS), ktorý umožňuje interpretovať dáta z rôznych hľadísk a s rôznymi stupňami podrobnosti (Silva, et. al., 2010).

OLAP podporuje prirodzený iteratívny analytický proces, pretože dovoľuje užívateľovi pohľad na dáta cez rôzne dimenzie, ktoré môžu mať rôznu úroveň detailu. To umožňuje rôzne kombinácie pohľadu na dáta, čo uľahčuje vznik nových hypotéz a znalostí (Glymour et al. 1997).

OLAP technológia pracuje s multidimenzionálnymi dátami, ktoré tvoria dátovú OLAP kocku. Multidimenzionálny model znamená, že dátová kocka môže mať viacero dimenzií, veľmi často sú to dimenzie produktu, času alebo geografického regiónu. Dimenzie sú tvorené atribútmi, ktoré sú logicky rozdelené do hierarchií. Napríklad dimenzia času je často rozdelená na: rok, kvartál, mesiac, týždeň a deň. V kocke sa tiež nachádzajú fakty, ktoré predstavujú kvantitatívne hodnoty, ktoré majú byť v databáze analyzované. Sú to atribúty, ktoré sú hlavným dôvodom evidencie z pohľadu množstva, ceny alebo inej zmysluplnej kvantitatívnej hodnoty.

## 2 OLAP

Rôzny pohľad na dáta umožňujú základné analytické operácie OLAPu (Dohnal, Pour, 1997) ako napríklad:

- **Drill-Down**: postupovanie v hierarchií smerom na dol a tak získanie menšej miery agregácie teda vyššieho detailu.
- **Roll-Up**: je opakom Dril-Down, kedy sa v hierarchií pohybujeme smerom hore.
- **Drill-Across**: je spojenie niekoľkých faktových tabuliek s rovnakou granularitou.
- **Slice-and-Dice**: je rez multidimenziálnou kockou a obmedzenie výberu dát.
- **Pivot**: je menenie uhla pohľadu na dátovú kocku na prezentačnej úrovni.

Najväčšia výhoda OLAPu je v tom, že naše požiadavky dokáže vykonať s minimálnou odozvou a práve dopredu agregované štruktúry sú základnou technikou ktorá túto funkcionalitu umožňuje (Gupta et. al.,1995). Hlavný problém je v podmienke zvanej „summarizability“, ktorá hovorí o schopnosti využiť hierarchicky nižšie dátové agregáty pre výpočet vyššie hierarchických agregátov.

Tieto operácie sú hlavným dôvodom pre vznik OLAPu, pretože štandardnými databázovými dotazmi je táto funkcionalita OLAPu dosiahnuteľná iba veľmi ťažko.

Treba si tiež uvedomiť, že OLAP spracúva u DWH štandardne dáta o veľkosti niekoľkých terabytoch, a preto potrebuje špecifickú štruktúru, do ktorej ukladá pred agregované dáta pre potreby špecifických OLAP operácií. Túto špecifickú štruktúru si môžeme predstaviť ako multidimenzionálnu dátovú kocku, nad ktorou vykonávame štandardné OLAP operácie.

Agregácia je funkcia, ktorá sumarizuje vlastnosti dátového súboru cez nejakú časť dimenzie, ako napríklad čas alebo administratívne usporiadanie. Jej výsledkom sú odvodené dáta, ktoré sú zosumarizované. Základné agregačné funkcie sú definované už v základných štandardoch jazyka SQL, napríklad SUM, COUNT, MAX, AVG atď.

Priestorové agregácie je možné vykonať aj v štandardných GIS aplikáciách, problém však nastáva pri väčších objemoch dát, kedy GIS aplikácie dokážu agregovať dáta iba zo základnej bázy dát, teda nedokážu

efektívne využívať dopredu agregované štruktúry. Tým sa výrazne limituje doba odozvy a možnosť nazerať na dáta viacerými pohľadmi (Pedersen, et, al., 2001).

V prípade OLAPu sa môžeme stretnúť taktiež s termínom „plná predagregácia“, ktorá hovorí o totálnej agregácií všetkých položiek v dátovom súbore. Skutočná „plná predagregácia“ je však nezmyselná, pretože výsledné predagregované dáta budú 200 až 500 krát väčšie ako primárne dáta a ich informačná hodnota bude nevyužiteľná (Shukla, et. al., 1996). Preto je vždy úlohou analytika vyberať vhodné dáta pre agregácie.

Preto je potrebné vytvoriť podmienky, v ktorých je možné dopredu pripraviť agregované štruktúry s ktorých sa neskôr na dotaz analytika, vypočítavajú agregované výsledky. Ak bude OLAP využívať priestorové dáta, tak je skoro nemožné aby dokázal spočítal všetky možné agregácie pre priestorové objekty, pretože výsledná veľkosť pred agregovanej štruktúry by bola extrémne veľká. Preto Bédard a kol. navrhli spôsoby, ktoré sa snažia zvoliť optimálny pomer medzi nízkou odozvou, informačnou hodnotou a veľkosťou vygenerovaných pred agregovaných dát.

### 2.1 MDX

MDX je jazyk, ktorým sa vytvárajú dotazy nad multidimenzionálnou dátovou kockou. Môžeme povedať, že všetky vyššie uvedené operácie OLAPu sa dajú vykonať aj pomocou MDX.

MDX je svojou syntaxou podobný SQL, ale slúži výhradne na prácu s viac dimenziálnou dátovou štruktúrou a jeho výsledok vráti vždy iba nejaký viac rozmerný objekt. Často je aplikovaný nad nejakou OLAP technológiou.

```
SELECT
  {[Measures].[Population]} on columns,
  Filter(
    {[Unite geographique].[Region economique].members},
      ST_Distance([Unitegeographique].CurrentMember.Properties("geom"),
      [Unite geographique].[Province].[Ontario].Properties("geom")) < 2.0
  ) on rows
FROM [Recensements]
WHERE [Temps].[Rencensement 2001 (2001-2003)].[2001]
```

**Obr. 1.** Ukážka MDX s priestorovým rozšírením. zdroj(http://tinyurl.com/6ghpj68)

Na **obr. 1**. je ukážka MDX dotazu, ktorý dokáže využiť pre svoj výpočet priestorovú podmienku vzdialenosti.

## 3 PRIESTOROVÝ OLAP

### 3.1 EXISTUJÚCE RIEŠENIA

Existuje viacero riešení, ktoré spájajú výhody GIS aplikácií a OLAPu, ja sa zmienim iba o niekoľkých z nich.

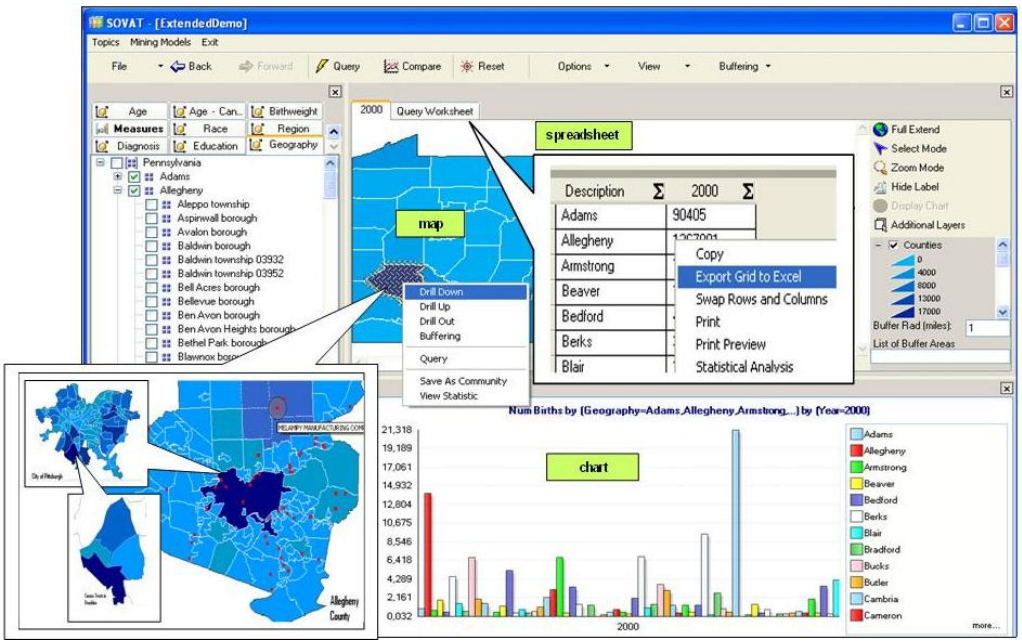
OLAP for ArcGIS je extenzia do ArcGIS Desktop, ktorá sa dokáže napojiť na analistické služby (OLAP) komerčných databázových riešení. Keďže sa pripája priamo na OLAP, dokáže pracovať iba so spracovanou dátovou kockou.

Celá architektúra využitia OLAP for ArcGIS je ukážkou relatívne jednoduchého prepojenia výsledkov multidimenzionálnych OLAP analýz s GIS aplikáciou, ktorá umožňuje pokročilé reportovanie výsledných analýz. Extenzia na úrovni ArcGIS umožňuje OLAP funkcionalitu v mapovom prostredí. Agregované dáta sú lokalizované na mape, najčastejšie cez atribúty PSČ, Okres, Kraj a pod.

Podobnú funkcionalitu poskytuje aj program PostGeoOLAP, ktorý sa pripája na PostgreSQL + PostGIS. Priestorovú zložku dát môže využiť ako obmedzujúce pravidlo, napríklad: „vyber do analýz iba dáta, ktoré sú vzdialené od záujmového bodu do určitej vzdialenosti“.

Komplexným nástrojom sa postupne stáva GeoMondrian, ktorý je odvodený od systému Pentaho, má podporu priestorových dátových typov. Zatiaľ dokáže pracovať iba s PostGIS, pričom v budúcnosti by mali byť podporované aj iné SRBD. Poskytuje priestorové rozšírenie jazyka MDX.

Ďalší z radov experimentálnych nástrojov je SOVAT, kombinuje geografický informačný systém spolu s multidimenzionálnymi štruktúrami. Bol navrhnutý ako systém pre výskum časovo priestorových dát zo zdravotníckej oblasti.



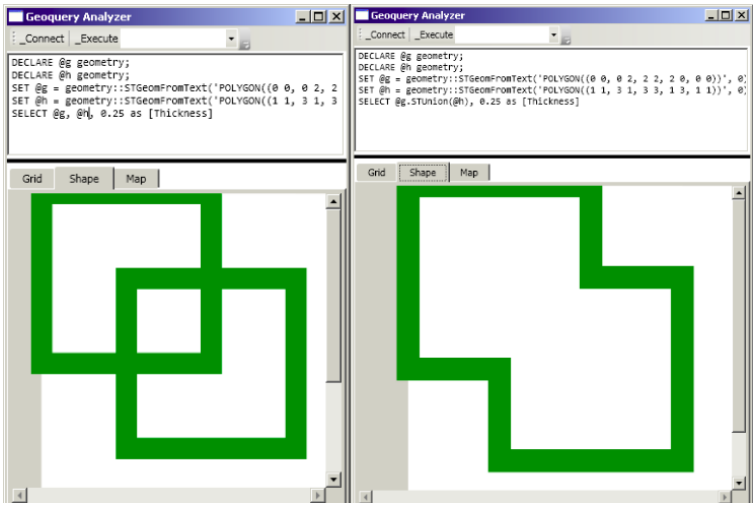

**Obr. 2.** SOVAT zdroj (http://tinyurl.com/cmqmhzj)

Architektúra SOVATu počíta s integráciou socioekonomických dát spolu s dátami o zdravotnom stave obyvateľstva a ich transformácie do multidimenzionálnych štruktúr. Pre vlastné analýzy dokáže využiť integrované dátové dolovanie dát (Scotch, 2005).

## 4 PROBLÉMY APLIKÁCIE PRIESTOROVÉHO OLAPU

### 4.1 PRIESTOROVÉ DÁTA

Spoločným problémom viacerých priestorových OLAP systémov, je okrem iného aj nedostatočné využívanie priestorových dát. Do výpočtu agregovaných hodnôt by bolo napríklad vhodné zaradiť aj konkrétne priestorové objekty, ktoré by boli agregované podľa určitých priestorových funkcií.




**Obr. 3.** V ľavo objekty pred agregáciou, v pravo objekty po agregácií. zdroj (http://tinyurl.com/5rbx23h)

Na **obr. č. 3.** agregujem priestorové objekty na základe funkcie „priestorové zjednotenie“.

Ďalším významným problémom priestorových dát v OLAP systémoch je využívanie vzťahov medzi priestorovými objektmi v procese OLAP priestorových analýz, teda využívanie priestorových a nepriestorových atribútov.

Vzrastajúci objem dát, kde sú zaznamenané priestorové objekty v pohybe, ako napríklad sledovanie mobilných zariadení, si vynútili vznik úplne nového prístupu k OLAP analýze. V novom prístupe potrebujeme definovať nové algoritmy, ktoré dokážu aplikovať koncepty OLAP na takýto druh dát.

### 4.2 PODPORA PRIESTOROVÉHO OLAPU

Významným zlepšením priestorového OLAPu, by bola natívna podpora priestorových dátových typov SRBD pri priestorových OLAP analýzach. Táto pokročilá požiadavka súvisí priamo s potrebou definovať určitý pracovný rámec, ktorý by zapuzdroval jednotlivé analytické požiadavky, napríklad v podobe konkrétnych dotazov na priestorový OLAP. Jednotlivé testovacie dotazy na priestorový OLAP môžu vypadať napríklad takto:

- Vyhľadaj všetky miesta v ČR, ktoré splňujú podmienku že v okruhu 30 km je počet voľných pracovných miest v kategórií vysokoškolák väčší ako 50 a súčasne miera znečistenia ovzdušia je menšia ako 10.
- Vyhodnoť podľa polohy a ponúkaného platu, oblasti v ktorých potenciálny zamestnanec zarobí najviac peňazí. Výsledok chcem zoradiť od najlepšej oblasti po najhoršiu.
- Vyhľadaj aké sú počty voľných pracovných miest vzdialené DO 50 km od Prahy a z agregované podľa vzdialenostného kroku 10 km + z agregované podľa minimálneho vzdelania, alebo oboru alebo aj času vzniku voľného pracovného miesta.

Podstatný problém predstavuje optimalizácia dotazov z hľadiska výpočtovej náročnosti výpočtu dotazov a zložitosti zápisu dotazov.

Iba po dôkladnej analýze celého radu analytických dotazov je možné pristúpiť k tvorbe pracovného rámca pre priestorový OLAP.

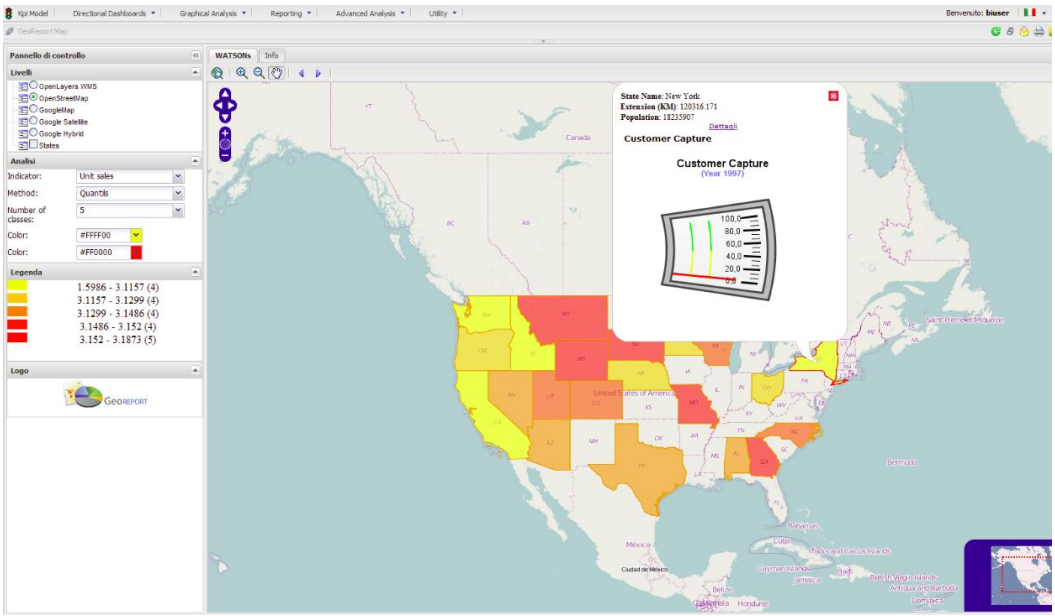
## 5 VIZUALIZÁCIA

Vizualizácia OLAP analýz je často v podobe kontingenčných tabuliek, ktoré sú v mnohých prípadoch aktívne. To znamená že užívateľ môže pomocou klikania a pohybu myši veľmi ľahko meniť pohľad na výsledné informácie.

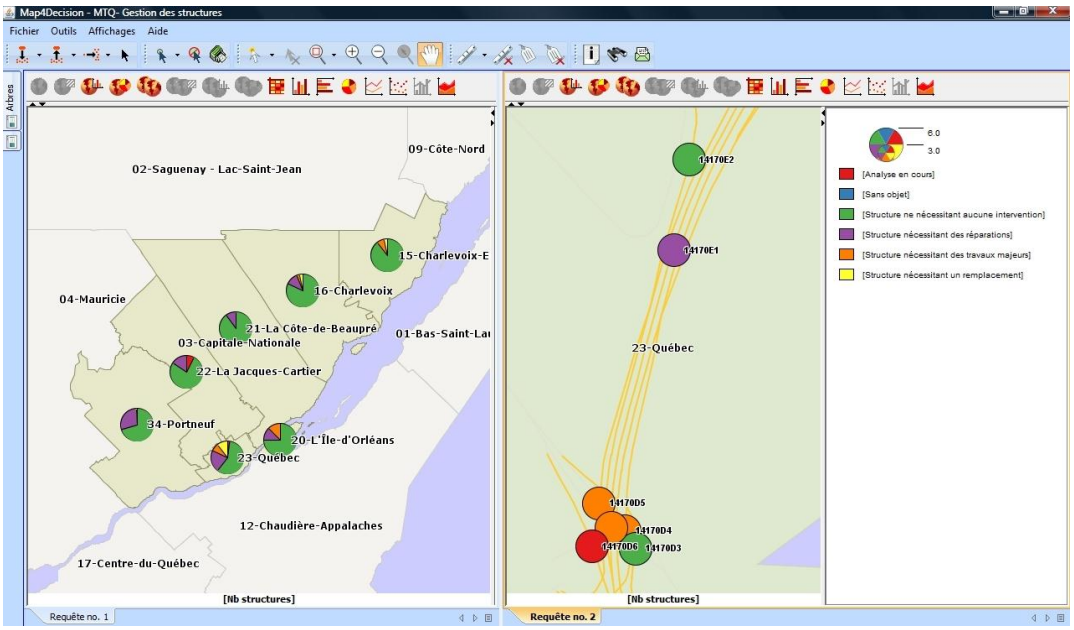
| | | Yr | | | |
|---|---|---|---|---|---|
| | | 2005 | 2006 | 2007 | ALL |
| EmpId | 1 | GROUP BY (EmpId, Yr) | | | GROUP BY (EmpId) |
| | 2 | | | | |
| | 3 | | | | |
| | ALL | GROUP BY (Yr) | | | GROUP BY () |

**Obr. 4.** Tvar kontingenčnej tabuľky

V prípade priestorového SOLAPu sa k vizualizácií využíva aktívna mapa, na ktorej užívateľ môže využívať OLAP analytické operácie, napríklad mapový drill-down.




**Obr. 5.** Aktívna mapová aplikácia. zdroj (http://preview.tinyurl.com/d6d5aj6)




**Obr. 6.** Ukážka Map4Decision. zdroj (http://tinyurl.com/c5e7z4c)

Pre vizualizáciu agregovaných hodnôt sa často využíva kartogram a kartodiagram spolu s ďalšími informačnými prvkami, ako napríklad s rôznymi typmi grafov, ktoré sú často prepojené s mapovým výstupom.

## ZÁVER

Priestorové rozšírenie jednotlivých súčastí BI je relatívne nová oblasť pre skúmanie. Väčšina projektov, ktoré sa zaoberajú touto problematikou, rieši možnosti priestorového rozšírenia na úrovni nových softvérových riešení. To znamená, že väčšinou sa jedná o experimentálny softvér, ktorý je niekedy odvodený z komerčného predchodcu. Problémy, ktoré prinášajú priestorové dáta do oblasti BI, sú príliš zložité a stále sa hľadá vhodné využitie a implementácia v praktickom využití.

**LITERATÚRA**

Baltzer O. 2011. *COMPUTATIONAL METHODS FOR SPATIAL OLAP.* Halifax : Dalhousie University, 2011.

Bédard Y., M. Proulx, S. Rivest, T. Badard. *Merging Hypermedia GIS with Spatial On-Line Analytical Processing: Towards Hypermedia SOLAP.* Berlin : Springer Berlin Heidelberg, 2006. 978-3-540-34238-0.

Joel da Silva, Anjolina G. de Oliveira, Robson N. Fidalgo, Ana Carolina Salgado, Valeria C. 2010. *Modelling and querying geographical data warehouses.* s.l. : Twenty-second Brazilian Symposium on Databases, 2010.

Glymour, C., D. Madigan, D. Pregibon and P. Smyth. Statistical themes and lesson for data mining. Data mining and knowledge discovery, 1997, 1:pp. 11 -28.

DOHNAL, J., POUR,J. Řízení podniku a IS/IT v informační společnosti. Praha: VŠE, 1999. ISBN 80-7079-023-7.

Gupta, V. Harinarayan, and D. Quass. Aggregate Query Processing in DataWarehousing Environments. s.l. : In Proceedings VLDB , 1995.

Pedersen T. B., Tryfona N. Pre-aggregation in Spatial DataWarehouses. s.l. : Department of Computer Science, Aalborg University, Denmark, 2001.

Shukla, P. M. Deshpande, J. F. Naughton, K. Ramasamy. 15. .Storage Estimation for Multidimensional Aggregates in the Presence of Hierarchies. s.l. : In Proceedings of VLDB , 1996.

M. Scotch and B. Parmanto. Development of SOVAT: A numerical-spatial decision support system for community health assessment research. International Journal of Medical Informatics, 75(10–11):771–784, 2005.

Bimonte, S., Miquel, M.: When Spatial Analysis Meets OLAP: Multidimensional Model and Operators. IJDWM(2010) 33-60