MODELING MULTIDIMENSIONAL SPATIO-TEMPORAL DATA WAREHOUSES IN A CONTEXT OF EVOLVING SPECIFICATIONS

Maryvonne Miquel [a,b], Yvan Bédard [a], Alexandre Brisebois [a], Jacynthe Pouliot [a], Pierre Marchand [a], Jean Brodeur [a,c]

[a] Centre for Research in Geomatics Laval University, Quebec City Canada G1K 7P4 {maryvonne.miquel, yvan.bedard, alexandre.brisebois, jacynthe.pouliot, pierre.marchand }@scg.ulaval.ca,
[b] LISI-INSA de Lyon Bât B. Pascal - 7 av. Capelle, 69621 Villeurbanne Cedex France
[c] Natural Resources Canada, Center of Topographic Information, 21144 King Street West, Sherbrooke Canada J1J2E8 brodeur@RNCan.gc.ca

Commission IV, Working Group IV/1

KEY WORDS: Spatial On-Line Analytical Processing, Multidimensional Data, Data Modelling, Forestry

ABSTRACT:

In this paper, we study some problems linked to the integration of data in a spatio-temporal data warehouse. In many cases, the specifications of the data sets have evolved over time, especially when the observed period is large. Under those circumstances, data sources have temporal, spatial and semantic heterogeneity. In order to explore and analyse spatio-temporal data sets in a SOLAP (Spatial On Line Analytical Processing) application, we propose two approaches to model heterogeneous data in multidimensional structures. The first solution consists in a unique temporally integrated cube with all the data of all epochs. The second solution consists in creating a specific cube (data mart) for each specific view that users want to analyse. The final objective is to support geographic knowledge discovery through data exploration of detailed data for an epoch and of integrated comparable data for time-variant studies. Using a practical example in the field of forestry, we evaluate the implementation of these two models.

RESUMÉ :

Dans cet article, nous abordons les problèmes liés à l'intégration de données spatio-temporelles au sein d'un entrepôt de données. Dans de nombreux cas, notamment lorsque la période d'étude est relativement longue dans le temps, les spécifications des jeux de données évoluent. Dans ce cas, les données sont hétérogènes à la fois des points de vue temporel, spatial et sémantique. Afin d'explorer et d'analyser des jeux de données spatio-temporels dans une application SOLAP (Spatial On Line Analytical Processing), nous proposons deux approches pour modéliser ce type de données dans des structures multidimensionnelles. La première solution consiste à intégrer toutes les données dans un seul cube. La deuxième solution propose de créer un cube spécifique (un marché de données) pour chaque vue que l'utilisateur veut analyser. L'objectif final est de permettre l'extraction de connaissances géographiques par l'exploration des données détaillées associées à une époque et des études temporelles sur les données intégrées et comparatives. A partir d'un exemple pris dans le domaine de la foresterie, nous évaluons l'implémentation de ces deux modèles.

1. INTRODUCTION

On-Line Analytical Processing (OLAP) technology enables users to quickly analyse large sets of data. Hence decision-making is facilitated. OLAP systems are generally based on a three-tiers architecture including a data warehouse with integrated data, an OLAP server for the dimensional view and an OLAP client, i.e. a user interface for the rapid and easy exploration of data (Han, 2001). Similarly, SOLAP (Spatial On-Line Analytical Processing) systems are built to support the rapid and easy spatio-temporal analysis as well as the exploration of data according to a multidimensional approach typical of data warehouses (Bédard, 1997). This approach is comprised of aggregation levels supporting cartographic displays as well as tabular and diagram displays at various levels of detail. SOLAP systems provide the exploration and navigation tools required to analyze and explore spatial data, identify potential clusters, discover potential trends and build hypothesis (Rivest, 2001). However, building spatio-temporal data warehouses for SOLAP applications implies significant work of data integration especially when the data acquisition specifications have evolved over time. In this case, databases sources differ from one epoch to another not only in data coding and structures, but also in semantic contents. In order to permit temporal comparative studies, data must be integrated following a compatible set of temporal, spatial and semantic definitions. In this paper, we focus on some of the data warehouse supply difficulties in the case of conventional multidimensional models coupled to high heterogeneity issues. We expose a practical example in the field of forestry that considers forest maps of three 10-years periods elaborated following different acquisition specifications. We propose two solutions that are implemented and evaluated. At last, we discuss their advantages and disadvantages.

2. MULTIDIMENSIONAL MODELING OF SPATIAL DATA IN A DATA WAREHOUSE

To perform spatio-temporal analysis, multidimensional database modelling is very useful. Multi-dimensional views are produced

Miquel, M., Y. Bédard, A. Brisebois, J. Pouliot, P. Marchand & J. Brodeur, 2002, Modeling Multidimensional Spatio-temporal Data Warehouse in a Context of Evolving Specifications, Joint Int. Symp. ISPRS Commission IV, SDH 2002, 95th Annual CIG Conference, July, 8-12, Ottawa, Canada

when measures are analysed against the different dimension categories of a cube (Marchand, 2001). The most popular multidimensional model for relational OLAP (ROLAP) is certainly the star schema (Kimball, 1996). This model is centred on a fact table containing measures with related dimension tables, which characterize these facts. Each dimension has a number of attributes used for selection or grouping. A dimension is usually organized in hierarchies supporting different levels of data aggregation as well as multiple inheritances. The snowflake schema is a variant model where the hierarchies in the dimensions are explicit following normalized tables. Pedersen and al. (2001) analyse 14 multidimensional data models including star and snowflake models and show that these models do not support requirements such as multi dimension in each dimension, non strict hierarchies, handling change and time and handling different levels of granularity. They define an extended multidimensional data model for these requirements; their model is adapted for imprecise data. Some papers propose solutions for multidimensional structures with spatial data. Han and al. (1998) use a star/snowflake model to build a spatial data cube. They propose the idea of spatial measures with a method to select spatial objects for materialization. Papadias and al. (2001) use the star-schema with spatial dimensions and present methods to process arbitrary aggregations. In both cases, hierarchies of the spatial dimension are unknown at design time and are arbitrarily created by the user. Allowing change in aggregation hierarchies, Eder (2001) introduces a temporal multidimensional data model allowing the registration of temporal versions of dimension data. This solution is based on structure version and supports functions to transform data from one structure version to another. Espil (2001) and Hurtado (2001) also study multidimensional schemas with redefinition of aggregation hierarchies and heterogeneous schemas. In their solutions, they describe a new framework for modelling dimensions.

3. DATA INTEGRATION ISSUES

Spatial information is often organized according to spatial objects with geometric and descriptive attributes. For example, forest inventory information is usually vector-based and represents the boundaries of forest stands with their associate attributes. A forest stand is defined as a part of territory with homogeneous characteristics (species, height, age and so on). In Quebec, Canada, for each inventory, a new map is made by aerial photo-interpretation. Thus, each inventory is a complete new set of spatial data with no reference to the precedent set. Furthermore, from one inventory to another (usually every 10 years), the definition, the acquisition mode and specifications can change significantly, which make temporal, geometric and semantic integration as well as multidimensional modelling a difficult task. To integrate and model spatio-temporal data in a multidimensional structure with evolving specifications, several inter-related problems appear. We present the ones we have encountered once we have described our data sets.

3.1 The Data Sets Used

Three inventories of Montmorency forest compose the data set used in this study. Table 1 gives the characteristics of each inventory.

Year	Number of forest stands
1973	≈ 1 700
1984	≈ 2 400
1992	≈ 3 800

Table 1. Composition of inventories

The number of forest stands varies according to the evolution of the characteristics used for the photo-interpretation because the legislation changes. As specifications evolve over time, the definition and the number of attributes and classes vary from one inventory to another. Over the 20 years period covered by these 3 inventories, we have determined that only 12 % of data types can be compared over time and temporal analysis is hardly achievable.

For example, we present here the different classifications of the Age attribute for each inventory. In 1973, this attribute is composed of five qualitative classes (In regeneration, Young, Regular Mature, Irregular Mature, By stage). In 1984 classes have become quantitative data segmented in 20-years periods whereas in 1992 it is segmented in 15-years periods.

Descriptive data attributed to forest stand characterize the dominant trees. Even if the measured attributes of forest stands change over time, descriptive data can be grouped in time-invariant items such as Age, Density, Height, Diseases and slope.

3.2 Integration Problems

To support spatial and temporal analysis and exploration in the case of these forest inventories, we must cope with four issues: 1- each map is made independently of the previous one for the same territory, leading to delineating forest stands without relationship to the stands in the preceding epoch; 2- spatial referencing systems may change from a map to the other, sometimes without proper metadata, creating spatial matching problems 3- geometric heterogeneity caused by the temporal variation of forest stands definitions and 4- the descriptive heterogeneity caused by the evolution of the definitions and specifications.

Exploration of spatial dimension is the first requirement of our application. In order to supply time-invariant spatial reference, we propose to translate spatial data in an arbitrary regular tessellation. For each inventory, an overlay of the spatial vector-based objects and a fixed tessellation representation is produced. The cells become objects with invariant geometric characteristics and, as such, can be used as spatial reference. Figure 1 shows forest maps of the same area spanning the 3 studied epochs. At the top, we have initial forest stands as they were delimited in each inventory. Below we have the result of the overlay with a regular tessellation. The spatial reference is composed of regular cells as represented at the bottom of Figure 1. Each cell inherits the descriptive attributes of the dominant forest stand of a given inventory. With this solution, a new inventory is easy to add because it does not modify the existing data.

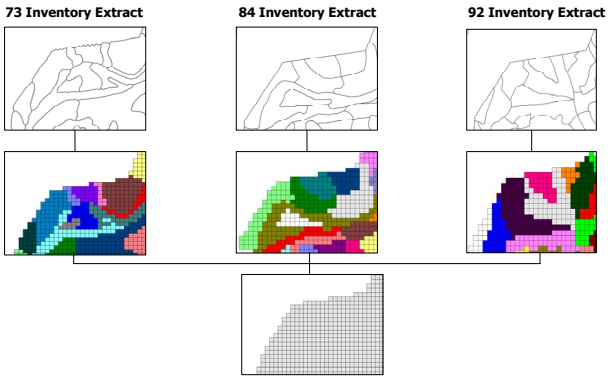

Figure 1. Data transformation using a regular tessellation

The hierarchy of the spatial dimension is derived from forest management rules. The hierarchy has three levels, by grouping cells in forest stands, then in compartments and in units or forest station (figure 2). A cell belongs to different forest stands depending on the inventory. The surface attributed to a cell is pondered in order to maintain correct surface measure of the forest stand. The results of different tests demonstrate that a grid of 20x20 meter is suitable to depict spatially the forest stands in Montmorency forest, it produces 162 000 cells to cover the territory of interest.

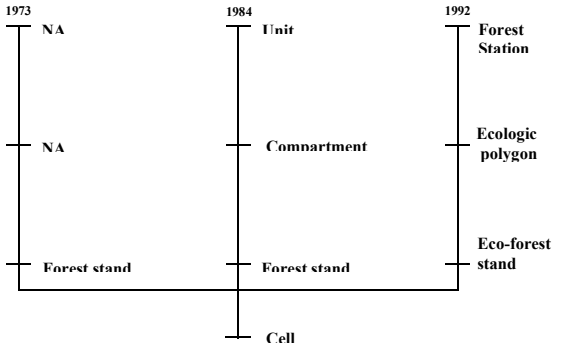

Figure 2. Hierarchy of the spatial dimension based on inventories

With this structure, users get access to the detailed data for each specific epoch i.e. 1973, 1984 and 1992 (it is especially useful for the most recent one). They also get access to evolution of the forest over the 20 of aggregated data. In other words, users access both detailed data for an epoch and to integrated comparable data for time-variant studies. So, we keep all the detailed data even if it is not homogeneous over time and it is up to the user to wisely navigate in the data set (or to the developer to restrain navigation).

Based on a multidimensional conceptual model of the data, data integration consists of building new data groupings by aggregating initial classes in a temporally compatible manner (Rebout, 1998). For each type of attributes, we define new generalized classes based on similar meanings. The aim of these new classes is to obtain comparable categories at the aggregated levels from one inventory to another. Each grouping is a level in a dimension hierarchy. At the lowest level (finest granularity) of a descriptive dimension, data is not time-comparable whereas the upper levels (coarser granularity) are time-comparable hierarchies with roll-up and drill-down functions. In Figure 3, the hierarchy of the dimension Age is shown. The attribute values are grouped in seven Age classes, and three Age Groups. At these two hierarchical levels, measures are time-comparable. For example the mature or out of age territory is identified as a category existing for all the inventories. Hence the evolution of mature or out of age areas during the last 20 years can be studied.

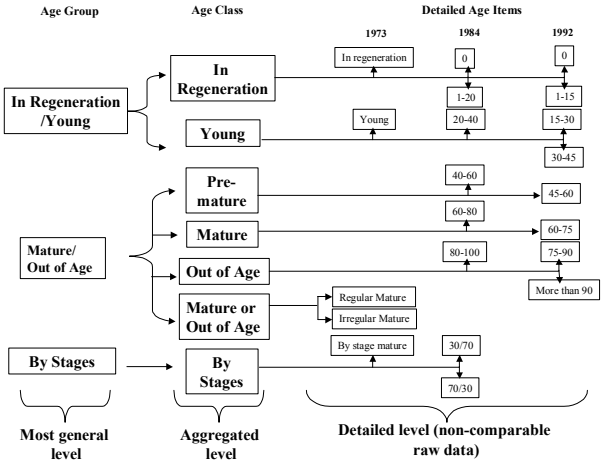

Figure 3. Hierarchies of Age dimension

Figure 4 shows other descriptive dimensions with detailed data at the lowest levels. The hierarchy labels have been chosen according to the vocabulary used in forestry. One can note that the Density dimension is a single level dimension. The codification of forest stand density was maintained over time. Hence there is no integration issue in this case. In all other cases, these new aggregated classifications integrate the data semantically and over time. All the raw data are cleaned and encoded with these classifications in a common database. Then, different cubes can be designed and produced.

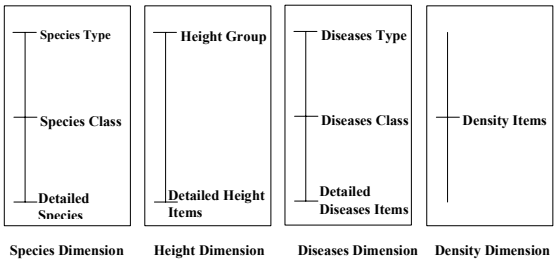

Figure 4. Hierarchy of the Descriptive Dimensions

4. MULTIDIMENSIONAL MODELS

Using these spatial and descriptive dimensions, we design two models that show the link between facts and dimensions. In the star multidimensional schema, the fact table groups measures at the finest granularities of all dimensions. For forest inventories data, facts are measured areas with given characteristics (species, height, age…). In order to allow easy and complete exploration and analysis of this large set of data, we propose two approaches for modelling the spatio-temporal multidimensional data. The first solution is based on a unique temporally integrated cube. The second solution is a combination of 4 multidimensional structures that share

dimension, one for each epoch and one for the 12% of data that are time-comparable at their finest level.

SOLUTION 1: One temporally integrated multi-epoch cube

In the first model, fact table is associated to a cell and to the descriptive dimensions (Figure 5). In this simplified schema, only Age, Height and Species are represented. The temporal variation is integrated within the descriptive and spatial dimensions.

This simple conceptual schema hides some implementation difficulties. The first lies in the storage of the large set of data resulting from the regular tessellation representation. Because of the homogeneous descriptive characteristics of a cell, the resulting multidimensional structure is sparse. The OLAP server will have to optimise the storage of this sparse structure.

The second difficulty lies in the management of the exploration paths (i.e. allowing a priori all dimension combinations). An additional user interface layer needs to be implemented to capture the user actions in order to control the exploration of the warehouse. Time varying queries using detailed descriptive dimensions granularities will be prohibited until allowed for upper granularities studies dealing with several epochs.

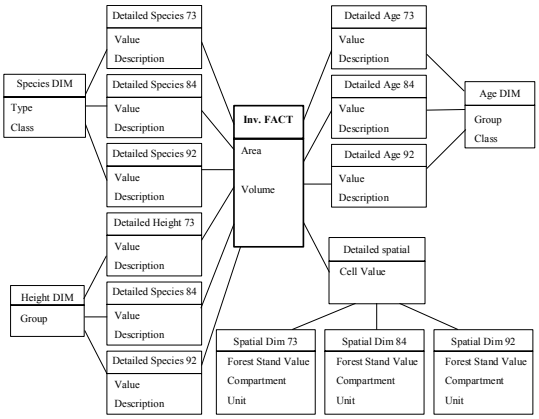

Figure 5. A temporally integrated cube

SOLUTION 2: spatially-specific/ mono-epoch cubes and spatially-unified/multi-epoch cube

The second model is based on the fact constellation schema (Figure 6). In this case, multiple fact tables share dimension tables. In our application, a fact table is implemented for each of our 3 ten-year inventories and the measures are associated to forest stands. The descriptive dimensions include the Age, the Height and the Species. The finest granularities correspond to the detailed attributes for one inventory.

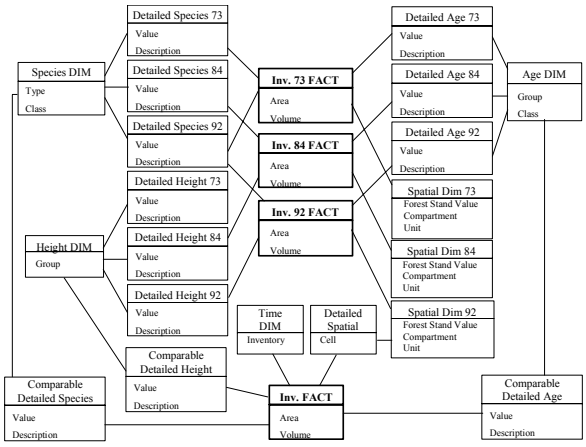

Figure 6. Fact Constellation Schema with 4 Multidimensional Cubes

The heterogeneity of dimensions hierarchies is explicit. A fourth fact table integrates the data which are comparable between the three inventories. This table shares a part of the hierarchical dimensions with the other fact tables. Measures are associated to single cells and one inventory (lowest levels of the spatial and time dimensions). This schema defines 4 multidimensional structures: one for each inventory and one for the integrated data. Each multidimensional cube associated to each inventory is based on forest stands as opposed to the multi-epoch cube which is based on the regular cells. The multidimensional cube with integrated data is unified both from a spatial and a time point of view. This schema optimises the data storage but requires the use of 4 multidimensional structures by the users. From a conceptual point of view, this model is a better representation of the complexity of the data and its intrinsic links.

5. PROTOPTYPE AND EVALUATION

The proposed solutions have been implemented using MS SQL Server and Analysis Services. These models are similar to the logical models previously presented. We defined each hierarchical dimension by one descriptive dimension for each inventory. All the multidimensional structures have been stored in a Multidimensional OLAP (MOLAP) database. In a MOLAP database, information (raw and aggregated data) is stored as series of multidimensional arrays. MOLAP databases respond faster to multidimensional queries as the structures are highly denormalized and stored in RAM.

The SOLAP prototype has been designed to support the spatio-temporal exploration of multidimensional structures using Proclarity and Intergraph Geomedia software components. A Visual Basic Application drives Proclarity functionalities to perform OLAP operators on descriptive data and drives Geomedia functionalities for spatial navigation and cartographic displays.

Figure 7 shows the interface of the SOLAP application. It is composed of two windows. The first one, on the right, is a classical OLAP interface with selection of dimensions to analyse, the navigation operators and the data displays. Using this interface, a user explores his data and obtains the

representation of the selected measures by histogram and tabular displays. On the left, the spatial window presents the map of the selected data. Between these two windows, the same graphic semiology codes are used (colour, pattern, etc.).

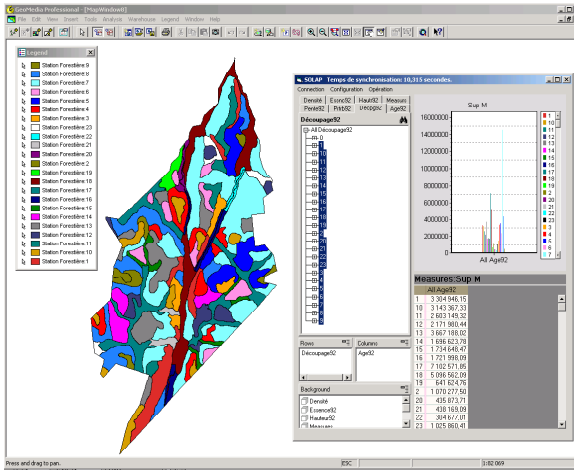

Figure 7. Interface of the SOLAP Application

Table 2 presents a quantitative evaluation of the two previous described solutions. The number of elements in the fact table depends on data organisation. The cube size is the storage size. Finally, performance gain is estimated by a ratio of pre-calculated data over stored data.

When we built cubes, we used an SQL-Server Analyse Services functionality to determine how many aggregations to create. The wizard adds aggregations until the performance gain reaches a specified percentage. We tried several values and chose a compromise between a percentage and a time running to optimise the structures.

	Elements of Fact Table	Cube Size (MB)	Performance gain
First Solution 1 temporally integrated cube	162 000	17.73	9 %
Second Solution 3 spatially-specific / mono-epoch cubes	From 1 700 to 3 800	From 0.15 to 0.65	100 %
1 spatially-unified /multi-epoch cube	496 000	18.5	50 %

Table 2. Quantitative evaluation of the implementation

With the first solution, the cube has a fixed number of elements into the fact table because the structure is stored by cell. By adding a new inventory, this value will not vary. Only the size of the multidimensional structure (actually 17.7 MB) will increase. The performance gain is very low (9%) indicating that almost any aggregation is pre-calculated. Here, the wizard failed to optimise structure in an acceptable delay.

With the second solution, four physical multidimensional structures are implemented. For the structure associated to each inventory, the number of table fact elements and the structure size depend on the number of forest stands because descriptive data are grouped according to forest stands. The fact table of the spatially-unified structure has an element for each cell and each inventory. Half the aggregated values are calculated and stored. In this solution, data are duplicated between the cubes associated to each inventory and the spatially-unified and time-integrated cube. More aggregations are pre-calculated, increasing the total storage size. The comparative table 1 displays the link between the size of the storage structure and the performance gain.

	Temporal Analysis	Spatial Analysis	Detailed data	Naviga -tion
First Solution 1 temporally integrated cube	Possible	Cell level	Yes	For expert
Second Solution 3 spatially-specific / mono-epoch cubes 1 spatial-unified/multi-epoch cube	Easy for the multi-epoch cube	Forest stand level Cell level	Yes	Easy and intuitive

Table 3. Qualitative evaluation of the implementation

Capabilities of the implemented solutions are evaluated in table 3. Both solutions allow analysis on detailed data. Temporal analysis is possible but difficult in the first solution because time is not a specific dimension. However all the necessary information is available in the stored cube. In this solution, spatial navigation in the smallest level (the cell) is possible allowing very abundant spatial analysis. Generally, navigation in this cube is quite difficult because a user can select any view he wants without any assistance and constraint. He can perform faulty analysis by comparing non-comparable detailed data and the result of complex queries depends on the order of the elementary queries. On the other hand, only possible temporal comparisons are implemented in the second solution, so the navigation is easy and secure in the multi epoch cube. To analyse data on one epoch, a user simply selects the cube associated to the inventory he wants to study. All the detailed data are easily available with classical OLAP operators. Users can navigate in hierarchical dimensions with different levels. These aggregated levels are used also in the spatially-unified cube in which temporal analysis can be performed.

6. CONCLUSION

In this work, we propose solutions to design multidimensional structures when source data sets have temporal, spatial and semantic variations. First, time-variant objects (forest stands) are converted in a fixed, invariant tessellation (cells) for proper spatial referencing. Then, new classifications of descriptive data are introduced by grouping initial detailed classes in comparable time-invariant super-classes. At last, several multidimensional structures are designed in order to explore data with a SOLAP application. Two opposed approaches have been implemented. The first solution consists in a temporally-

integrated cube with all the data (detailed and aggregated), even if a lot of comparisons are meaningless at the lowest level. This solution is valid only if the SOLAP front-end tools guide data exploration by imposing navigation constraints. The design of the OLAP server is very simple but front-end tools need to be specifically designed. The second solution consists in creating a specific cube (data mart) for each specific view that users want to analyse. In this approach, navigation constraints are taken into account immediately in the design of the multidimensional structures. The analysis and the exploration of data can be done with standard front-ends tools for the descriptive part of data.

REFERENCES

Bédard, Y. 1997. Spatial OLAP, Vidéoconférence. 2ème Forum annuel sur la R-D, Géomatique VI: Un monde accessible, 13-14 novembre, Montréal.

Eder J., Koncilia C., 2001 Changes of dimension in Temporal Data, 3rd International Conference on Data Warehousing and Knowledge Discovery - Munich, Sept 3-7

Espil M, Vaisman 2001 A. Efficient Intentional Redefinition of Aggregation Hierarchies in Multidimensional Databases, DOLAP

Han J., Stefanovic N., and Koperski K., 1998 Selective Materialization: An Efficient Method for Spatial Data Cube Construction, Proc. Pacific-Asia Conf. on Knowledge Discovery and Data Mining Melbourne, Australia, pp. 144-158.

Han, J., Kamber M., 2001 Data Mining: Concepts and Techniques, Morgan Kaufmann Pub

Hurtado C., Mendelzon A., 2001 Reasoning about Summarizability in Heterogeneous Multidimensional Schemas In Proc. ICDT'01

Kimball R., 1996 The Data Warehouse Toolkit. J. Wiley and Sons

Marchand, P., Bédard, Y. and Edwards, G., 2001. A hypercube-based method for spatio-temporal exploration and analysis. GeoInformatica. Submitted, in correction.

Papadias D., Kalnis P., Zhang J., Tao Y., 2001 Efficient OLAP Operations in Spatial Data Warehouses, Symposium on Spatial and Temporal Databases (SSTD), pp 443-459

Pedersen T.B., Jensen C., Dyreson C., 2001 A foundation for capturing and querying complex multidimensional data, Information Systems 26 383-423

Rebout, C., 1998. Adaptation d'une base de données pour une application SOLAP pour l'aide à l'aménagement intégré des ressources forestières, Université Joseph Fourier, Grenoble, France.

Rivest, S., Bédard, Y. & Marchand P., 2001, Towards better support for spatial decision-making: Defining the characteristics of Spatial On-Line Analytical Processing (SOLAP), Geomatica, the journal of the Canadian Institute of Geomatics.

ACKNOWLEDGEMENTS

The authors wish to acknowledge the financial support of the GEOIDE Network of Centers of Excellence via the DEC#2 project: Designing the technological foundations of geospatial decision-making with the World Wide Web. Maryvonne Miquel thanks the Center of Research in Geomatics (CRG), Laval University and especially Yvan Bedard and his team for welcome during one-year sabbatical.