On-Line Analysis in Environmental Management Based on Multi-dimensional Modelling

ANNA G. ROZEVA

Department of Computer Systems and Informatics, University of Forestry, 10 Kliment Ohridski Bld., 1799 Sofia, Bulgaria,
E-mail: arozeva@ltu.acad.bg

Abstract. The paper deals with real-time complex analysis and decision support to environmental management by designing a suitable model for natural ecosystem's description. Environmental data values are obtained through different types of monitoring. Environmental management has to be provided with a multi-aspect reporting capability, powerful and informative visualization enabling comparisons of registered parameter values in different locations and time periods, on-line aggregations of data, analysis of interrelationships among different factors and forecasting, i.e. on-line analytical processing (OLAP). In order to provide for these needs, we suggested monitored data to be described by multidimensional technique. We applied the principles of multidimensionality to a part of a river ecosystem and created a model of water quality as a basis for providing environmental management with on-line analytical capability. When designing the model, we introduced hierarchical levels of dimensions, defined special type of dimensions such as attributes, organized parameter values as time series data. We developed a practical tool for on-line analysis based on the elaborated model. It provides for ad-hoc navigation through dimensions, attributes and time periods, powerful visualization and accessibility of data from heterogeneous sources (OLAP cubes and Geographic Information Systems). The tool has a broad applicability. It can be easily redesigned to any type of ecosystem.

Key words: environmental management, river ecosystem, on-line analytical processing (OLAP), multidimensional modelling, data cube, object-oriented application development.

INTRODUCTION

The performance of thorough and complex analysis of an ecosystem's state and behaviour can be achieved on the basis of suitable modelling of monitored data. The adequacy of the applied model turns out to be of main importance for the information value of the performed analysis. Qualitative analytical results provide steady support for further decision-making in the ecosystem's management. Monitoring of a part of the Iskar River ecosystem in Bulgaria has been performed and water quality data has been registered during a seven-year period. Monitoring of river water quality has been performed also in course of TEMPUS Project on Monitoring Bulgarian Rivers (JepTempus, 1997). The data obtained thereby has been elaborated using the object-relational approach (Tsankova, 2001). The approach, which we implemented here, is the multidimensional one. It provides a basis for powerful on-line information analysis.

Parameter values of river water quality are presented as cells in a multidimensional cube. The dimensions represent characteristics of the ecosystem that the measured parameters are related to, as well as different focuses and viewpoints for their analysis. For the part of the ecosystem under discussion, dimensions are: time, location (points, where measurements are performed), region (part ot the river's flow), point type (by city, by village), river, pollution (medium, high, very high), monitoring type (mobile or stationary chemical laboratory). The multidimensional model, discussed by Agrawal (1997), Inmon (1992) and Kimball (1996), has been chosen. It provides the basis needed for on-line analytical processing (OLAP). OLAP is a "dynamic analysis of historical data for flexible creation, manipulation, animation and synthesis of information, according to consolidation approaches" (Codd, 1993; Vassiliadis, 1999). OLAP technology and design of data cubes have been implemented mainly in the area of business intelligence and business solutions (Rozeva, 2000).

The aim of the paper is to broaden OLAP applicability to the area of environmental management in order to raise the quality and value of information provided to ecosystem's management.

MULTIDIMENSIONAL MODEL OF RIVER WATER ECOSYSTEM

We use an iterative analytical process for designing the model. Our goal is, from the data available in Relational Data Bases (RDB), spreadsheets and text files, to elaborate qualified decision supporting information. Fig. 1 shows the structure of a RDB where monitoring data, obtained through measurements of physical, chemical and biological parameters of river water quality, has been stored.

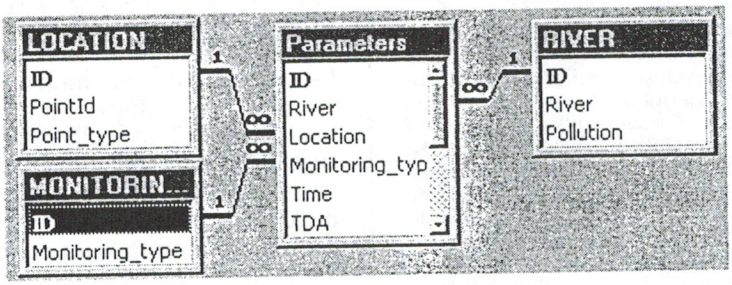

Fig. 1. RDB for river water quality data

Data kept in multiple related tables in the RDB can be modelled as a multidimensional cube with cells, containing measurements of monitored parameter values (Fig. 2).

Multidimensional modelling process consists of several iteration steps.

Identification of dimensions. The dimensions are characteristics (descriptions) of the part of the ecosystem being considered. In the initial RDB except for time, they are of text data type. We have elaborated the following dimensions:

1. **River**: Iskar;
2. **Location**: P1, P2, P3, P4, P5, P6, P7.
3. **Region**: Upper flow, UpMiddle flow, Middle flow, MidLower flow, Lower flow, etc.;
4. **Time**: Mar93, Apr93, May93, Sep93, Oct93, Nov93, Mar94, etc., i.e. spring and autumn;
5. **Pollution**: medium, high, very high;
6. **Monitoring Type**: mobile, stationary;
7. **Point Type**: by city, by village.

The elements of each dimension are enumerated. A cell in the cube contains data for a monitored parameter. In JepTempus (1997), the following water quality parameters have been monitored: hydrogen ion concentration PH, water temperature TDA, amonium NH_4, colibacteria, NO_3, dissolved oxygen OKISL, solid substance content TDS, biochemical oxygen demand BPK5.

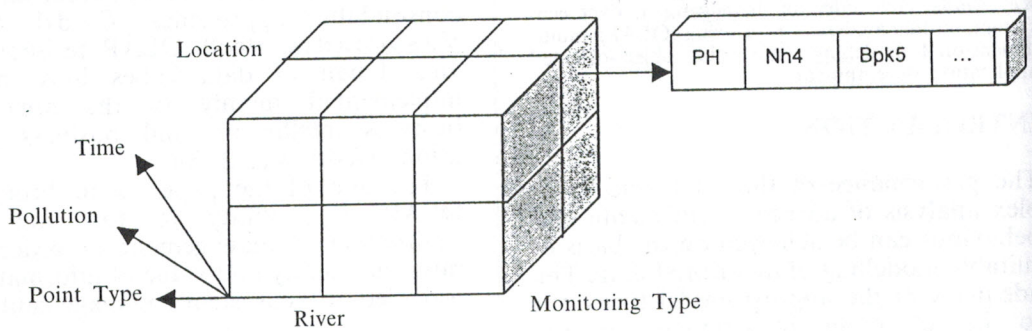

Fig. 2. Cube of river water quality data

Reduction of the number of dimensions by defining hierarchies. We examine the dimensions Location and Region. Both of them reveal characteristics of the monitored parameters, referring to geography. Location concerns the points where measurements are performed (P1, P2, P3, etc.). Region refers to the different parts of the river flow - Upper, UpMiddle, Middle, Lower, etc. If the points of performing monitoring and the parts of the river flow are defined as separate dimensions the designed model will be with cardinality of 7. This cube will be sparse (it will have a lot of empty cells). For example, since points P1, P2 and P3 belong to the Upper flow only, the cells at the intersection of the rest of the points (P4, P5, P6,...) and Upper flow will have no values. Thus the cube, resulting out of the modelling process, will be with high cardinality and sparsity at the same time. In order to avoid this, we discard Region as a separate dimension and define Location as consisting of several scale layers. These layers are Point and Region, which roll up into the Total layer (Fig. 3).

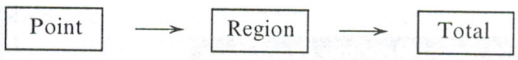

Fig. 3. Dimension Location with hierarchical structure

Data that populates the Point layer of the cube, is obtained in the process of monitoring. Data at the Region layer is consolidated, i.e. calculated through aggregation of the Point layer data. The Total layer contains aggregated information from the elements of the Region layer. The type of aggregation has to be specified. Typical aggregation operations are: sum, average, minimum, maximum, etc. For aggregating data values of water quality parameters, we consider average, minimum and maximum as appropriate ones. The hierarchical structure of dimensions provides for storing aggregated as well as detailed information in the cube. Management in general requires generalized view of data. The cube with the stated structure of dimensions enables access to aggregated information, which is the most suitable for analytical purposes.

Reduction of the number of dimensions by introducing attributes. The higher the multidimensionality of the model is, the larger space and processing time is needed. Reduction of the number of dimensions in the cube is very important for obtaining a model with reasonable size and enhanced productivity. Each dimension has to be analysed for the nature of its elements. When examining Pollution and Point Type, we find out that each one of them has a small number of elements with values, known in advance. They appeare in combination with the different values of another dimension's elements. Pollution values appear with the elements of the River dimension and Point Type values – with the Location element values. Pollution and Point Type qualify, distinguish River and Location dimensions and hence do not exist independently. We define each of them as an attribute of another dimension. When attributes of dimensions are defined, the model's size is further reduced.

Optimisation of the time dimension. Measurements of water quality parameters are performed during definite time periods. This means that the data values obtained thereby are historical, i.e. time-dependent. For the sake of reducing the cube's cardinality, we decided to attach Time as a property (meta data) to the value in each cell. As a result of this, data values in the cells of the cube turn into series (arrays). Fig. 4 shows the structure of parameter values as series with their corresponding meta data. Meta data determine the period of time when data values appear, their periodicity, the rule for converting data from one periodicity to another (monthly data to quarterly – spring/autumn, yearly, etc.). In this paper, the conversion rule is accepted to be Average. In case that same data value appears several times in a series, it is kept only once while the meta data shows the number of times it is repeated in the series. Data series of the type, shown in

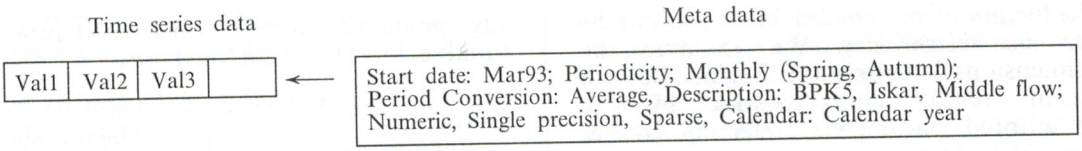

Fig. 4. Cell content as time series data

Fig. 4, is created for each one of the monitored water quality parameters. Their corresponding meta data may be different. Thus, the optimization of Time dimension consists in eliminating it from the model by converting data values into time series with support of proper meta data.

Establishment of aggregations. Typically, ecosystem's management requires aggregated instead of detailed data for reporting and analytical purposes and, hence, support for proper decision-making. The response time for processing queries for aggregating huge amounts of data in the cubes (consolidation) usually is achieved by raising the number of cells in the cube to a certain extent in order to keep aggregated values. It is important to determine the dimensions, for which consolidations make sense and the proper consolidation rule. The dimensions with hierarchical structure are suitable for performing consolidations. For the part of the river ecosystem being modelled, this is only the Location dimension. The average consolidation rule has been chosen for it. River and Monitoring type dimensions are not suitable for consolidation.

After performing the so stated steps of the iterative analytical process, we obtained the following final multidimensional model (Fig. 5).

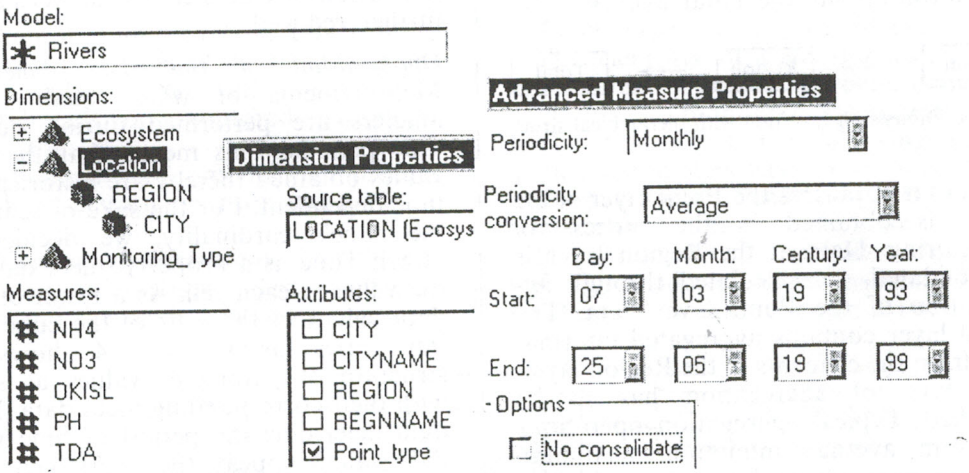

Fig. 5. Optimized multidimensional model for part of a river ecosystem. The Location dimension has two hierarchical layers and an attribute called Point type. River and Monitoring type are not hierarchical. Consolidation of parameter values (measures) is performed for the hierarchical dimension

turns out to be significant. It can be reduced by performing aggregations once and storing them in the data cube, together with the detailed data stored in the RDB. This reduction of response time

IMPLEMENTATION

We developed a practical tool for on-line analysis of river water quality monitored data stored in the optimized cube struc-

ture. It has been implemented in an object-oriented design environment (Pilot Designer, Version 6.2 1999, Pilot Software, Inc.). This environment enables design of applications with full implementation of the advantages of the OLAP technique based on the multidimensional model. The application provides the main OLAP characteristics, such as:

- ♦ report generation in the most intuitive table-like across/down/page form with on-line calculation capability;
- ♦ live ad-hoc data access and analysis – on-line change of the focus on data, month, season, navigation through hierarchical dimensions, etc;
- ♦ visualisation through powerful charting capability;
- ♦ evaluation of trends and forecasting capability;
- ♦ accessibility of data from heterogeneous sources such as data cubes, Geographic Information Systems (GIS), etc.

The object-oriented design environment implies definition of objects and actions they fulfil on occurrence of certain events. The application contains several sheets that are connected and navigation facilities are provided. The main sheet enables table-like report generation, on-line change of the focus to data which are available through the toolbar buttons and menu item "Parameters", time-based analysis through the menu item "Season", calculations with data and charting the displayed data. It has been developed by using the following object types:

- table with source data from the River multidimensional model, shown in Fig. 5. Time dimension (month March with yearly periodicity) and parameter BPK5 are viewed across, Location dimension for the Point layer is viewed down. The name of the Iskar river and the monitoring type (mobile) are viewed on each page. Refresh, scroll and select events are assigned for the whole table on occurrence of which two variables are created – for the data, displayed in the table and the page title;
- chart with source data from the variables, defined for the table object;
- selection, which contains a list of chart types with a click event assigned to it for changing the chart type;
- menu for navigation among application sheets, selection of time period and parameter/parameters to display;
- toolbar with buttons enabling drilling up/down and changing the focus on data.

The sheet designed by these objects is shown in Fig. 6.

The menu item Forecast provides for forecasting parameter values based on the values obtained through monitoring. The result of BPK5 spring forecast at monitoring point P4 for years 2000 – 2002,

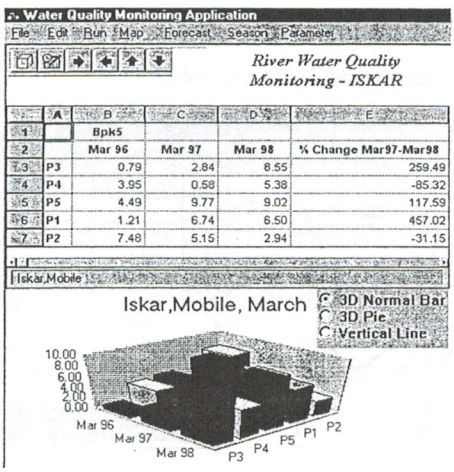

Fig. 6. On-line analysis with multi aspect reporting and flexible charting capabilities

implementing adaptive response smoothing and Quadratic forecast method, is shown in Fig. 7. The selection of smoothing and forecast methods depends upon the existing trends in data. This smoothing method has been chosen because the yearly data series included in the analysis does not display a trend. Adaptive response smoothing is one, which is responsive to changes in the pattern of the data. We have chosen the type of forecast method to fit the smoothed and forecasted points when the values for criteria Goodness of Fit and R Squared have been adjusted to the lowest values. The results of the forecast differ from the measured values about 7 %.

The sheet "Map Analysis", opened by the menu item "Map", performs analysis of GIS data by implementing MapInfo (MapInfo Professional 4.5. Supplement. 1997. MapInfo Corporation) object. By clicking on a map river name, the parameter values measured at all points for chosen season (spring or autumn) are displayed. MapX object provides access to GIS data and point by point analysis (Fig. 8).

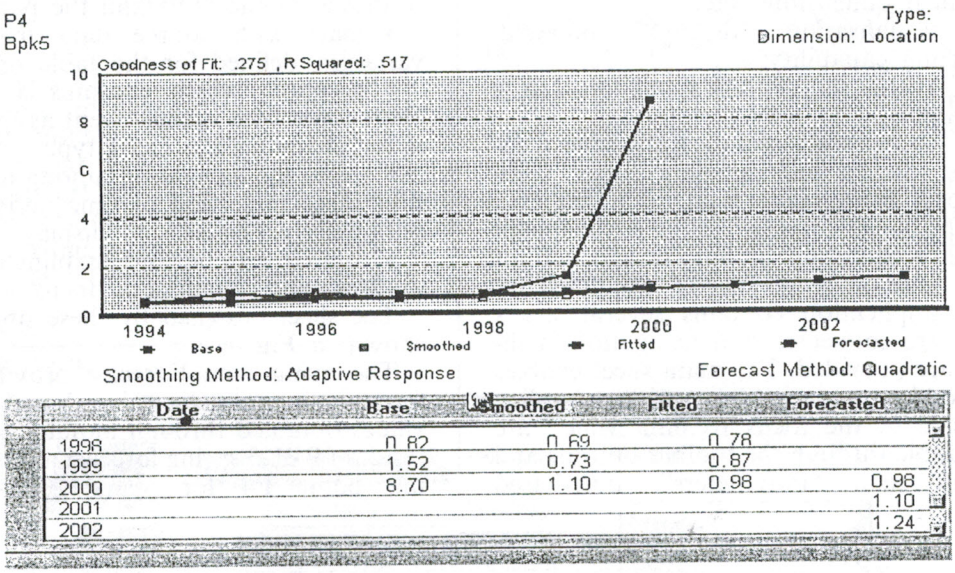

Fig. 7. Analysis with trend evaluation and forecast capability

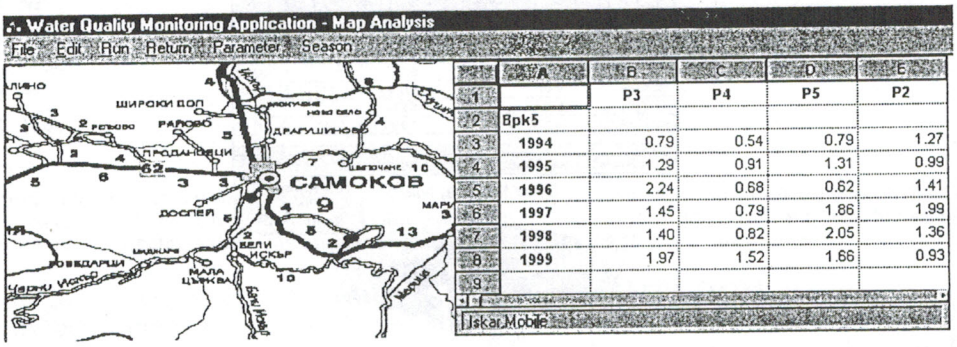

Fig. 8. Point by point analysis with access to GIS data type

CONCLUSION

We designed a data cube to serve as a basis for providing on-line analysis and decision support capabilities to environmental management. In the model's design process, we achieved significant advantages concerning space and query processing time. They refer to reducing the size of the cube through transforming some of the dimensions into hierarchical layers or attributes of other dimensions. Thus, we obtained a structure with reduced total number of cells and a raised population with data. Further advantage of the model concerns the storing of series of data in the cube's cells, containing water quality parameter values over a time period. We performed consolidation of detail data for the layers of hierarchical dimensions and stored the resulting aggregated data in the cube for optimizing the query processing time. We developed a practical application in an object-oriented design environment for analyzing the data stored in the cube. It is with enhanced performance and has the advantage to provide management with the ability to view and analyze all aspects of the ecosystem at once. The main requirements for OLAP analysis are implemented therein and management is provided with valuable and adequate information derived out of the monitored data in a useful decision supporting format. The tool can be applied to other types of ecosystems, as well.

REFERENCES

Agrawal, R., A. Gupta, S. Sarawagi. 1997. Modelling Multidimensional Databases. – Proc. 13th Int. Conf. on Data Engineering, Birmingham, UK. IEEE, 232–248.

Codd, E. F., S. B. Codd, C. T. Salley. 1993. Beyond Decision Support. – Computerworld, **27**, 87–95.

Inmon, W. H. 1992. Building the Data Warehouse. – QED Press/John Wiley, p. 292.

JepTempus 07521. 1997. Environmental Education and New Information Technologies. – Final Report, IPN, Kiel, Germany.

Kimball, R. 1996. Practical Techniques for Building Dimensional Data Warehouses. – John Wiley, New York, p. 287.

Rozeva, A. 2000. Data Warehouse – Contemporary Technology for Information Support of Business Management. – Management and Sustainable Development, University of Forestry, Sofia, **3** (1–2), 32–38.

Tsankova, R. 2001. Geoinformation Model for River Water Quality Management. – Journal of Balkan Ecology, **4** (1), 25–34.

Vassiliadis, P., T. Sellis. 1999. A Survey of Logical Models for OLAP Databases. – SIGMOD Record, **28** (4), 64–77.