Seismological Data Warehousing and Mining:
A Survey

Gerasimos Marketos,University of Piraeus, Greece

Yannis Theodoridis, University of Piraeus, Greece

Ioannis S. Kalogeras, National Observatory of Athens, Greece

ABSTRACT

Earthquake data composes an ever increasing collection of earth science information for post-processing analysis. Earth scientists, local or national administration officers and so forth, are working with these data collections for scientific or planning purposes. In this article, we discuss the architecture of a so-called seismic data management and mining system (SDMMS) for quick and easy data collection, processing, and visualization. The SDMMS architecture includes, among others, a seismological database for efficient and effective querying and a seismological data warehouse for OLAP analysis and data mining. We provide template schemes for these two components as well as examples of their functionality towards the support of decision making. We also provide a comparative survey of existing operational or prototype SDMMS.

Keywords: *data mining; decision making; data warehousing; seismological databases*

INTRODUCTION

For centuries, humans have been feeling, recording and studying earthquake phenomena. Taking into account that at least one earthquake of magnitude $M < 3$ ($M > 3$) occurs every one second (every ten minutes, respectively) worldwide, the seismic data collection is huge and rapidly increasing. Scientists record this information in order to describe and study tectonic activity, which is described by recording attributes about geographic information (epicenter location and disaster areas), time of event, magnitude, depth, an so forth.

On the other hand, computer engineers specialized in the area of Information & Knowledge Management find an invaluable "data treasure", which they can process and analyze helping in the discovery of knowledge from this data. Recently, a number of applications for the management and analysis of seismological or, in general, geophysical data, have been proposed in the literature by Andrienko and Andrienko (1999), Kretschmer and Roccatagliata (2000), Theodoridis (2003), and Yu (2005). In general, the collaboration between the data mining community and physical scientists

has been only recently launched (Behnke & Dobinson, 2000).

Desirable components of a so-called *seismic data management and mining system* (SDMMS) include tools for quick and easy data exploration and inspection, algorithms for generating historic profiles of specific geographic areas and time periods, techniques providing the association of seismic data with other geophysical parameters of interest, such as geological morphology, and top line visualization components using geographic and other thematic-oriented (e.g., topological and climatic) maps for the presentation of data to the user and supporting sophisticated user interaction.

In summary, we classify users that an SDMMS should support in three profiles:

- **Researchers of geophysical sciences**, interested in constructing and visualizing seismic profiles of certain regions during specific time periods or in discovering regions of similar seismic behavior.
- **Public administration officers**, requesting for information such as distances between epicenters and other demographical entities (schools, hospitals, heavy industries, etc.).
- **Citizens ("Web surfers")**, searching for seismic activity, thus querying the system for seismic properties of general interest, for example, for finding all epicenters of earthquakes in distance no more than 50Km from their favorite place.

The availability of systems following the proposed SDMMS architecture provides users a wealth of information about earthquakes assisting in awareness and understanding, two critical factors for decision making, either at individual or at administration level.

The rest of the article is organized as follows. Initially, we sketch a desired SDMMS architecture, including its database and data warehouse design. The section that follows, presents querying, online analytical processing (OLAP) and data mining functionality an SDMMS could offer, putting emphasis on the support of decision making. Furthermore, we survey and compare proposed systems and tools found in the literature for the management of seismological or, in general, earth science data. Conclusions are drawn in the last section.

THE ARCHITECTURE OF A SEISMIC DATA MANAGEMENT AND MINING SYSTEM

Earthquake phenomena are instantly recorded by a number of organizations (e.g., Institutes of Geodynamics and Schools of Physics) worldwide. The architecture of a SDMMS might allow for the integration of several remote sources. The aim is to collect and analyze the most accurate seismic data among different sources. Obviously, some sources provide data about the same earthquakes though with slight differences in their details (e.g., the magnitude or the exact timestamp of the recorded earthquake). SDMMS should be able to integrate the remote sources in a proper way by refining and homogenizing raw data.

Collected data can be stored in a local database and/or a data warehouse (for simple querying and analysis for decision making, respectively). In general, data within the database is dynamic and detailed, while that within the data warehouse is static and summarized (this is because the modifications of the former are continuous, while the latter are subjected to periodical updates).

Figure 1 presents the proposed abstract architecture that serves the task of collecting data from several sources around the world and storing them in a local repository (database and/or data warehouse). A mediator is responsible for the management of the process from the extraction of data from their sources until their load into the local repository, the so-called *extract-transform-load* (ETL) approach.

Formats for storing seismic data include SEG-Y and SEG P1-P4. SEG-Y (Barry et al., 1975) is used by the U.S. Geological Survey and consists of a header and a trace data block. SEG P1-P4 (SEG, 2006) has been developed

by SEG Subcommittee on Potential Fields and Positioning Standards in order to standardize data exchange formats. The oil and gas industry in conjunction with PPDM, a not-for-profit organization that develops and maintains standards for the resource industry, run a Data Exchange Project that will guarantee interoperability between businesses within the energy industry (PPDM, 2006). The aim of the project is to replace SEG-Y and SEG P1-P4 formats with new ones based in open source technologies (like XML and SOAP).

In the following subsections, we present efficient design proposals for the two components of the local repository of a SDMMS, namely the seismological database (SDMMS database subsection) and the seismological data warehouse (SDMMS data warehouse subsection).

SDMMS Database

Remote sources provide SDMMS with a variety of seismological information to be stored in the local database. Figure 2 illustrates the conceptual design (Entity-Relationship diagram) of a local database proposed for SDMMS purposes.

QUAKE contains the minimum information required to describe an earthquake event includes *timestamp* of its appearance, *location* (latitude / longitude coordinates) and *depth*. On the other hand, this information only is not adequate for user-friendly querying and further data analysis as one wish to know more about the geographical areas where an earthquake occurred. For this purpose, the addition of FLINN_AREA assists on the geographical positioning of both the earthquake epicenter and the affected sites using the Flinn & Engdahl geographical terminology (Young et al., 1996) that partitions world in disjoint polygons. Moreover, FAULT includes details about the *seismogenic fault* related with an earthquake (name of the fault, its characterization, strike, slip and rake of plates, etc.), extracted from bibliography, e.g. (Kiratzi & Louvari, 2003); see Figure 3 for an illustration of faults and plates worldwide.

SITE stores demographical and other information about the primitive *administrative partitions* of a country (e.g., counties or municipalities) with information about population and so forth, while GEOLOGY describes the *geological morphology* of a site so that we can discover how the different morphological classes are affected by earthquakes.

EFFECT records *macroseismic intensity* observed at a site as a result of an earthquake. Other attributes of this entity might include the

Figure 1. A general SDMMS architecture proposed for seismological data management

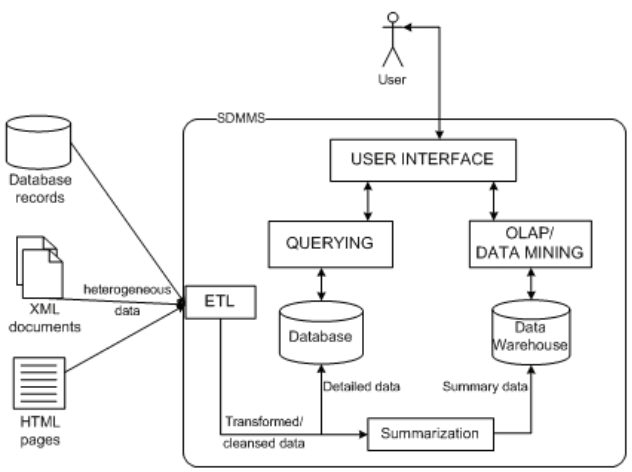


Figure 2. The proposed E-R diagram of a seismological database for SDMMS purposes

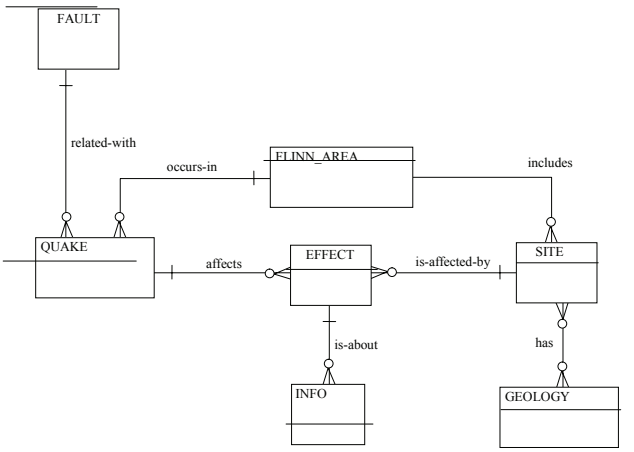
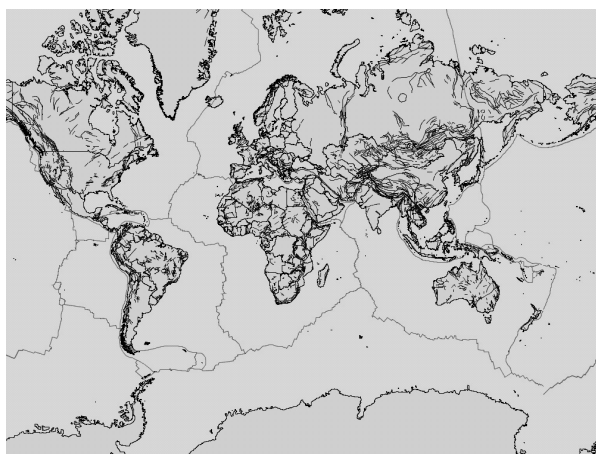


Figure 3. Faults and plates worldwide (Seismo-Surfer, 2006)



epicentral and *hypocentral distance* and the *azimuth* (the angle between the site-epicenter line and the line of North). Finally, an auxiliary entity (INFO) might include complementary *multimedia material*, such as pictures, audio/video descriptions, references and so forth. about earthquake effects.

SDMMS Data Warehouse

A *data warehouse* is defined as a subject-oriented, integrated, time-variant, non-volatile collection of data in support of management decision-making process (Inmon, 1996). Data warehouses are usually based on a multi-dimensional data model, which views data in the form of a data cube (Agarwal et al., 1996). A data cube allows data to be modeled and viewed in multiple dimensions and is typically implemented by adopting a star (or snowflake) schema model, according to which the data warehouse consists of a *fact table* (schematically, at the center of the star) surrounded by a

set of *dimensional tables* related with the fact table. For SDMMS purposes, dimensional tables should maintain at least information (e.g., hierarchies) about *magnitude*, *intensity*, *geography*, *time dimension*, and so forth. (the so-called *dimensions* of the data cube), while the fact table should contain measures on seismological data, such as the *number of earthquakes*, *minimum/maximum depth*, and so forth, as well as keys to related dimensional tables (Figure 4). Since geography is a key issue in SDMMS, involved in dimensions and/or measures, what we propose here is a spatial data warehouse (Stefanovic et al., 2000).

In particular, dimension *time* consists of a hierarchy that represents time periods in which an earthquake happened. Dimensions *magnitude*, *intensity* and *depth* consist of intervals rather than hierarchies. They represent classes of *magnitude*, *intensity* and *depth* so that we can categorize the earthquake phenomena. Dimensions *geography* and *geology* represent the geographical area in which an earthquake happened and the geological morphology of this area, respectively. As for the fact table, the cardinality of a certain type of earthquake events (*number_of_quakes*) together with *min/max* and *average* information are stored.

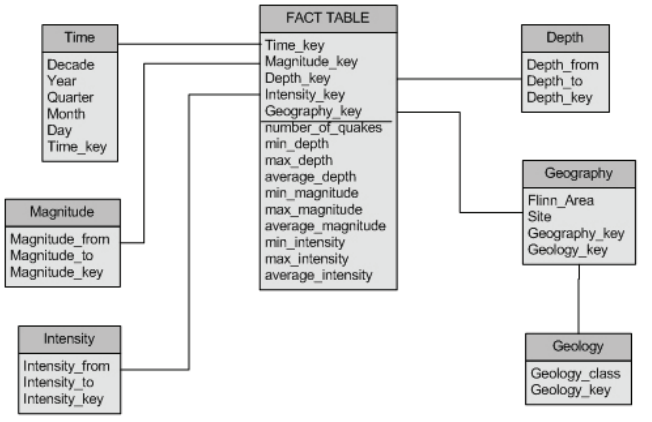
In the following section, we present examples of operations that illustrate the usefulness of a database and a data warehouse that follow the schemes of Figure 2 and Figure 4, respectively.

QUERYING, OLAP ANALYSIS, AND MINING

Traditional database management systems (DBMS) are known as operational database or OLTP (online transaction processing) systems as they support the daily storage and retrieval needs of an information system. Apart from querying, they support three main operations (insertions, updates and deletions) that can be formalized and executed over a DBMS using a structured query language (SQL).

Nevertheless, maintaining summary data in a local data warehouse can be used for data analysis purposes. Two popular techniques for analyzing data and interpreting their meaning are OLAP analysis and data mining. An important aspect in decision making is the level of details that the decision-maker needs. Middle and upper management make complex and important decisions and therefore detailed data can not satisfy these requirements. Summarized data and hidden knowledge acquiring from the stored data can lead to better decisions. Similarly, summarized seismological data are of particular interest to earth scientists because they can study the phenomenon from a higher level and search for hidden, previously unknown knowledge.

Figure 4. A spatial data warehouse design proposed for SDMMS purposes



Querying the Database

Querying seismological databases involves spatiotemporal concepts like snapshots, changes of objects and maps, motion and phenomena (Pfoser & Tryfona, 1998; Theodoridis, 2003). In particular, SDMMS should provide at least the following database querying functionality:

- **Retrieval of spatial information given a temporal instance:** This concept is used, for example, when we are dealing with records including position (latitude and longitude of earthquake epicenter) and time of earthquake realization together with attributes like magnitude, depth of epicenter, and so on.
- **Retrieval of spatial information given a temporal interval:** This way, evolution of spatial objects over time is captured (assume, for example, that we are interested in recording the duration of an earthquake and how certain parameters of the phenomenon vary throughout the time interval of its duration).
- **Overlay of spatial information on layers given a temporal instance or interval:** The combination of layers and time information results into snapshots of a layer. For example, this kind of modeling is used when we are interested in magnitude thematic maps of earthquakes realized during a specific day inside a specific area (temporal instance) or modeling the whole sequence of earthquakes, including pre- and aftershocks (using the notion of layers in time intervals).

Examples of typical queries involving the spatial and the temporal dimension of seismological data are the following (Theodoridis, 2003):

- Find the ten epicenters of earthquakes realized during the past four months, which reside more closely to a given location.
- Find all epicenters of earthquakes residing in a certain region, with a magnitude M >5 and a realization time in the past four months.
- (Assuming multiple layers of information, e.g., corresponding to main cities' coordinates and population) find the five strongest quakes occurred in a distance of less than 100Km from cities of population over one million during the 20th century.

OLAP Analysis

Additional to (naïve or advanced) database queries on detailed seismological data, a data warehouse approach utilizes online analytical processing (OLAP). We illustrate the benefits obtained by such an approach with two examples of operations supported by spatial data warehouse and OLAP technologies:

- A user may ask to view part of the historical seismic profile, that is, the ten most destructive quakes in the past twenty years, over Europe, and, moreover, he/she can easily view the same information over Greece (more detailed view, formally a *drill-down* operation) or worldwide (more summarized view, formally a *roll-up* operation).
- Given the existence of multiple thematic maps, perhaps one for quake magnitude and one for another, non-geophysical parameter such as the resulting damage, these maps could be overlaid for the exploration of possible relationships, such as finding regions of high, though non-destructive, seismicity and vice versa.

Further to roll-up and drill-down operations described above, typical data cube operations include *slice* and *dice*, for selecting parts of a data cube by imposing conditions on a single or multiple cube dimensions, respectively (Figure 5), and *pivot*, which provides the user with alternative presentations of the cube (Figure 6).

Another important issue in data warehousing is the physical representation of a cube. Relational OLAP (ROLAP) and multidimensional OLAP (MOLAP) are the two principal models proposed in the literature. ROLAP actually uses relational tables that a relational DBMS is designed to handle, while MOLAP makes use of specialized structures (multi-

Figure 5. Selecting parts of a cube by filtering a single (slice) or multiple dimensions (dice)

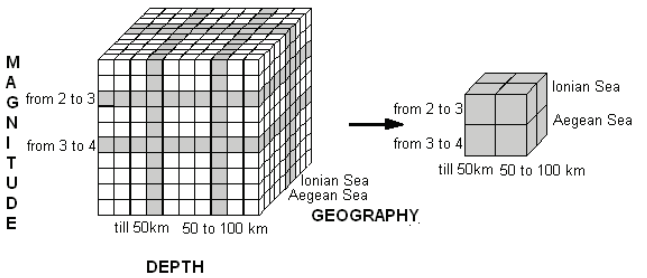
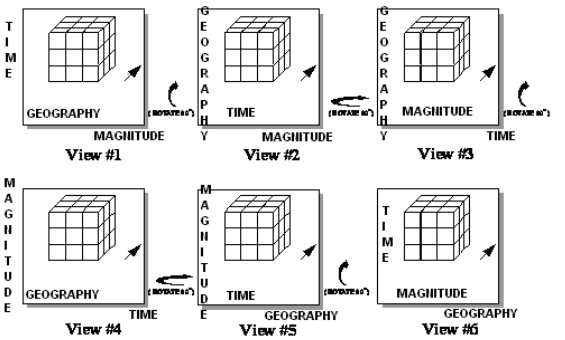


Figure 6. Alternative presentations: Views of a cube (pivot)



dimensional arrays) designed especially for OLAP purposes. The advantage of ROLAP is that it can handle large volumes of data (since relational DBMS

are perfect for this task). On the other hand, MOLAP is much faster for performing OLAP operations due to the extensive use of main memory structures.

For SDMMS purposes, where both requirements are there (large volume of data and fast OLAP operations), either ROLAP or MOLAP could be adopted for the implementation of the seismological data warehouse, or even a combination of the two (Hybrid OLAP – HOLAP), which has been recently supported by commercial DBMS, would be an alternative.

Data Mining

Integrating data analysis and mining techniques into an SDMMS ultimately aims to the discovery of interesting, implicit and previously unknown knowledge. The *knowledge discovery in databases* (KDD) process consists of the following steps, from the storage of interesting information in a data warehouse until the extraction, interpretation and understanding of useful, possibly hidden knowledge (Fayad et al., 1996; Han & Kamber, 2000):

1. Building a data warehouse from one or more raw databases (data warehouse building step)
2. Selecting and cleansing data warehouse contents to focus on target data (selection and cleansing step)

3. Transforming data to a format convenient for data mining (transformation step)
4. Extracting rules and patterns by using data mining techniques (data mining step)
5. Interpreting and evaluating data mining results to produce understandable and useful knowledge (interpretation and evaluation step)

Examples of useful patterns found through KDD process include clustering of information (e.g., shocks occurred closely in space and/or time), classification of phenomena with respect to area and epicenter, detecting phenomena semantics by using pattern finding techniques (e.g., characterizing the main shock and possible intensive aftershocks in shock sequences, measuring the similarity of shock sequences, according to a similarity measure specified by the domain expert, etc.). Recently, there have been proposals that expand the application of knowledge discovery methods on multi-dimensional data (Koperski & Han, 1995; Koperski et al., 1998).

Association Rule Mining

Association rule mining aims at discovering interesting correlations among database attributes (Agrawal et al., 1993). Association rules are implications of the form $A \Rightarrow B\,[s, c], A \subset J, B \subset J$ where A, B and J are sets of items (i.e., attributes), characterized by two measures: *support* (s) and *confidence* (c). The support of a rule $A \Rightarrow B$ expresses the probability that a database event contains both A and B, whereas the confidence of the rule expresses the conditional probability that a database event containing A also contains B.

As an example, an association rule on seismological data would be like the following (cf. discussion in SDMMS database subsection for attribute meanings):

location in $L \wedge$ *depth* $\geq$ 100 Km $\Rightarrow$ *magnitude* $\geq$ 5R [1%, 50%]

which is interpreted as follows: *whenever an earthquake occurs in location L at a depth of over 100 Km its magnitude is likely to be greater than 5R with a probability of 50%; this combination occurred in 1% of all recorded events*.

An interesting variation is that of temporal association rule mining (*sequencing*), which detects correlations between events with time as in the following example:

location in $L_1 \wedge$ *magnitude* $\geq$ 7R $\Rightarrow$ *location* in L_2 within [0, 30 days] [0.1%, 30%]

which is interpreted as follows: *whenever an earthquake occurs in location L_1 with a magnitude greater than 7R it is likely that another earthquake occurs in location L_2 within a month after the first event with a probability of 30%; this combination occurred in 0.1% of all recorded events*.

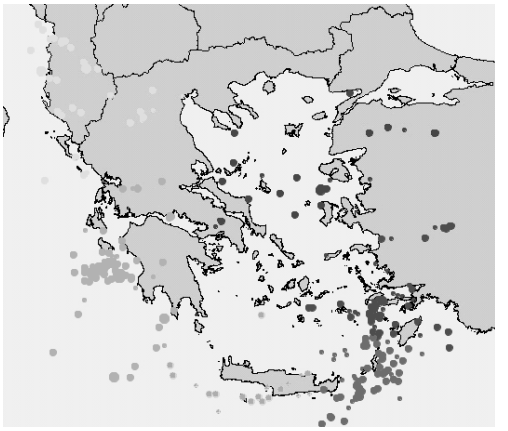
By identifying and analyzing event sequences (*seismic sequences*) seismologists can be assisted in studying this kind of earthquake behavior.

Clustering

Data clustering (Kaufman & Rousseeuw, 1990; Jain et al., 1999) is the unsupervised process of grouping together sets of objects into classes with respect to a similarity measure. Thus, it is the behavior of groups rather than that of individual events that is detected. Applications on seismic data could be for the purpose of finding densely populated regions (according to the Euclidean distance) between the epicenters, and, hence, locating regions of high seismic frequency or dividing the area of a country into zones according to seismicity criteria (e.g., low/medium/high seismic load) as illustrated in Figure 7.

Several clustering methods have been proposed in the literature. Using multi-dimensional correlations, local spatio-temporal clusters of low magnitude events can be extracted (Dzwinel et al., 2003). Also, correlations between the clusters and the earthquakes are recognized. Signal processing techniques can be applied to spatial data if they are considered as multidimensional signals (Sheikholeslami et al., 2000). A clustering approach based on wavelet

Figure 7. Discovering clusters of earthquake epicenters (Theodoridis, 2003)



transforms can identify clusters by finding dense regions in the transformed data. Finally, hybrid methodologies have been proposed (Guo et al., 2003) where spatial clustering is combined with high-dimensional clustering.

Data Classification

Classification is one of the most common supervised learning techniques. The objective of classification is to first analyze a (labeled) training set and, through this procedure, build a model for labeling new data entries (Han & Kamber, 2000). In particular, at the first step a classification model is built using a *training data set* consisting of database records that are known to belong in a certain class and a proper supervised learning method, e.g. decision trees or neural networks. In case of decision trees, for example, the model consists of a tree of "if" statements leading to a label denoting the class the record it belongs in. At the second step, the built model is used for the classification of records not included in the training set. Many methods have been developed for classification, including decision tree induction, neural networks and Bayesian networks (Fayad et al., 1996).

As an example, the (hypothetical) decision tree illustrated in Figure 8 tries to "predict" the macroseismic intensity at a site given the depth and the magnitude of an earthquake, the geographic area and the local geology. Such an implication uncovers correlations among the attributes of the seismological database and decision trees of such a type are already used by local authorities to prioritize actions for response and relief of the population after a strong earthquake.

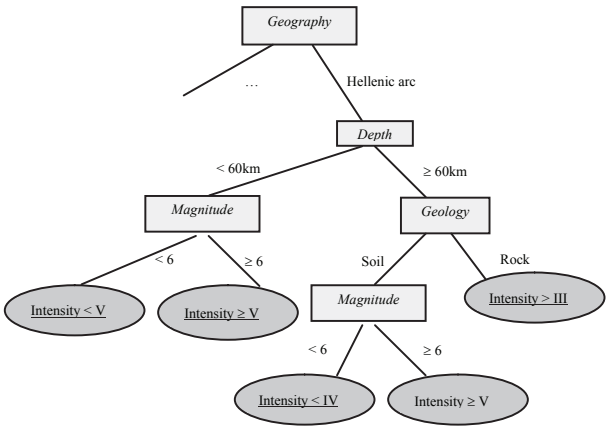
In this section, we presented querying, OLAP and data mining that could be used for extracting useful conclusions about seismological data stored in a SDMMS. These operations can be part of a system that manages seismological data in order to support the decision-making process.

SDMMS for Decision-Making Purposes

After having discussed the components of a SDMMS, we present alternative usage of such a system with respect to user profile:

- Citizens find a portal useful for getting information about past earthquakes and about protection against earthquakes.
- Geophysicists make data analysis for constructing and visualizing seismic profiles of certain regions.
- Public administration officers utilize information to improve emergency response

Figure 8. An example decision tree for seismological data



and make decisions about the structural rules.

A system with these characteristics can be characterized as a Decision Support System (DSS) that will provide users with aggregated information and, even more, useful, interpretable and easily understood knowledge. Examples of decision making through collecting and analyzing seismological data are the following:

- Public administration officers utilize information to improve emergency response and make decisions about the structural rules. For example, PEADAB is a related EU-funded project towards this direction (Gerbesioti et al., 2001; PEADAB, 2006).
- Seismological events can not be isolated from the movement of plates that cause faults, the volcano activities, the site effects and many others. Seismologists need an integrated environment in which all this information can be presented and analyzed. So they can reach conclusions by collecting and analyzing seismological data using SDMMS, which can automate this process and provide strong analytical tools. As an example, an integrated seismic network, called CISN, has been developed in California with ShakeMap and HAZUS being the two core tools of this network (Goltz & Eisner, 2003). ShakeMap is used to provide details about the earthquakes within five minutes after they happen. HAZUS, a methodology for earthquake, flood and wind hazards, generates estimates of population impacts in terms of deaths and injuries. Other provided estimates are damages to buildings, critical facilities and transportation lifelines.

A real challenge for the future could be the use of all the collected information to manage emergencies. Assume a DSS that could predict the level of destruction in urban areas and by taking under consideration the particular infrastructure (mass transport means, utility networks, locations of hospitals and schools, etc.) assist officers guide an emergency operation.

PROTOTYPE SYSTEMS AND TOOLS: A SURVEY

In this section, we present a number of prototype tools that have been proposed to collect, process and analyze seismological or, in general, spatial and earth science data. We also provide a short comparison from the perspective of SDMMS architecture and objectives.

Descartes/Kepler

Andrienko and Andrienko (1999) proposed an integrated environment (Descartes/Kepler) where data mining and visualization techniques are used to analyze spatial data. Their aim is to integrate traditional data mining tools with cartographic visualization tools so that the users can view both source data and results produced by the data mining process.

Descartes provides mapping and visualization features. Furthermore, it supports some data transformations effective for visual analysis, and the dynamic calculation of derived variables. On the other hand, Kepler incorporates a number of data mining methods. It is an open platform and through an interface new methods can be added. Kepler supports the whole Knowledge KDD process including data input and format transformation tools, access to databases, querying, management of (intermediate) results, and graphical presentations of various kinds of data mining results (trees, rules, and groups).

Figure 9 illustrates a composite screenshot of the tool with maps and charts visualization.

CommonGIS

CommonGIS (Kretschmer & Roccatagliata, 2000; CommonGIS, 2006) deals with geographical data and supports the visualization and analysis of statistical data that are related with spatial objects. The main features of CommonGIS are the following:

- Supports a variety of standard formats of map and table data
- Adopts a flexible client-server architecture that optimizes download time and supports integration of data from remote servers
- Combines interactive mapping techniques with statistical graphics displays and computation
- Includes comprehensive tools for analysis of spatial time-series
- Includes information visualization tools (dynamic query, table lens, parallel coordinate plots, etc.) dynamically linked to maps and graphics via highlighting, selection, and brushing
- Supports interactive multi-criteria decision making and sensitivity analysis
- Helps users to follow problem solving scenarios
- Applies multivariate graphics to the analysis of spatial data
- Displays spatio-temporal events and other kinds of multidimensional data
- Includes tools for interactive aggregation of grid data tightly coupled with dynamic visualization of aggregation results

Figure 9. Descartes/Kepler functionality (Andrienko & Andrienko, 1999)

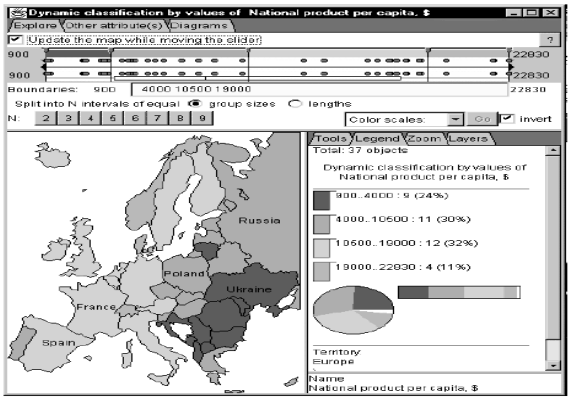


Figure 10. Visualization capabilities in CommonGIS (CommonGIS, 2006)

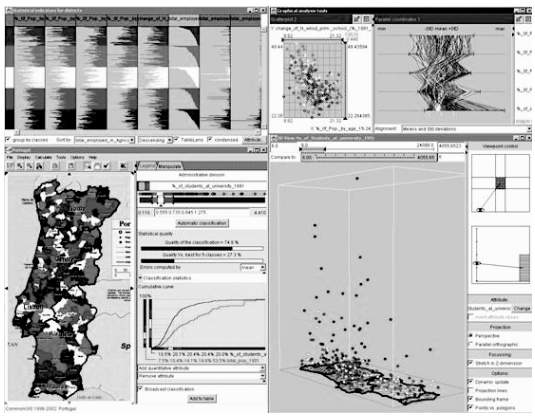


Figure 10 illustrates a collection of visualization results supported by CommonGIS.

GEODE

Geo-Data Explorer (GEODE) is an ambitious and highly promising application developed by the USGS for providing users with geographically referenced data. The project aims in developing a portal which will provide real-time data and will support data analysis independently from special hardware, software and training (Levine & Schultz, 2002). The main features of GEODE include: simultaneous display of many data formats, possibility of downloading specific parts of datasets, illustration of data in real-time, support of multiple scales, unlimited dataset size; maps customization and support of image export.

Figure 11 illustrates the functionality of GEODE through a representative screenshot.

Seismo-Surfer

Last but not least, Seismo-Surfer is a tool for collecting, querying, and mining seismological data following the SDMMS concept (Theodoridis, 2003; Kalogeras et al., 2004; Seismo-Surfer, 2006). Its database is automatically updated from remote sources; querying on different earthquake parameters is allowed, while data analysis for extracting useful information is limited to a data clustering algorithm. Querying and mining results are graphically presented via maps and charts.

Seismo-Surfer architecture, in general, follows the SDMMS architecture illustrated in Figure 1. A number of filters cleanse and homogenize the datasets (mainly concerning about duplicate entries), which are available from remote sources and pre-processed datasets are stored in the local database. In its current version, Seismo-Surfer supports links with two remote sources: one at a national level for Greece (GI-NOA, 2006) and one worldwide (NEIC-USGS). Users interact with the database via a graphical user interface. Querying and data mining results are presented in graphical mode (maps, charts, etc.). Querying on earthquake parameters includes variations of spatial queries, such as range, distance, nearest-neighbor and top-N queries (illustrated in Figure 12).

A Comparison of SDMMS Prototypes

Table 1 presents a comparison between the different prototypes that were presented in this section. According to this table, all support dynamic loading of up-to-date information from remote sources, querying and visualization facilities; OLAP functionality is not provided at all, whereas data mining techniques are included in Descartes/Kepler and Seismo-Surfer.

Figure 11. GEODE functionality (GEODE)

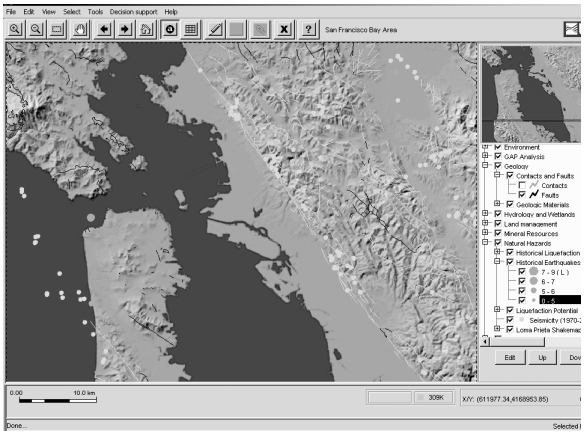
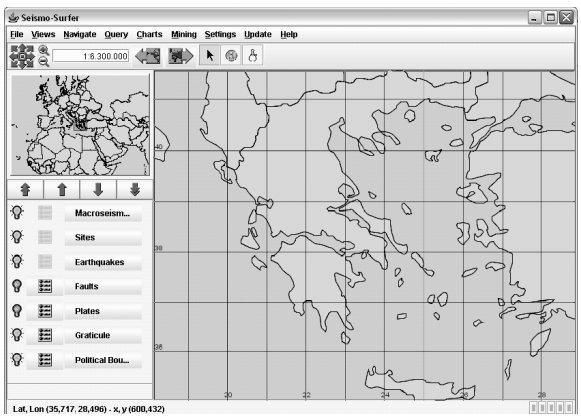


Figure 12. Querying capabilities of Seismo-Surfer (Seismo-Surfer, 2006)



CONCLUSION

In this article, we discussed the architecture of a so-called *seismic data management and mining system* (SDMMS) for quick and easy data collection, processing (generating historic profiles of specific geographic areas and time periods, providing the association of seismic data with other geophysical parameters of interest, etc.), and visualization supporting sophisticated user interaction.

The core components of this architecture include a seismological database (for querying) and a seismological data warehouse (for OLAP analysis and data mining). We provided template schemes for both components as well as examples of their functionality. Emphasis was put on the decision-making, since SDMMS could be used as a DSS by specialized earth scientists and public administration officers. We also provided a survey of existing operational or prototype systems following (at a low

Table 1. Comparing SDMMS prototypes

	Descartes/Kepler	**CommonGIS**	**GEODE**	**Seismo-Surfer**
Web interface	No	Yes	Yes	Yes
Dynamic information (through Internet or Map load/retrieval)	Both	Internet	Both	Both
Commercial / prototype systems exploited	Descartes Kepler	Descartes PGS map server (Lava/Magma) Vizard	Informix PGS map server (Lava/Magma)	Oracle OpenMap
OLAP functionality	No	No	No	No
Data pre-processing	Cartographic visualization (Descartes) ETL techniques (Kepler)	Data characterization scheme	Image compression (MR SID), data transform unit	Filters for data cleansing & integration
Data mining techniques	Clustering, Classification, Assoc. rules	None	None	Clustering
Visualization techniques	Maps, Charts	Maps, Charts	Maps	Maps, Charts
Query formulation (via Interface or Query language)	Interface	Interface	Interface	Interface

or high percentage) the proposed SDMMS functionality.

Interestingly, OLAP and data mining functionality vary from absent to quite limited. This is a hint for future work on the surveyed as well as new tools for seismological data management.

REFERENCES

Agarwal, S., Agrawal, R., Deshpande, P., Gupta, A., Naughton, J., Ramakrishnan, R., et al. (1996). On the computation of multidimensional aggregates. In *Proceedings of the 22nd International Conference on Very Large Databases, VLDB'96*, Bombay, India.

Agrawal, R., Imielinski, T., & Swami, A. (1993). Mining Association Rules between Sets of Items in Large Databases. In *Proceedings of ACM SIGMOD International Conference on Management of Data, SIGMOD'93* (pp. 207-216), Washington DC, USA.

Andrienko, G., & Andrienko N. (1999). Knowledge-based visualization to support spatial data mining. In *Proceedings of the 3rd Symposium on Intelligent Data Analysis, IDA'99*, Amsterdam, The Netherlands.

Barry, R., Cavers, D., & Kneale, C. (1975). Recommended standards for digital tape formats. *Geophysics, 40*, 344-352.

Behnke, J., & Dobinson, E. (2000). NASA workshop on issues in the application of data mining to scientific data. *ACM SIGKDD Explorations Newsletter, 2*(1), 70-79.

CommonGIS. (2006). *GIS for everyone…everywhere!*. Retrieved from http://commongis.jrc.it/index.html

Dzwinel, W., Yuen, D., Kaneko, Y., Boryczko, K., & Ben-Zion, Y. (2003). Multi-resolution clustering analysis and 3-D visualization of multitudinous synthetic earthquakes. *Visual Geosciences, 8*(1), 12-25.

Fayad, U., Piatetsky-Shapiro, G., Smith, P., & Uthurusami, R. (1996). *Advances in Knowledge Discovery and Data Mining*. MIT Press.

Gerbesioti, A., Delis, V., Theodoridis, Y., & Anagnostopoulos, S. (2001). Developing decision support tools for confronting seismic hazards. In *Proceedings of the 8th Panhellenic Conference in Informatics, PCI'01*, Nicosia, Cyprus.

GI-NOA. (2006). *Earthquake catalog*. Retrieved from http://www.gein.noa.gr/services/cat.html

Goltz, J., & Eisner R. (2003). Real-time emergency management decision support: The California integrated seismic network (CISN). In *Proceedings of the Disaster Resistant California 2003 Conference, DRC'03*, San Jose, CA, USA.

Guo, D., Peuquet D., & Gahegan, M. (2003). ICEAGE. Interactive clustering and exploration of large and high-dimensional geodata. *GeoInformatica, 7*(3), 229-253.

Han, J., & Kamber, M. (2000). *Data mining: Concepts and techniques.* Morgan Kaufmann.

Inmon, W. (1996). *Building the data warehouse,* 2nd ed. John Wiley & Sons.

Jain, A., Murty, M., & Flynn, P. (1999). Data clustering: A review. *ACM Computing Surveys, 31*(3), 264-323.

Kalogeras, I., Marketos, G., & Theodoridis, Y. (2004). A tool for collecting, querying, and mining macroseismic data. *Bulletin of the Geological Society of Greece, vol. XXXVI*.

Kaufman, L. & Rousseeuw, P. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley & Sons.

Kiratzi, A., & Louvari, E. (2003). Focal mechanisms of shallow earthquakes in the aegean sea and the surrounding lands determined by waveform modeling: A new database. *Journal of Geodynamics, 36*, 251-274.

Koperski K., & Han J. (1995). Discovery of spatial association rules in geographic information databases. In *Proceedings of the 4th International Symposium on Large in Spatial Databases, SSD'95*, Portland, MA, USA.

Koperski, K., Han, J., & Adhikary, J. (1998). Mining knowledge in geographical data. *Communications of the ACM, 26*(1), 65-74.

Kretschmer, U., & Roccatagliata, E. (2000). CommonGIS: A European Project for an Easy Access to Geo-data. In *Proceedings of the 2nd European GIS Education Seminar, EUGISES'00*, Budapest, Hungary.

Levine, M., & Schultz, A. (2002). *GEODE (Geo-Data Explorer)—A U.S. geological survey application for data retrieval, display, and analysis through the Internet*. U.S. Geological Survey, Fact Sheet 132-01, Online Version 1.0. Retrieved from http://pubs.usgs.gov/fs/fs132-01/

NEIC-USGS. *Earthquake search*. Retrieved from http://neic.usgs.gov/neis/epic/epic_global.html

PEADAB. (2006). *Post-earthquake assessment of building safety*. Retrieved from http://europa.eu.int/comm/environment/ civil/prote/cpactiv/cpact08a.htm

Pfoser, D., & Tryfona, N. (1998). Requirements, definitions and notations for spatiotemporal application environments. In *Proceedings of the 6th International Symposium on Advances in Geographic Information Systems, ACM-GIS'98*, (pp. 124-130). Washington DC, USA.

PPDM. (2006). *The data exchange project*. Retrieved from http://www.ppdm.org/standards/exchange/index.html

SEG. (2006). The Society of Exploration Geophysicists. http://www.seg.org

Seismo-Surfer. (2006). Seismo-Surfer Project. http://www.seismo.gr

Sheikholeslami, G., Chatterjee, S., & Zhang, A. (2000). WaveCluster: A Wavelet-based Clustering Approach for Spatial Data in Very Large Databases. *The VLDB Journal, 8*(3-4), 289-304.

Stefanovic, N., Han, J., & Koperski, K. (2000). Object-based selective materialization for efficient implementation of spatial data cubes. *IEEE Transactions on Knowledge and Data Engineering, 12*(6), 938-958.

Theodoridis, Y. (2003). Seismo-surfer: A prototype for collecting, querying and mining seismic data. In *Advances in Informatics—Post Proceedings of the 8th Panhellenic Conference in Informatics* (pp. 159-171*)*. Berlin: Springer Verlag.

Young, J., Presgrave, B., Aichele, H., Wiens, D., & Flinn, E. (1996). The Flinn-Engdahl Regionaligation

Scheme: The 1995 Revision. *Physics of the Earth and Planetary Interiors, 96*, 223-297.

Yu, B. (2005). Mining earth science data for geophysical structure: A case study in cloud detection. In *Proceedings of 2005 SIAM International Conference on Data Mining, SIAM'05*, Newport Beach, CA, USA.

Gerasimos Marketos is a PhD candidate at the Department of Informatics, University of Piraeus (UniPi), Greece. Born in 1981, he received his Bachelor of Science degree (2003) in informatics from University of Piraeus and his Master of Science degree (2004) in information systems engineering from University of Manchester Institute of Science and Technology (UMIST), UK. His research interests include spatiotemporal data warehousing and mining, pattern management and scientific databases. He is member of BCS. [URL: http://isl.cs.unipi.gr/db/people/marketos/]

Dr. Yannis Theodoridis is assistant professor with the Department of Informatics, University of Piraeus (UniPi). Born in 1967, he received his Diploma (1990) and PhD (1996) in electrical and computer engineering, both from the National Technical University of Athens, Greece. His research interests include spatial and spatiotemporal databases, geographical information management, knowledge discovery and data mining. Currently, he is scientist in charge for UniPi in the EC-funded GeoPKDD project (2005-08) on geographic privacy-aware knowledge discovery and delivery, also involved in several national-level projects. He has co-authored three monographs and over 50 articles in scientific journals (including Algorithmica, ACM Multimedia and IEEE TKDE) and conferences (including ACM SIGMOD, PODS, VLDB and ICDE) with over 400 citations in his work. He participates in the steering committee for the Int'l Symposium on Spatial and Temporal Databases (SSTD) and in the editorial board for the Int'l Journal on Data Warehousing and Mining. He is member of ACM and IEEE. [URL: http://www.unipi.gr/faculty/ytheod/]

Dr. Ioannis Kalogeras is senior researcher in the Institute of Geodynamics (IG), National Observatory of Athens. He was born in Athens in 1956 and he received his Diploma (1981) in Geology and PhD (1993) in Seismology both from the Athens University, Greece. In 2006 he received an MSc in geoinformatics from the National Technical University of Athens, Greece. His research interests include the strong ground motion study, the development of seismological databases, the seismic tomography and the development of strong motion networks. Since 1986 he is in charge of the strong motion networks of the IG. Recently he updated the macroseismic observation collection system of the IG. He is involved in various national and international research projects as scientific leader or as senior researcher. He has co-authored over 40 articles published in scientific journals (Natural Hazards, BSSA, SRL) and congress proceedings (EGS, IASPEI). He is member of SSA and AGU. [URL: http://www.gein.noa.gr/Kalogeras/cveng.html]