Data Analysis of Spatio-Temporal Sensor Data as a Contribution to the Model Analysis for Water Resources

Sanja Veleva[1], Kosta Mitreski [2]
University SS Cyril and Methodius, Faculty of Electrical Engineering and Information Technologies
Skopje, MACEDONIA

Abstract

The quality of the information is measured by its accuracy and its relevance over time. Therefore, the process of data analysis of the sensor eco-data is of a great importance to the detection and prediction of the eco-hydrology phenomena. The existing models for data mining do not relate to the continuously changing characteristics of the sensor eco-data. Furthermore, most of the monitoring systems are based on event alert services, which do not answer to the continuous variations of the measured parameters. Our approach embeds the nature of system characteristics into one dynamic model for data mining of continuously changing spatio-temporal characteristics of one eco-hydrology system. The continuously gathered sensor eco-data from the region of Lake Prespa consisted of 320 water samples, among them 224 from the lake gauging stations and 96 from the river gauging stations. Considering the recommendations from the Water Framework Directive (WFD), the sensor eco-data were grouped into three types: physical, chemical and biological, corresponding to their aspect of water quality. All of these types convey the same class definition in the form of value, spatial and temporal information. To define our sensor data mining model we contribute to three segments: outlier analysis, pattern analysis, and prediction analysis. The suggested sensor data analysis model should be of a useful asset in obtaining knowledge for certain aquatic phenomena.

Keywords: sensor eco-data, model analysis, eco-hydrology, Water Framework Directive, Lake Prespa

Introduction

The significance of the sensor eco-data is more and more perceptible considering the continuously varying conditions in the today's climate. Nowadays, the presence of the sensor eco-data is evident and even increasingly noticeable in different areas of application. Therefore, it is of crucial importance to obtain fast and effective extraction of the information of a certain eco-phenomena, but also with acceptable reliability.

The continuously gathered sensor eco-data from the region of Lake Prespa consisted of 320 water samples, among them 224 from the lake gauging stations and 96 from the river gauging stations. The eco-data was organized in the integrated database system, in order to provide a convenient, easy-to-use, and an intuitive way of storing the captured data. The integrated database system represents a centralized storage facility which enables the users to better organize, control, manage and use the data, create reports, perform statistical analysis, establish patterns in the model of the data, etc. Considering the recommendations from the Water Framework Directive (WFD), the sensor eco-data were grouped into three types: physical, chemical and biological, corresponding to their aspect of water quality. For the needs of this paper, we introduce the process of data mining as a part of the data analysis of eco-hydrology phenomena. The defined types of eco-data are characterized by the same class definition in the form of value, spatial and temporal information. To define our sensor data mining model we contribute to the three segments of data analysis: outlier analysis, pattern analysis, and prediction analysis. Our approach embeds the nature of system characteristics into one dynamic model for data mining of continuously changing spatio-temporal characteristics of one eco-hydrology system. All of these eco-data analysis potentially lead toward an integrated and sustainable model of ecological and environmental structure of a certain water body.

Sensor eco-data

The researchers form Albania, Greece and Macedonia in the framework of FP6 Project TRABOREMA collected water samples of the Lake Prespa from several specific designated points and areas of the lake and its rivers. The continuously gathered sensor eco-data from the region of Lake Prespa consisted of 320 water samples, among them 224 from the lake gauging stations and 96 from the river gauging stations. Using appropriate methods the teams extracted data values for the physical, chemical and biological parameters of the lake. The values for the parametars of the measured points or the points were the sampling has been carried out were presented on maps with the following annotations LX and RX, where:

- L represents measured point which is located on the surface of the Lake Prespa
- R represents measured point which is located on the surface of the river, and
- X is a number-indicator of the measured point.

In the following map different layers present the total phosphor [$\mu g \cdot dm^3$] at the locations of the measuring points.

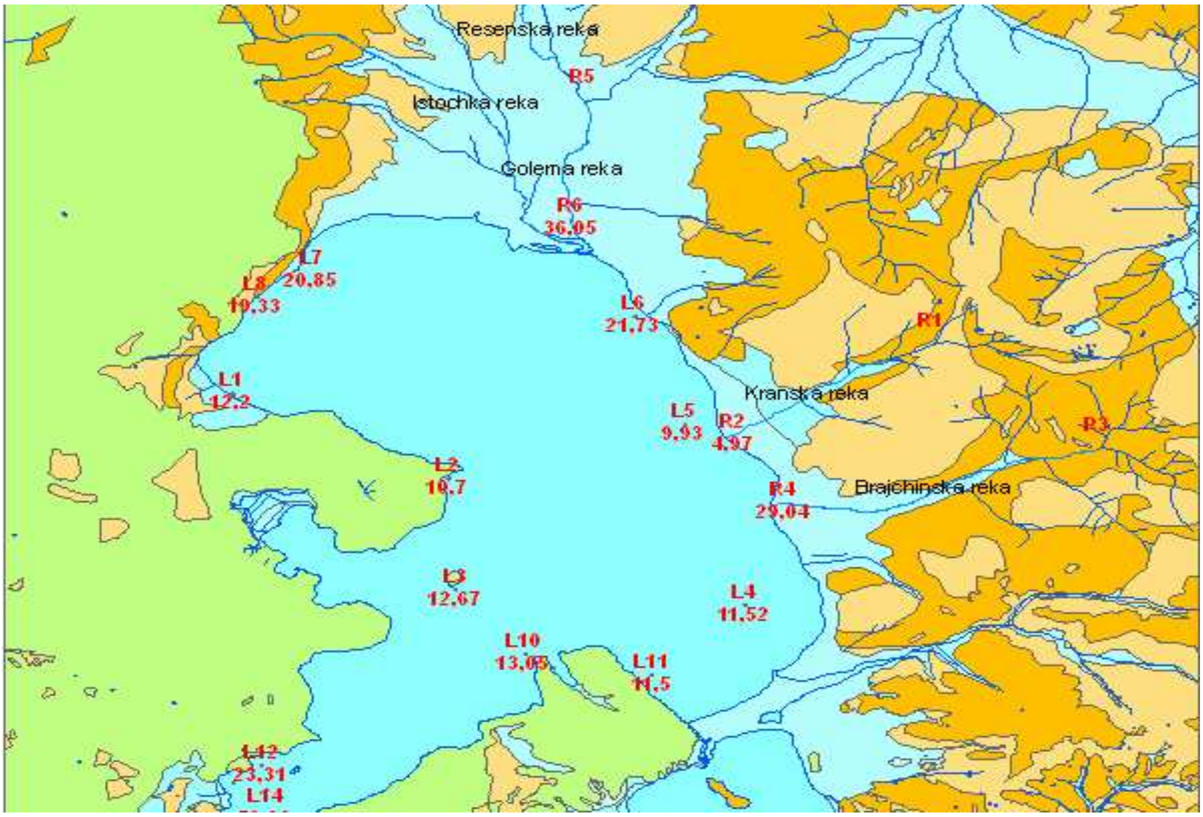

Figure 1. Measured points – total phosphor [$\mu g \cdot dm^3$] at the location of measured points

Categorization of sensor eco-data

Each team monitored these characteristics of the gauging points over a period of time and stored the values captured for further use. Corresponding to the aspect of water quality, and taking in consideration the recommendations from the Water Framework Directive (WFD), the sensor eco-data were categorized into three types: physical, chemical and biological.

Attributes for physical sensor eco-data

The physical data gives an overview of the microclimate of the lake such as the temperature changes throughout the year, visibility of the water, pH factor, etc. These parameters are measured at the sampling locations and entered directly into the database.

Attributes for chemical sensor eco-data

The chemical data shows in more detail the chemical properties of the lake itself and chemical changes caused by the human interactions i.e. changes in the chemical characteristics of the lake originated by the human activities in the region surrounding the lake. These activities include farming, domestic animals keeping, producing hard waste and waste waters, etc. Regular monitoring and analysis of the chemical data is very important both for the people living in the settlements surrounding the lake and for the flora and fauna of the lake. Using appropriate measurement methods, the data values are gathered from the collected samples and entered directly in to the database. The chemical data shows in more detail the chemical properties of the lake itself and chemical changes caused by the human interactions i.e. changes in the chemical characteristics of the lake originated by the human activities in the region surrounding the lake. These activities include farming, domestic animals keeping, producing hard waste and waste waters, etc. Regular monitoring and analysis of the chemical data is very important both for the people living in the settlements surrounding the lake and for the flora and fauna of the lake. Using appropriate measurement methods, the data values are gathered from the collected samples and entered directly in to the database.

Attributes for biological sensor eco-data

The biological data gives an inside view of the characteristics of the living world in the lake, its changes and fluctuations over the year. The biological data is a good indicator of the changes that are going on in the lake, caused both by the environmental influences and the periodical variations of the natural lifecycle. The biological data was first analyzed before entering it in the database. First, a selection is made of all the life form found in the specific water sample taken at a specific location and specific period. Only the life forms that dominate in that period are entered in the database. The dominant life forms were selected by the following criteria: if the specific life form is presented with more than 10% of all life forms in that period, than that is considered to be a dominant life form. For samples where no life form is presented with more than 10%, than the highest 3-4 are taken into account. Besides the list of the dominant life forms in a specific period at a specific location, diagrams of life form fluctuation are made. Based on the previously measured data, the team created bar charts, which give a statistical distribution of life form at a sampling location throughout one year cycle. Also, the team created pie-charts with average data of all measurements in the selected period. The data of the dominant life forms is entered in the database and the trophic and saprobity index charts are attached with the corresponding documents.

Figures 2 and 3 show an example of the representation data values for the trophic index and saprobity index for the same gauging point.

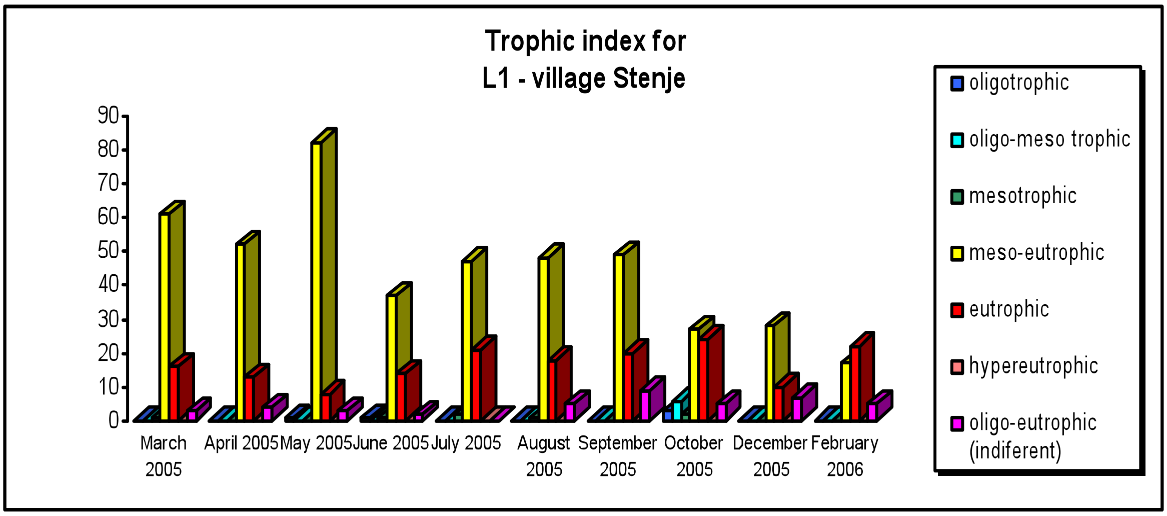

Figure 2. Example of the Trophic index for location L1 – village Stenje

The results in these figures are presented in a form of temporal data analysis for a period of one year. Depending on a certain pattern of behavior of the values, the interval of observation and research can be extended or fragmented into smaller intervals.

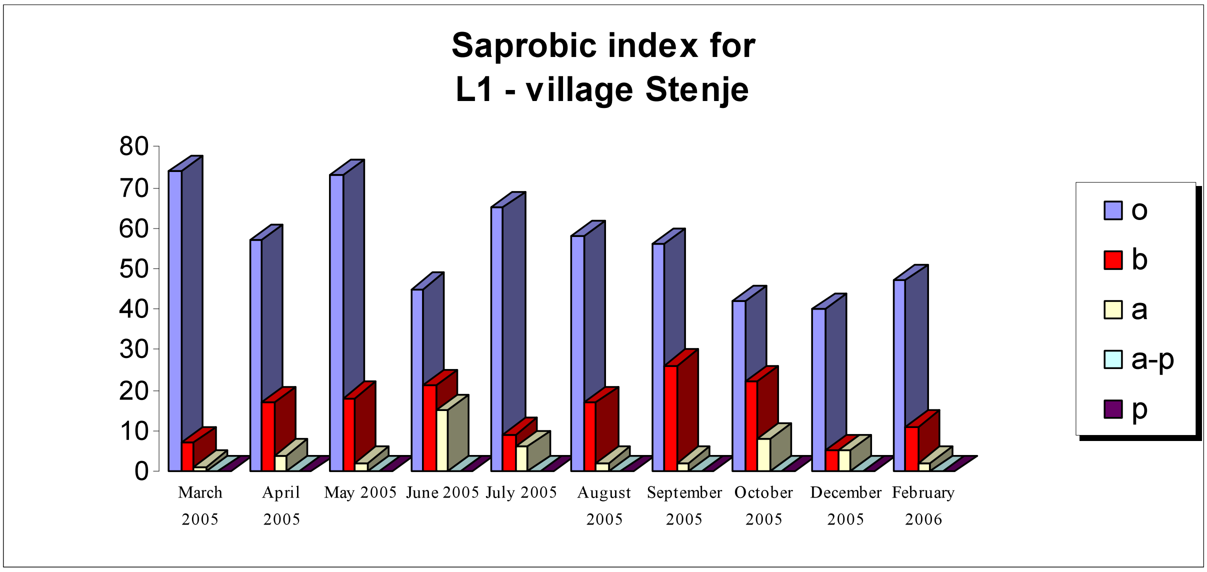

Figure 3. Example of the Saprobic index for location L1 – village Stenje

In Figure 4 and Figure 5, graphs represent a different approach of temporal analysis. The following are the sample diagrams for statistical analysis of the temperature and PH values during the analyzed period of one year.

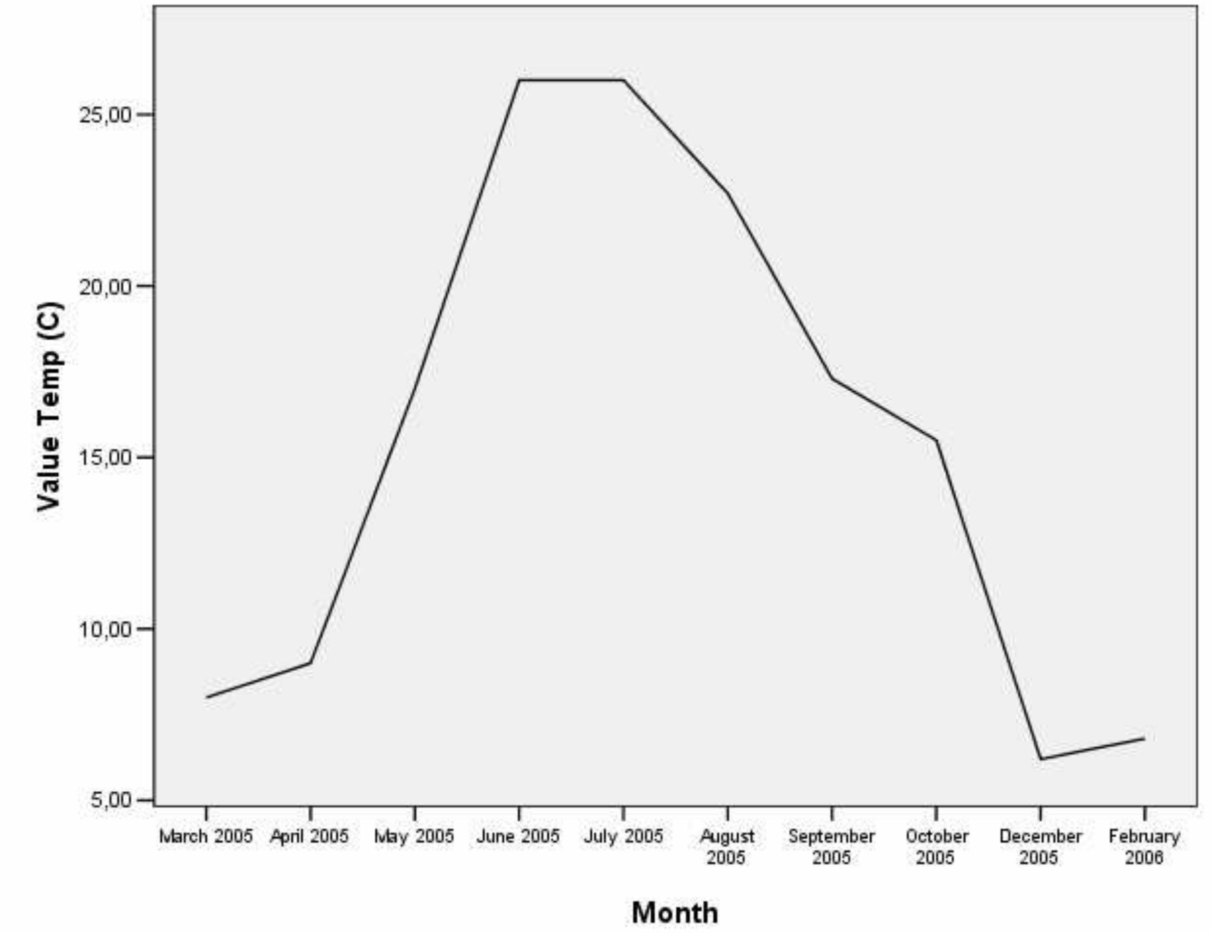

Figure 4. Diagram for temperature values during the measurement period

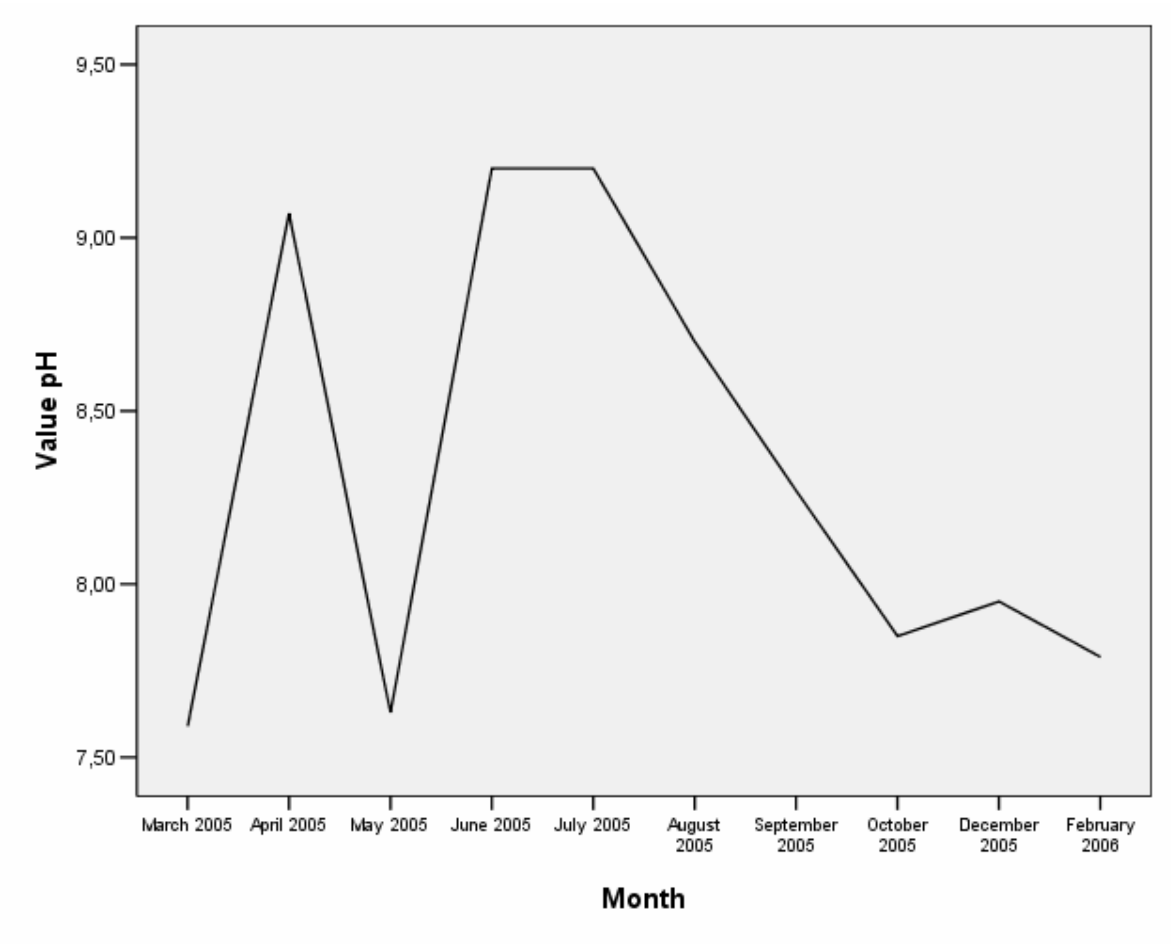

Figure 5. Diagram for pH values during the measurement period

Data storage

The values for the attributes of the water samples collected from Lake Prespa are stored in the designed integrated database system (sensor database - sensor DB). With the help of the centralized database system the teams are able to record these data values directly into the database system and to track easily the history of changes of these parameters. Another advantage is that, the teams have easily accessible information about the data values of the water samples of the other teams, not only their own. This helps them compare, monitor and evaluate the data collected on different locations and areas and collected at different point of time.

The software application for the database system was developed on Lotus Domino platform. Applications developed in Lotus Domino platform consist of one or more so called Notes databases. A Notes database is a unique combination of a storage facility (database) and a customable user interface for working with the data values (application).

Notes databases store information in so called documents and therefore, they are document oriented databases. Another type of databases is relational databases, which store data in tables. The advantage of the document oriented database is that it provides more intuitive, common and effective data representation. On the other hand, the data contained in the Lotus Domino database can be easily represented as a relational database. This can be done by using an adapter (driver) that will transfer the structure of the data so that the user will perceive the database as a relational one that can be represented by the UML methodology.

A Notes database is a single file that contains multiple documents. In addition, some databases can be customized to be used as "discussion databases," where users can discuss and post responses to particular topics. Documents in Notes databases contain structured text, rich text, pictures, objects, and many other types of information.

Databases may be stored on one or more Domino servers, accessible by many users. For the purpose of this paper we used shared and centralized database.

Data mining of sensor eco-data

Data mining as a process has already been received as an effective way to reduce time in the time consuming and memory consuming processes of obtaining knowledge from a given set of attribute values. In our paper, data mining is treated as an algorithmic process that has a sensor eco-data as an input, and as a result generates patterns for future prediction of hydrologic phenomena. The fundamental concepts that we use for the sensor data mining model are: sensor class, time interval for sampling, and threshold value. The sensor class is used to determine the sensor type by its location and sensing type. The time interval can have discrete sampling values or continuous interval values. The threshold value is given for the narrowing of the interesting values from the total set of values.

In order to develop an effective sensor data mining model we subsequently used the three basic data analysis:

- outlier extraction,
- pattern generation, and
- prediction analysis.

Outlier extraction

In the set of the measured values, certain values outstand from the expected interval around the average value. The process of the outlier extraction segregates these false values from the ones that convey the standard by creating the rules for classification of sensor eco-data. The outlier extraction can be done by mainly two criteria: the threshold value or the probability of outstanding from the set of average value.

Figure 6 illustrates the process of outlier extraction starting from the basic measurements up to the generation of filtered values ready for more effective further analysis. The sensor eco-data organized in the sensor class are selected from the sensor database. In the process of data selection, the additional classification criterion is extracted from the legacy database. The spatial information identifies the location of the sensor that is treated in this selection procedure. The selected sensor data, the threshold value and the rule for classification create the outlier set that is written in the knowledge database, so that can be used whenever a classification is needed with a same classification rule. After the successive generation of outliers, the filtered sensor classes go through a process of generation of pattern summaries.

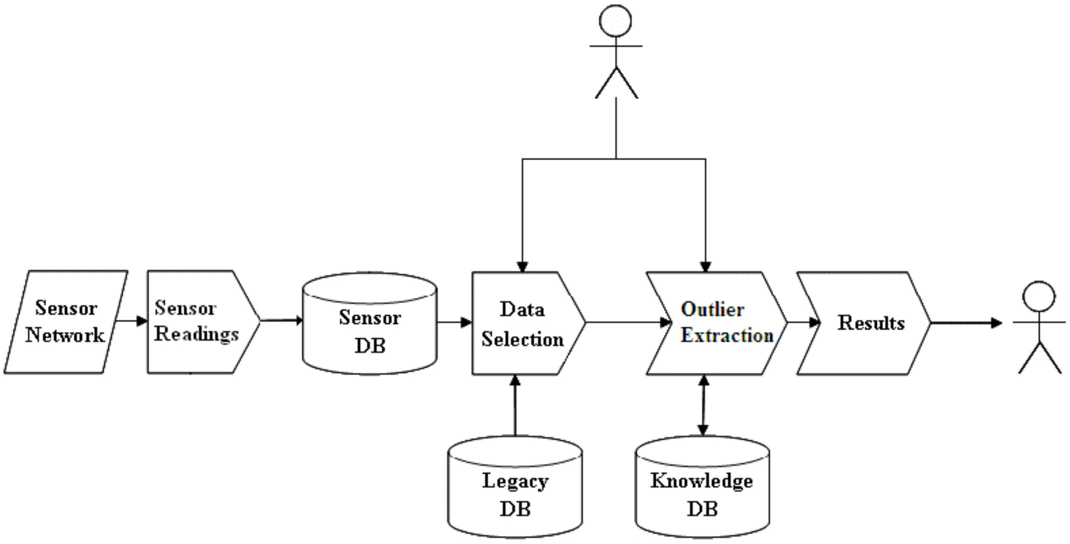

Figure 6. Diagram for outlier extraction

Pattern generation

Pattern generation is a process for determining patterns in the set of filtered sensor classes from the sensor database. The number of rules for pattern generation is very big. The decision for selecting the proper rule for particular domain and/or requirements is very difficult to make. Among the criteria is the kind of rule that we expect to be generated from the data mining procedure. If the result is a classification or an association rule, then the rules comply with these three basic criteria: objective, subjective and semantic measures. If a summary is expected as an output, then the rules are either objective or subjective measures. Depending on a nature of the eco-phenomena, the patterns can be analyzed as temporal, spatial or spatio-temporal. The process of spatial pattern generation and temporal pattern generation are illustrated in figure 7 and figure 8 accordingly.

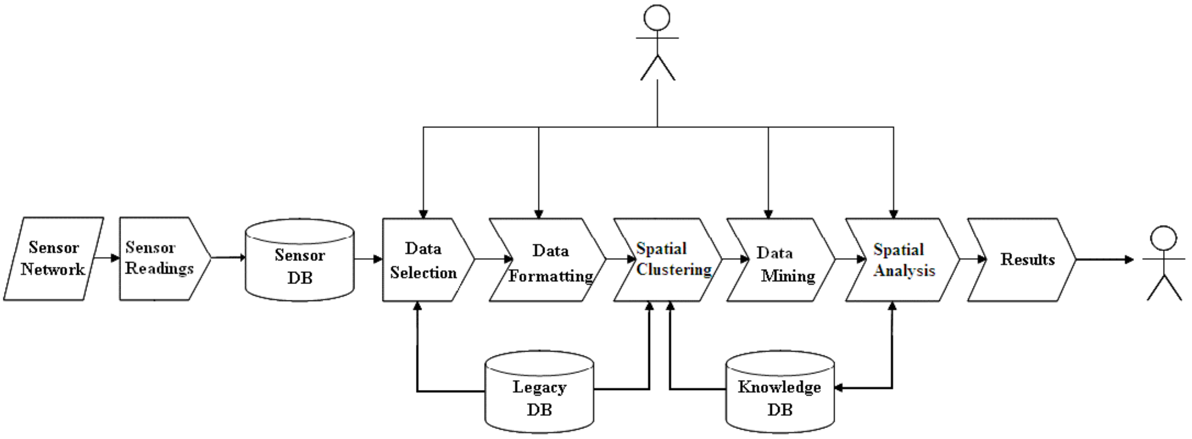

Figure 7. Diagram for spatial pattern generation

For the purpose of spatial correlation between the eco-data, the spatial pattern generation is implemented. The patterns are generated on the criteria of spatially closeness of the neighboring sensor data, density of the sensor data and similar behaviors of the neighboring sensor data. In the process of generating spatial patterns, after the formatting procedure, the values for the attributes of the sensor data are clustered by the rule of nearness of location. The mining procedure is more efficient when applied on the already spatially clustered data. As a result, the spatial patterns are generated based on the frequency of the appearance of the clusters.

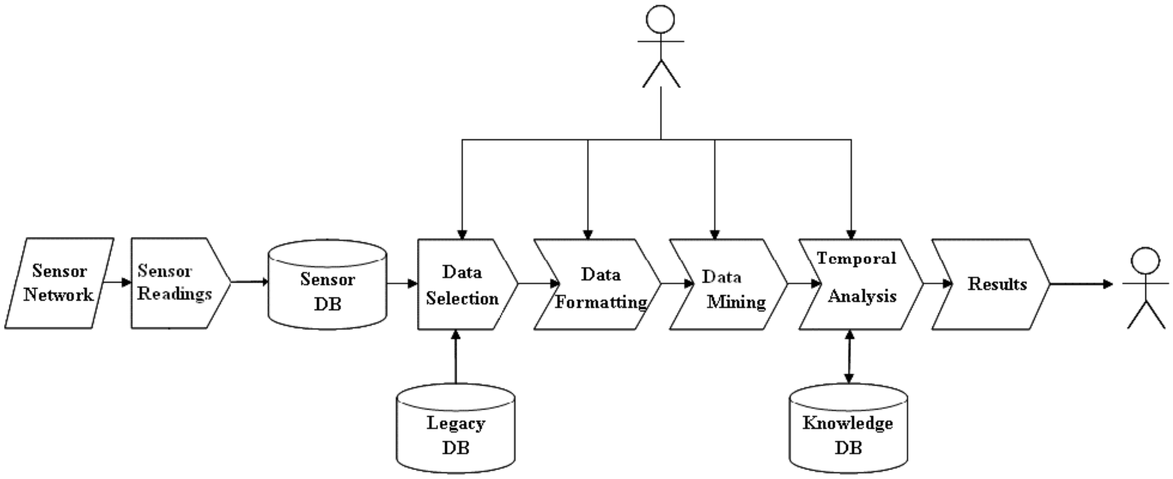

Figure 8. Diagram for temporal pattern generation

Temporal pattern generation is used for generating patterns that are recurring, sequential and that have the characteristics to be typical for a certain period. In comparison to the generation of spatial patterns, the generation of temporal patterns, lacks the procedure of clustering the sensor eco-data.

Prediction analysis

Prediction analysis is a process that makes a possible prediction of the future values of the eco-data based on the similarity of the previously generated patterns. It follows the procedure of generating patterns, and therefore if the patterns are carefully and correctly chosen, than the prediction analysis will be easily obtained. Figure 9 illustrates the process of prediction analysis.

The firstly generated prediction scenario is stored in the knowledge database as a reference for testing its validity by comparing it with the following prediction scenario.

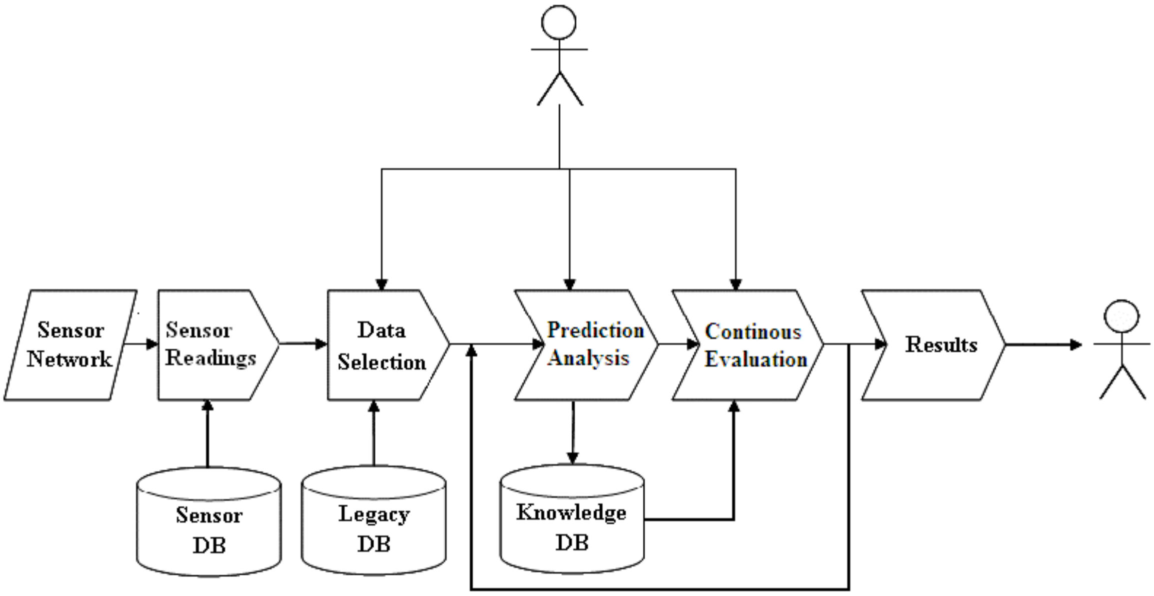

Figure 9. Diagram for prediction analysis

The firstly generated prediction scenario is stored in the knowledge database as a reference for testing its validity by comparing it with the following prediction scenario. If the generated scenario is not as correct as it should be according to the exceeding interval of the error value, than the prediction analysis improves the prediction by accepting the new and deleting the past and not precise scenario. The procedure of continuous evaluation, as a part of the prediction analysis has a positive effect because of the constant improvement in the accuracy of the generated scenario.

Conclusions

The goal of this paper was to present the developed hydrological monitoring system for spatio-temporal analysis for sensor eco-parameters. In order to provide better understanding of the natural phenomena of the aquatic bodies a monitoring system for the transboundary region of Lake Prespa has been developed. The development of the monitoring system for the Prespa region was analyzed through its three phases: data acquisition level, data handling and data management.

As a part of the data acquisition phase, teams from three neighboring countries Albania, Greece and Macedonia collected water samples of the Lake Prespa from several specific designated points and areas of the lake and its rivers. After the sampling was done with the unified methods, the extracted data values were grouped according to their physical, chemical and biological characteristics, following the guidelines from the deliverables of the Water Framework Directive (2000/60/EC).

The data handling was done with the help of the centralized database system, by which the teams are now able to record these data values directly into the database system and to track easily the history of changes for these parameters. Furthermore, the integrated database system provides the teams to have easily accessible information about the data values of the water samples of the other teams, not only their own. As a result of developed database teams can compare, monitor and evaluate the data collected on different locations and regions, at different point of time.

Due to the phase of data management, the integrating role of the monitoring system gives the project another dimension, a strong link between the research teams, a direct way of communication and collaboration, and can also serve as a solid and common standing ground for further projects and initiatives for preserving the environment in the Prespa region, and for building lasting partnership between the involved teams. This type of network collaboration strengthens the relationship between the research teams, raises the level of cooperation and overcomes the barriers of the international borders for the cause of the sustaining the Lake Prespa's ecological and environmental structure. Through this multilateral collaboration, each team is now provided with a complete set of results for further processing, analysis, reporting and monitoring.

From the global perspective, the developed DB model for monitoring Lake Prespa allows great geographical dispersion of partners in the project. Moreover, the DB model aims to encourage consistency in data structures to facilitate data sharing, which again contributes to the deliverables of the Water Framework Directive (2000/60/EC).

Biodiversity, physic-chemical data of the Lake Prespa and its catchments, including groundwater (BOD, COD, etc.) obtained from the monitoring system were subjected to various data analyses, systematization, analytical and statistical tools, hydro-geological analysis, etc. for obtaining the unified and reliable data base. Data collection procedures and standards (protocols etc.) were defined according to WFD for use in the project to ensure data collected by the various team members is compatible.

The created database organizes the data in so called documents. Instead of using tables, entering the data was performed using forms and already created templates. On the other hand, the data contained in the database can be easily represented as a relational database. This can be done by using an adapter (driver) that will transfer the structure of the data so that the user will perceive the database as a relational one that can be represented by the UML methodology.

Furthermore, with the spatio-temporal analysis for sensor data we managed to improve the monitoring system by establishing past and present baseline conditions in order to confirm the problem, by providing a reference against which progress can be assessed, by identifying significant information gaps, and finally by developing a cost-effective monitoring program in order to quantify the characteristics and patterns of transboundary water resources such as Lake Prespa and its associated river basins.

References

Arctur, D., and Zailer, M.,*. Designing geodatabases: Case studies in GIS data modelling, Redlands, Calif.: ESRI Press, 2004.*

Hluchy, L.; Habala, O.; Ciglan, M.; Tran, V.D.: *"Mining and Integration of Environmental Data", IEEE International Conference on Computational Cybernetics, ICCC 2008, 27-29 Nov. pp. 247 – 252, 2008.*

Jorgensen, S.E., Chon T.-S., and Rechnagel, F.: *„Handbook of Ecological Modelling and Informatics", WIT. Southampton. 2009.*

TRABOREMA Project No. INCO-CT-2004-509177: *"Concepts for integrated trans boundary water management and sustainable socioeconomic development in the cross border region of Albania, Former Yugoslav Republic of Macedonia (FYROM) and Greece".*

Veleva Sanja, Kosta Mitreski, Danco Davcev: *"Geo-spatial analysis for prediction of river floods", ICT Innovations 2009, 28–29 September, Ohrid, Republic of Macedonia, 2009.*

Veleva Sanja, Kosta Mitreski, Danco Davcev: *"Spatial modelling of the river surface water flow as a tool for managing the sustainability of rivers", International Society for Ecological Modelling ISEM 2009 Conference "Ecological Modelling for Enhanced Sustainability in Management", 6 - 9 October, Quebec City, P.Q., Canada, 2009.*

Walter. W., Piegorsch A., Bailer. J.: *"Analyzing Environmental Data", John Wiley & Sons, Ltd ISBN: 0-470-84836-7 (HB), 2005.*