Dynamic Information Model for Oceanographic Data Representation

Natalia A. Zhukova[1], Dmitry I. Ignatov[2], Oksana V. Smirnova[1]

[1]Institute for Informatics and Automation of the Russian Academy of Sciences, St. Petersburg, Russia
nazhukova@mail.ru, sov@oogis.ru
[2]National Research University Higher School of Economics, Moscow, Russia

Abstract. This paper treats issues of dynamic information model for oceanographic data representation construction. Proposed model includes three submodels – statistical model of data description, logical model of data relation description, model description of processes of change of water environment parameters state. As well statistical and intellectual methods used for automation of data processing and analysis are presented. Use of these methods will allow reducing processing time, to provide possibility of adaptive dynamic data processing generation, to improve processing which assumes data handling not at the level of measured values, but at the level of knowledge about measurements, parameters, and their relationship and also knowledge about subject domain.

Keywords. dynamic information model, intelligent processing of oceanographic data, geoscience, data processing.

1 Introduction

At the present time interest to problems relating to research of environment conditions significantly increased. It is, first of all, due to changes in the atmosphere, ocean and earth's surface caused by different factors. Secondly, methods of data processing and analysis, that were developed, are oriented on use by subject domain experts. Generally data processing and analysis are performed by hand using special tools. Today three are three main problems – first, the low speed and quality of newly received data acquirement, secondly, complexity and low speed of data processing in delayed mode, thirdly, complexity of the task solution of forecasting water environment state. At the stage of operative data processing preliminary estimation of data quality is performed. Quality rating is held with the use of test set specialized for different data sources and regions and takes about a day.

The most difficult operations are operations of analysis in the delayed mode. The procedure of delayed data processing provides removal of noise and outliers, that don't differ much from measurements, and restoring of missing values, calculation of offsets, exposure of trends, comparison with statistical data for detection of data correctness.

Experts have to analyze in details data when performing processing in the delayed mode taking into account all earlier received data on the area of interest, data received using intended and similar data sources, and also knowledge of physical features of the environment of the studied region. Complexity of problems of the delayed processing constantly increases as the volume of data which must to be processed increases. So measurements or result of their processing are available to end users on the average in half a year after receiving measurements. Also, part of errors is removed well after, and the general time of identification and removal of errors can take about two years.

Users of oceanographic systems (for example, tools for hydroacoustic monitoring of the water environment) have to deal with all these problems. For the analysis of oceanographic data ready-made products of the analysis are used that are usually updated two times in a year. Thus access to operational data isn't provided. It leads to decrease in accuracy of estimate of water environment state and, respectively, decrease in operating benefits of hydroacoustic tools.

The considered problems can be effectively solved at the expense of use of dynamic information model for oceanographic data representation that is reflective to actual state of water environment and also state of subject domain objects. The basis of proposed dynamic information model is a set of three models. It is statistical model of data description, logical model of the data relations description, model of the description of processes of change of a water environment parameters states. The dynamic information model is constructed on the basis of set of data mining methods.

2 Description of Oceanographic Data

For 30 years the basic source of ocean data was data received from oceanographic stations and mooring buoy station. Total amount of data made was about 500 measurements per day. Argo project [3] was started in 2000. The target number of Argo buoys was 3000. Currently general number of buoys are 122, general number of measurements are 109050. Number of oceanographic stations, bathythermospheres, buoys constantly increases. From all sources about 2000 measurements are received each day. At the present time total number of stations is about 12 million. Total amount of the available data contains 14 million of temperature profiles and 5 million of salinity profiles. Each profile represents set that contains time, earth coordinates, depth level and related measurements. Figs. 1-3 illustrate examples of temperature and salinity measurements.

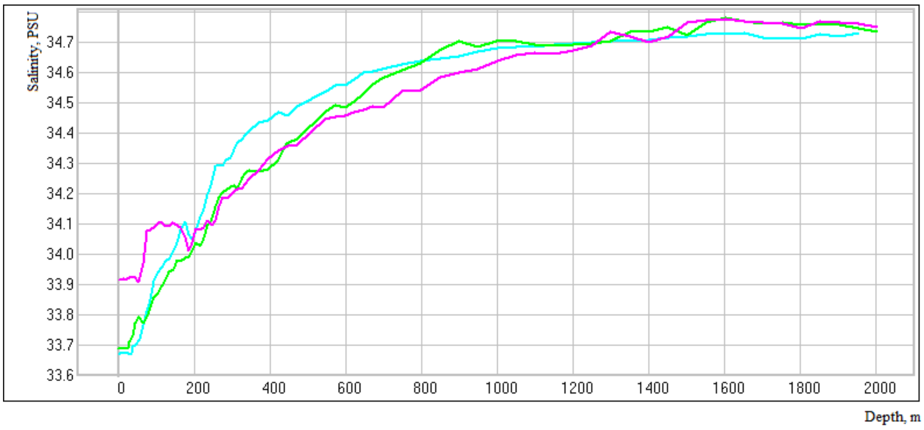

Fig. 1a. Salinity measurements.

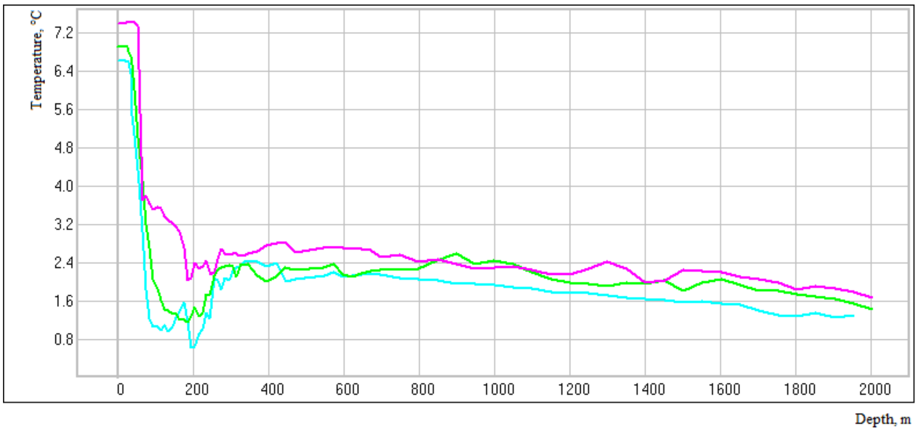

Fig. 1b. Temperature measurements.

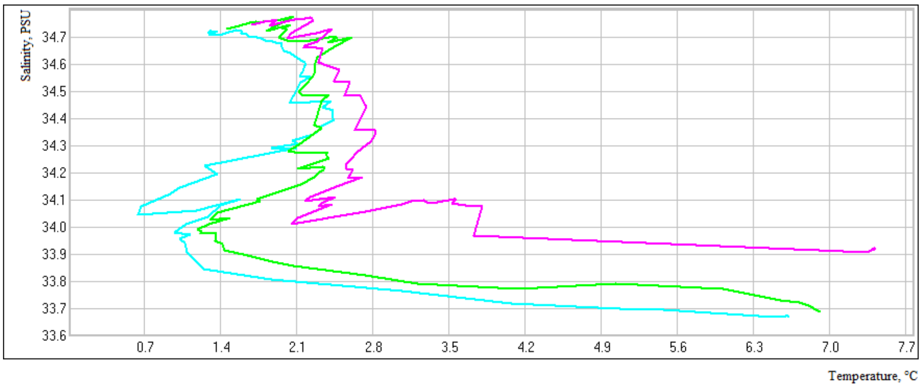

Fig. 1c. Temperature VS salinity.

Fig. 1. Measurements from Argo buoys (region is (49.5-50.5° S, 37.6-38.2° W)).

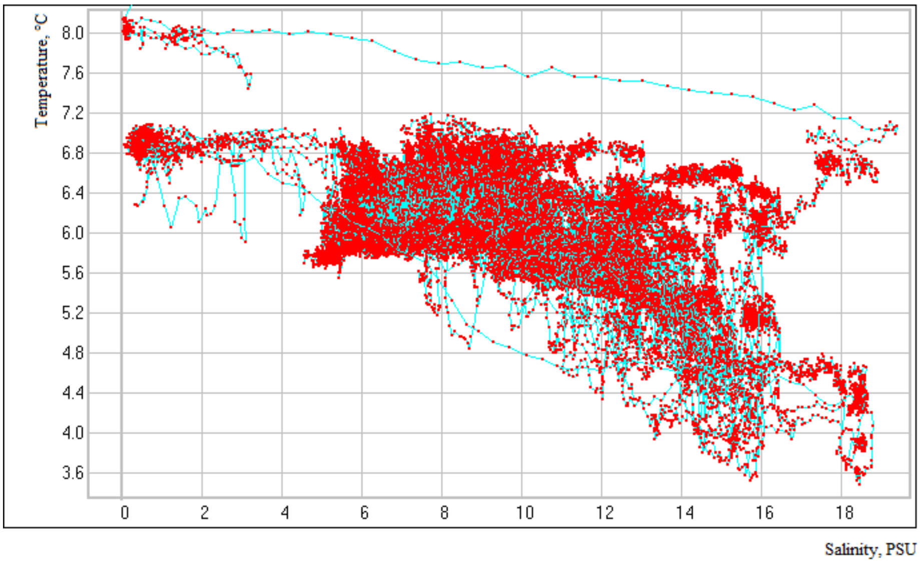

Fig. 2a. Temperature VS salinity.

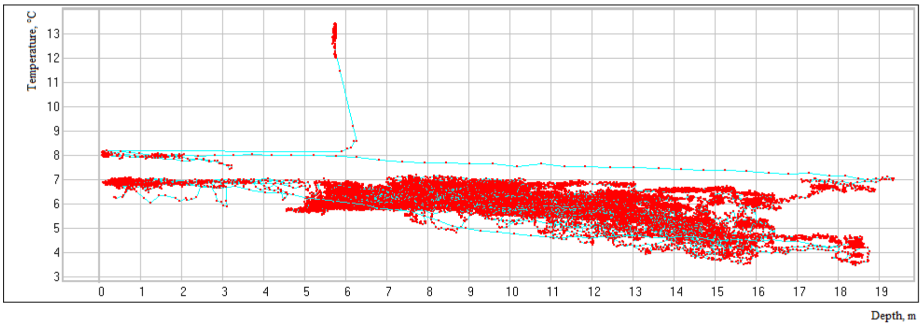

Fig. 2b. Temperature VS depth.

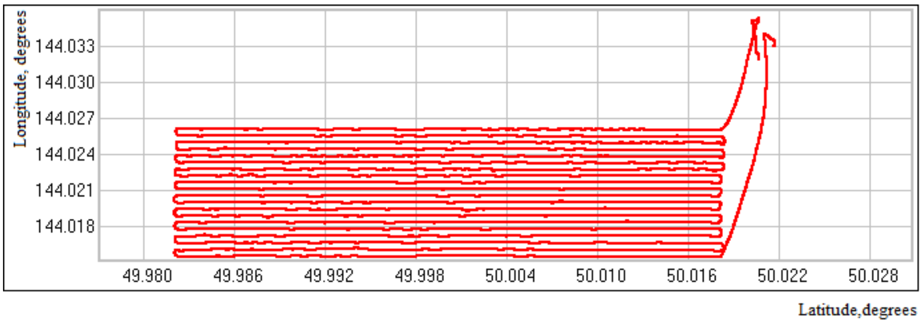

Fig. 2c. Trajectory.

Fig. 2. Measurements from autonomous underwater vehicle (region is (49.9-50.02° S, 144.0-144.033° W), time period is October).

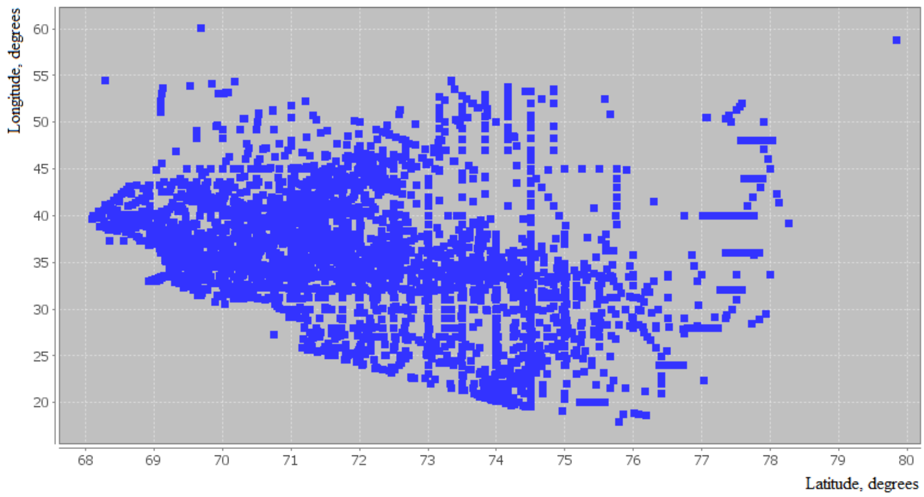

Fig. 3a. Latitude VS longitude (depth is (0-5).

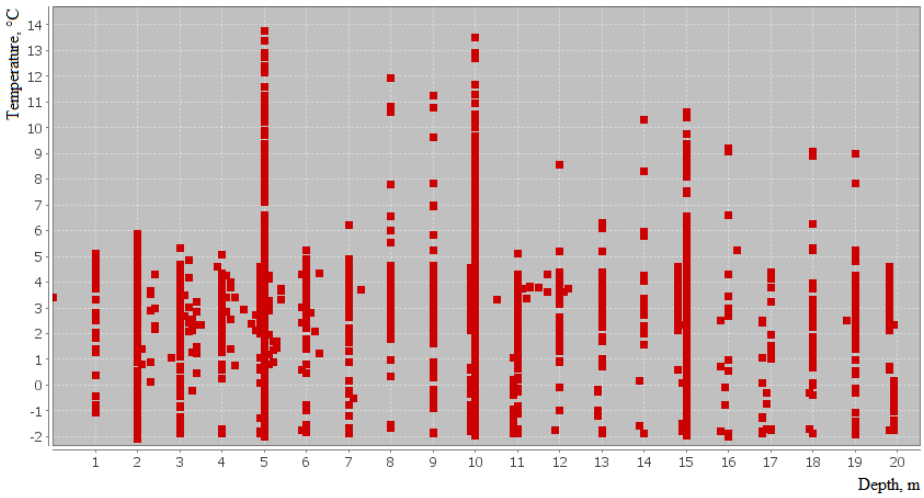

Fig. 3b. Depth VS Temperature.

Fig. 3. Measurements from various oceanographic stations (region is (60-80° N, 15-60° E), time period is January).

Analysis of data showed that measurements are not regular both by time and coordinates with the exception of data received from fixed stations. In additional data and its statistical characteristics for different regions are far from being similar and need special solutions for their processing. As well, processing of measurements assumes specialized methods, particular at the stage of data quality estimation, specific for each type of source.

Measurements have following particular characteristics:

- measurements are time series with different behavior. That is because they are received in different regions using different measuring tools. A set of external factors influence strongly on received values, for example, the seasonal phenomena, state of water environment of contiguous area.
- data contains considerable number of error values, for example, noise, outliers, gaps and also offsets and trends due to errors of measuring tools. Measurements on each data source and on each region demand application of specialized methods of processing. The majority of them requests participation in process of the expert.

Problem solution of automatic choice of methods and definition of their parameters is carried out by means of the use of adaptive approaches to data processing based on domain knowledge and statistical data.

3 Description of Dynamic Information Model for Oceanographic Data

Dynamic information model is integrated model of oceanographic data description developed on the basis of historical data and expert knowledge of subject domain. It provides actual data corresponding to environment and objects settings. Components of this model, which are based on [4, 5, 6], are [7, 8]:

- statistical model of data description is used for formalized description of separate measurements and their set, and also knowledge about the measurements received as a result of their processing. The following types of data and knowledge representation are used: the initial measurements representation including initial measurement representation model, results of data harmonization representation, including models of structural measurements representation, representation of results of data integration, including models of semantic representation of measurements;
- logical model of data relation description is a set of models that includes: models of representation of data integration results, including models of representation of multidimensional measurements, models of the qualitative and quantitative data description, representation of results of data fusion, including models of heterogeneous data combined representation;
- model of the description of processes of change of water environment parameters state represents relations between different processes on quantity and quality levels.

Dynamic information model for oceanographic data allows solving three main problems:

1. provides representation of actual information on various subject domain objects and states water environment parameters at a given moment of time and given point in space and possibility of operative improvement of information as a result of processing of the received data;

2. provides the short-time forecast of a state of basic parameters of the environment taking into account available data, knowledge and factors, that impact on state of parameters;
3. provides information on data relation and dynamics of the parameters change.

Primary properties of dynamic information model are:

- model is multilevel in the context of information content, it contains information of various levels - from initial data to knowledge about processes;
- model is multilevel and hierarchic and reflects the structure of subject domain– from separate measurements and group of measurements to measurements of separate regions;
- model is multidimensional (with different granularity);
- model is capable to accumulate all previously gathered data and knowledge;
- model is capable to provide rating and accounting of external factors, that influence directly or indirectly on state of the environment.

Harmonization, integration and fusion data [9, 10] and also statistical analysis and data mining are key technologies that are used in the dynamic model.

Data harmonization suppose definition of main concepts and their relationship on the corresponding subject domains and/or responsibility spheres. The general procedure of data integration assumes: an assessment of data quality from each source on the basis of specialized set of tests; search and exclusion of duplicating values; statistical data processing of each set of measurements, including denoising, removing outliers, identification of trends, filling gaps; interpolation of data. Data fusion is defined as process of data combination from various sources which allow to receive information of new quality and reduce its size. Statistical analysis and data mining provide task solution of system data processing and knowledge acquisition from data.

4 Description of Data Mining Technology

The general method of multidimensional measurements analysis using data mining methods is given in Fig. 4. Proposed stages are general and depending on data type and solved task stages can be skipped.

Stage 1. « Structure analysis». For initial data, that is a structured binary stream, that contains measurements, task of stream structure validation is solved.

Stage 2. «Measurements extraction». Measurements extraction assumes parameter measurement extraction from data stream according to its description.

Stage 3. «Definition of measurement types». For each measurement, received on the stage of measurements extraction, its type is defined. Parameters, which possess priori formed set of properties, refer to one type. The constant and spinner can be examples of measurement types.

Stage 4. «Data preprocessing». Cleaning measurements from noise or outliers, exclusion of trends, filling missing value is implemented on the stage of data preprocessing. In additional, statistical analysis of measurements is fulfilled, for example, statistical analysis of distribution parameters, regression analysis, spectrum analysis.

Stage 5. «Data segmentation». The stage assumes segmentation of time series, so that each segment has a defined set of constant properties. Segmentation can be realized by experts or using segmentation algorithms. Segments and their characteristics are saved in model. When new data is received, it is segmented taking into account results of segmentation of historical data.

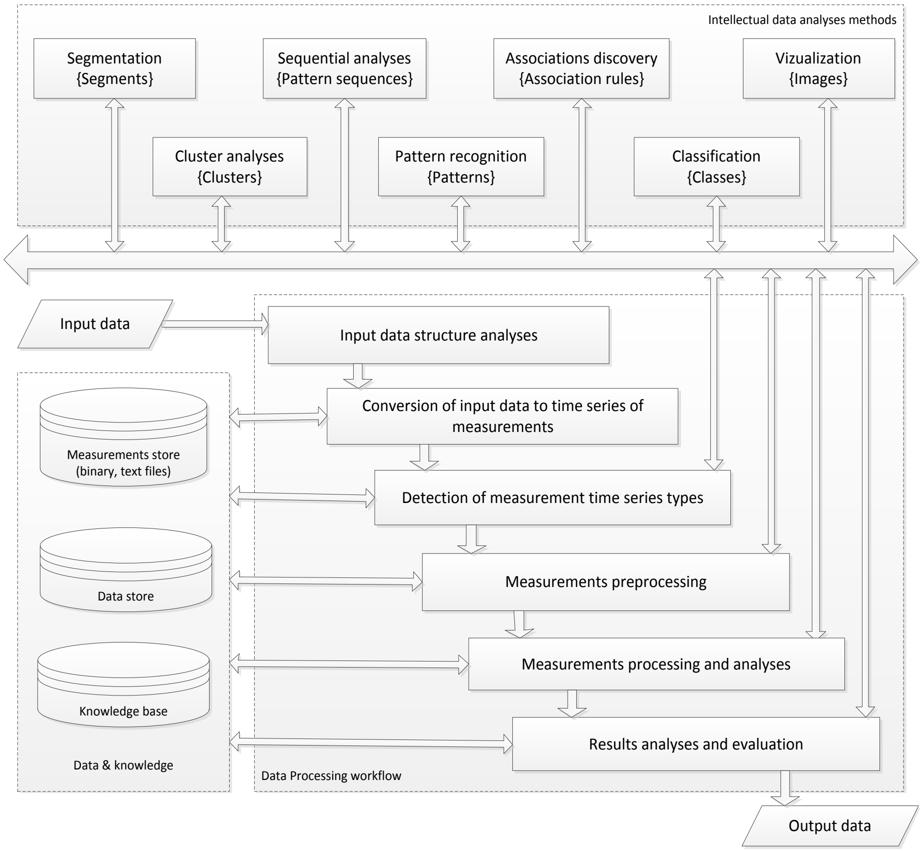

Fig. 4. General methods of multidemensional measurements analysis using methods of data mining.

Stage 6. «Cluster analysis». Clusterization problem consists of detection and description of confluence areas in analyzed space i.e. clusters are defined so, that distance between instances of one cluster is minimal and distances between instances of different clusters was maximal. Procedures for distances calculation are defined using specified criteria. When clustering time series first segmentation is made. Application of cluster analysis algorithms to time series allows revealing a set of possible time series states.

Stage 7. «Sequential analysis». This stage supposes searching time dependencies in sequence of segments. Time dependencies are represented in the form of a pattern sequences. Formed patterns are saved in model. When analyzing new data, match of new data to patterns is checked.

Stage 8. «Association analysis». The stage assumes search of association dependencies in interval and qualitative data in the form of association rules. The rules are mined in historical data and then they are located in knowledge base. Discovered rules are applied for analysis of new data.

Stage 9. «Pattern recognition». The stage is intended for generation of measurement pattern on the basis of single-type measurement. Recognition of new data is realized by comparing new data and patterns.

Stage 10. «Visualizing results». When working with historical data analysis of initial data and results of analyses at different stages are visualized. When analyzing new data discovered mismatches are visualized.

Stage 11. «Obtained results analysis». This stage supposes representation of data processing results, oriented on expert use. It assumes usage of cognitive graphics methods and other visualization tools. At this stage formation or extending of knowledge base is realized.

Automation and adaptation of data mining processes and analysis of multidimensional measurements is performed by means of use of exploratory analysis and mechanisms of processing control. Procedures of prospecting analysis, that allow to receive priori estimates of data. According to estimations and using classification of measurement type and rules for data and knowledge representation of different types effective form of measurements and knowledge representation can be chosen and appropriate processing methods can be used. Mechanism of processing control is one of the central element in data processing and analysis systems. It provides data processing processes construction and correction. Mechanism of processing control is described in [11, 12].

5 Presentation of Dynamic Information Model in Intelligent Geoinformation System

Dynamic information model for representation of oceanographic data is realized under system of lighting situation. It is oriented on solution of wide range of problems, for example, search, detection, classification, definition of different objects parameters and also solution of hydroacoustical problems. Description of architecture of intelligent geoinformation system (IGIS) gives in [13]. Figs. 5-6 display examples of processed data and result of regular data grid construction on the basis of dynamic information model.

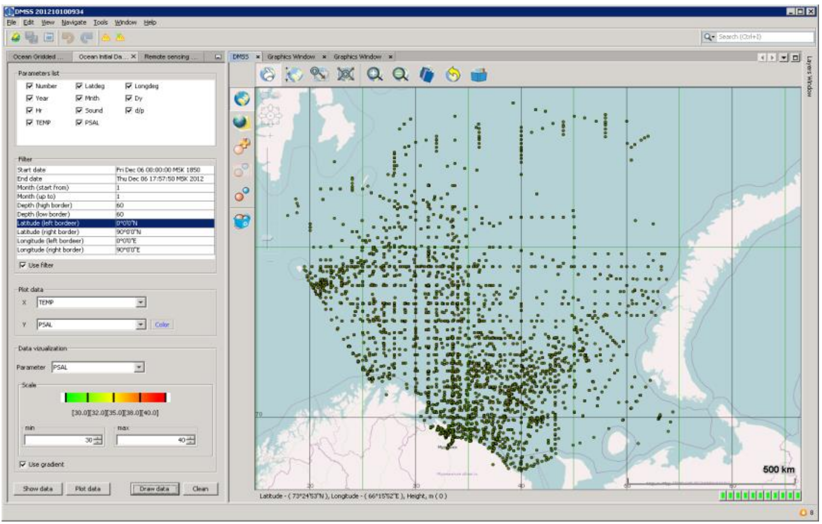

Fig. 5. Visualization of processed oceanographic data in IGIS.

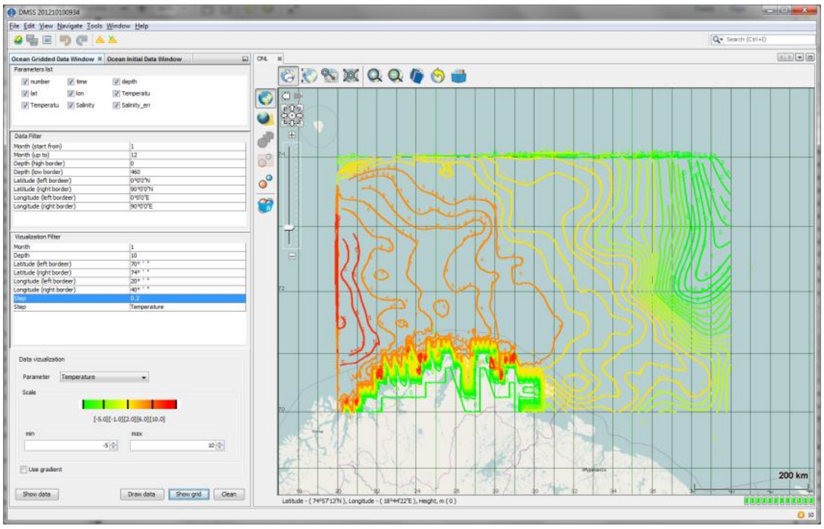

Fig. 6. Visualization of gridded data in IGIS.

6 Case Study

The dynamic information model of ocean data representation was constructed on the basis of data received from Arctic region during the period from 1876 up to now [13, 14]. Temperature and water salinity of Arctic region were measured at depths from 0 to 460 meters. Total number of performed measurements is about two million. Database of measurements is made and provided by the Arctic and Antarctic research institute on a grant of Office of Naval Research #62909-12-1-013 ("Decision Making Support System for Arctic Exploration, Monitoring and Governance"). In Fig. 7 temperature and water salinity values distribution and distribution of gathered data by years and depths is shown.

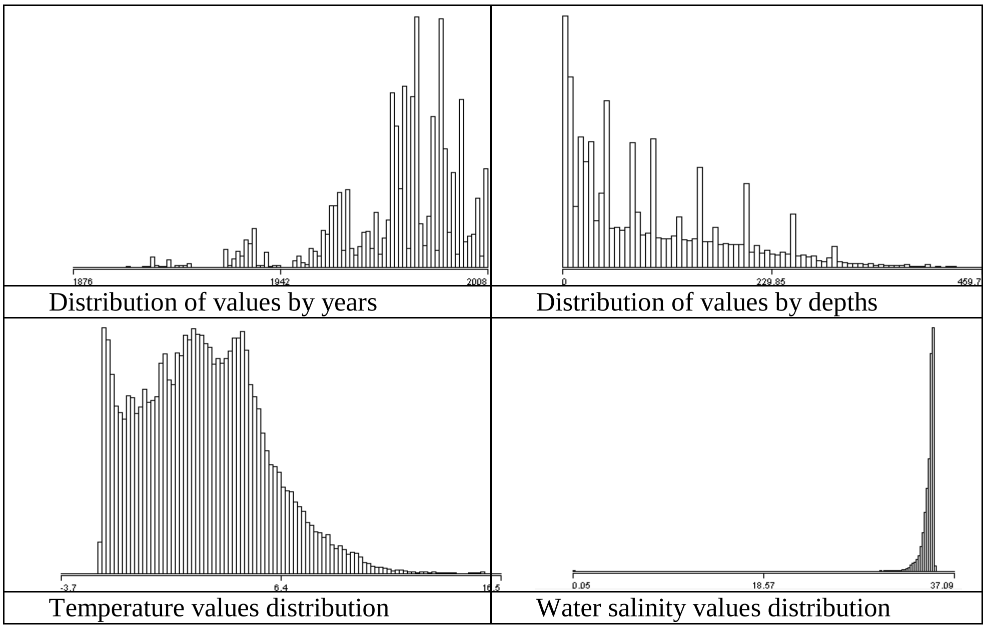

Fig. 7. Temperature and water salinity values distribution by years and by depths.

Example of application of data mining methods for solving task of operational data assessment obtained from external sources for the propose of decision making if it can be used at the next processing stages, particularly, for recalculating nodes of regular grid is given.

The task solution of operative data assessment is founded on comparison of received data with historical data of the same region at similar time intervals. As time line months in which measurements were received were considered. One of the most complex tasks is detection of stable regions in which values of analyzed parameters differ slightly. Task of region detection was solved using methods of cluster analysis. Below the description of procedure of region detection based on analyses of data received in various years in July is provided. As algorithm of cluster analysis SimpleKMeans algorithm was used, number of clusters was selected using estimation of result clusters compactness.

Step 1. Cluster analysis of initial data: time interval – from 1870 to 2008, time period – July, range of depths – from 0 to 460 meters, elements of feature space – latitude, longitude and depth of measurements, year of measurements conduction, values of temperatures and salinities. Results of cluster analysis are shown in Fig. 8, description of clusters is given in the Table 1. Borders of clusters take place at depths of 40 and 120 meters which is equivalent to border of water layers.

Table 1. Description of cluster centers developed for initial sample.

Attribute	1	2	3	4	5	6	7	8	9	10
Latitude, degrees	76.2	71.5	71.1	70.8	70.3	71.3	72.4	76.3	76.4	77.2
Longitude degrees	27.4	33.9	32.2	32.8	50.2	34.5	31.4	28.1	30.7	51.5
Depth, meter	77.4	110.5	28.2	12.3	22.7	98.8	246.9	17.4	179.4	58.5
Tempera-ture, celsius	0.6	3.2	5.9	6.6	3.4	2.8	2.3	1.8	0.7	-0.5
Salinity, PSU	34.6	34.7	34.1	33.7	32.7	34.7	34.9	34.1	34.8	34.4

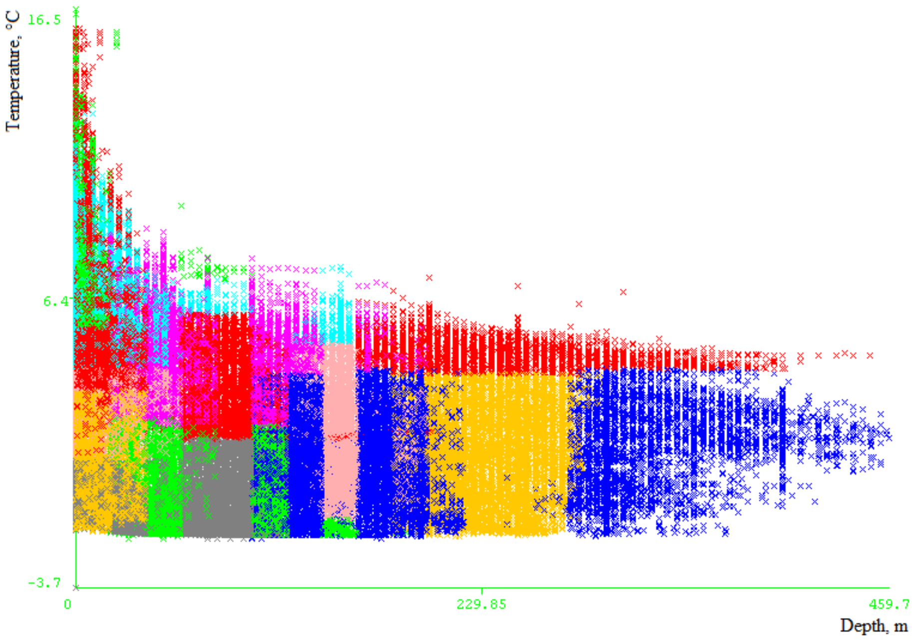

Fig. 8. Results of cluster analysis of initial samples.

Step 2. Cluster analysis of the data measured at depths of 0 - 40 meters. Results of the cluster analysis are shown in Fig. 9. Total number of the clusters are 5. At depth around 20 meters clear boundary of clusters is observed. It means that further data partitioning by parameter "depth" is to be done.

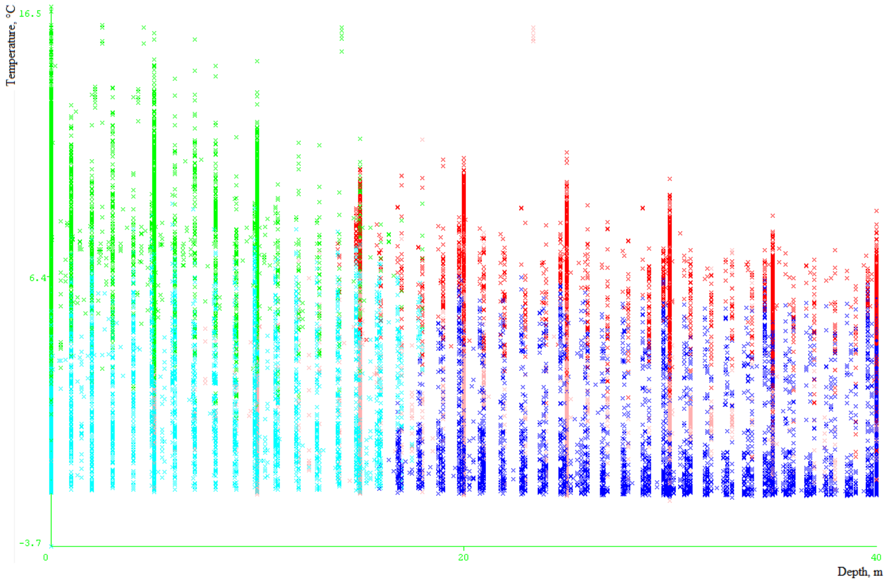

Fig. 9. Results of cluster analysis of data for depth 0-40 meters.

Step 3. Cluster analysis of data measured at depths of 0 - 20 meters. Results of the cluster analysis (Fig. 10) show that further data partitioning by parameter "depth" isn't expedient. However, clear clusters of measurements can be observed in the space of latitude and longitude (Fig.11) features. As boundary value latitude of 74 degrees is considered.

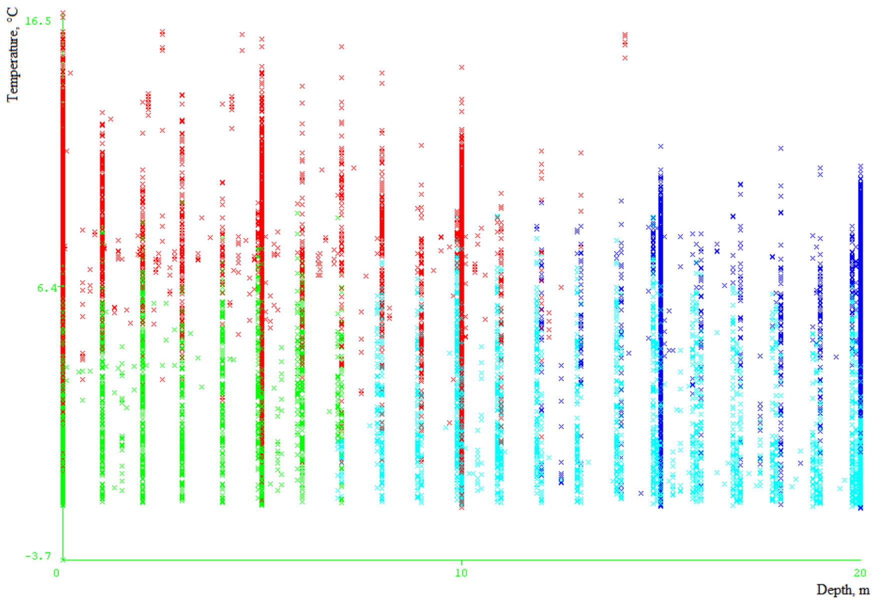

Fig. 10. Results of cluster analysis of data for depth 0-20 meters (Depth VS temperature).

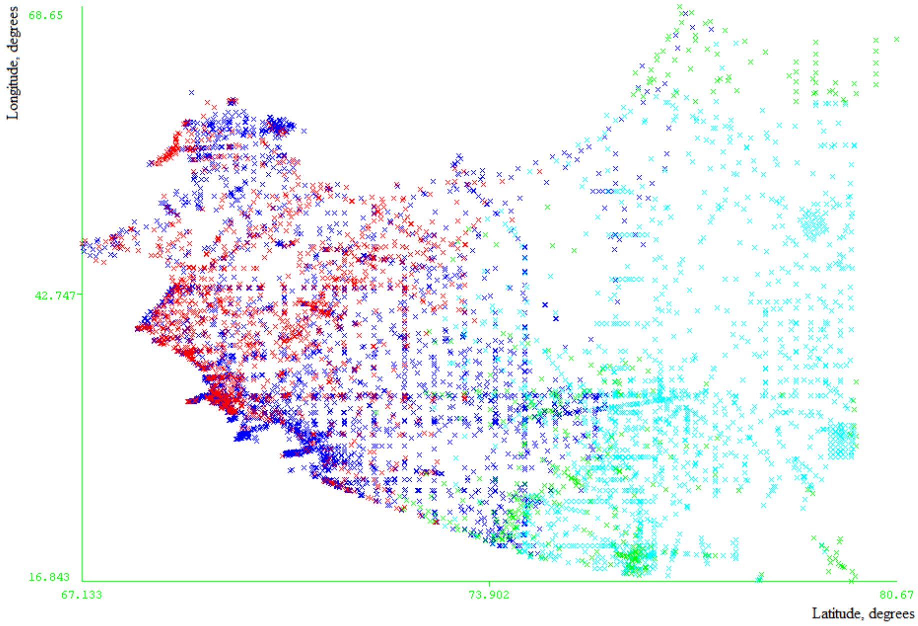

Fig. 11. Results of cluster analysis of data for depth 0-20 meters (Latitude VS longitude).

Step 4. Results of cluster analysis of data for depth from 0 to 10 meters and latitude more than 74 degrees is shown in Fig. 10. Further decomposition of data was done by parameter "longitude", for measurements with value of latitude more than 42 degrees.

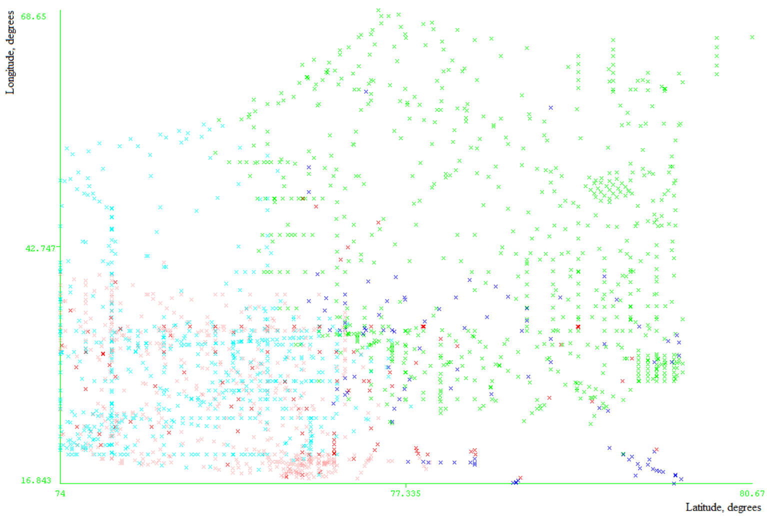

Fig. 12. Results of cluster analysis of data for depth 0-10 meters, latitude 74-81 degrees.

Step 5. Results of cluster analysis of data for depth from 0 to 10 meters and value of latitude more than 74 degrees and longitude more than 42 degrees are given in Fig. 13. At this stage data non-crossing clusters are formed.

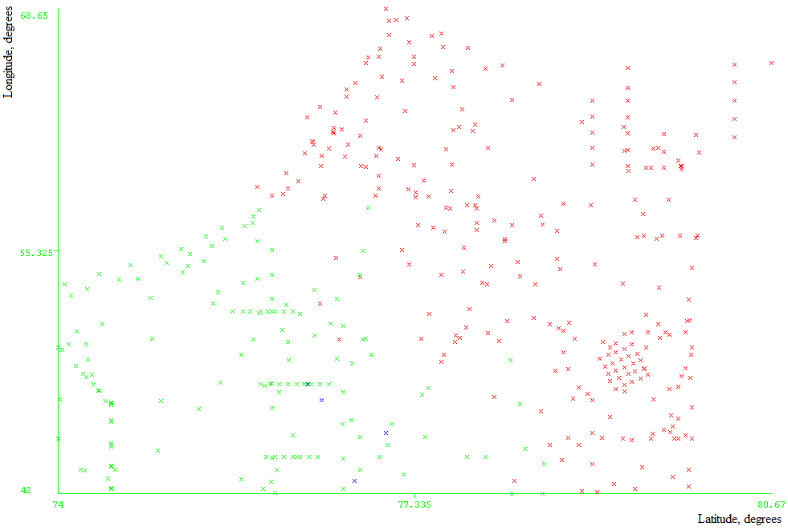

Fig. 13. Results of cluster analysis of data for depth 0-10 meters, latitude 74-81 degrees, longitude 42-69 degrees.

In the similar way all historical data on Arctic region was analyzed. All data space was decomposed on set of stable regions.

Results of cluster analyses were interpreted by specialists from Arctic and Antarctic Research Institute (Saint-Petersburg, Russia). Clusters border at depth of 20 meters corresponds to a wave-mixing zone. The zone exists during the time when there is no ice. Borders of clusters for depth of 0-20 meters are not quite clear because in July seasonal thermowedge is destroyed. Spatial distribution of data show zone of the Norwegian current, borders of distribution of Atlantic waters in Barents Sea.

7 Conclusion

The paper illustrates the dynamic information model for oceanographic data representation based on application of data mining methods and intelligent GIS technologies. Proposed model allowed to decrease processing time both in operational and delayed mode due to use of automated methods of data analyses, such as cluster analyses. That is important for different monitoring systems of water environment.

The further direction of researches is connected with application of biclustering and triclustering methods to oceanographic data. These methods are nowadays widely used in various spheres [1, 2, 3]. That allows take into account not only measurements but also time and location where measurements were received, so it can be expected that the rate of cluster compactness will increase.

Reference

1. Gnatyshak, D., Ignatov, D.I., Semenov, A. and Poelmans, J.: Gaining Insight in Social Networks with Biclustering and Triclustering. In.: Aseeva Natalia, Babkin Eduard, Kozyrev Oleg (eds.) Perspectives in Business Informatics Research, Lecture Notes in Business Information Processing. Volume 128, Part 4 (2012) 162-171
2. Ignatov, D., Poelmans, J., Zaharchuk, V.: Recommender System Based on Algorithm of Bicluster Analysis RecBi. In CEUR Workshop proceedings, CDUD'11 – Concept Discovery in Unstructured Data. Volume 757 (2011), 122-126
3. Ignatov, D.I., Kuznetsov, S.O., Magizov, R.A. and Zhukov, L.E.: From Triconcepts to Triclusters. In: Kuznetsov et al. (eds.) RSFDGrC 2011, LNCS/LNAI, vol. 6743/2011, pp. 257-264. Springer-Verlag Berlin, Heidelberg (2011)
4. The International Argo Project Homepage, http://www.argo.net/
5. W3G Geospatial Ontologies, http://www.w3.org/2005/Incubator/geo/XGR-geo-ont-20071023/
6. OGC Standards and Supporting Documents, http://www.opengeospatial.org/standards
7. Llinas, J., Bowman, C., Rogova, G., Steinberg, A., Waltz, E., White, F.: Revisiting the JDL data fusion model II. In: Proceedings of the Seventh International Conference on Information Fusion, Stockholm, Sweden (2004)
8. Liss, A.R., Zhukova, N.A.: Software System for Processing On-Line Information of Complex Dynamic Objects. In: Proceedings of Saint Petersburg Electrotechnical University "LETI". Issue 5 (2010), 67-72
9. Deripaska, A.O., Zhukova, N.A., Pan'kin, A.V.: Adaptive Selection of Processes of Handling and Analysis Multivariate Measurements in Intelligent Information Systems. In: Proceedings of 13th Russian conference on Artificial Intelligence with international participation, October 16 – 20, 2012, Russia, Belgorod (CAI-2012) (in Russian)
10. Pankin, A.V., Kuzeny, V.V.: Data Harmonization in CIS. In: Proceedings of International Conference of Information Fusion and Geographic Information Systems 2009, St. Petersburg, pp. 63-76, Springer, Berlin
11. Popovich, V.V., Potapichev, S.N., Sorokin, R.P., Pankin, A.V.: Intelligent GIS for Monitoring Systems Development. In: Proceedings of CORP2005, February 22-25, 2005, University of Technology Vienna
12. Zhukova, N.A., Pankin, A.V.: Principles of managing the processing and analysis of multidimensional measurements in IGIS. In: Proceedings of the Information technologies in management, St. Petersburg, October 9 – 11 (2012)
13. Smith, H., Fingar, P.: Business Process Management (BPM): The Third Wave, Meghan Kiffer Press (2003)
14. Popovich, V., Pankin, A., Galiano, F., Potapichev, S., Zhukova, N.: Service-Oriented Architecture of Intelligent GIS. In: SOMAP 2012
15. Korablev, A. A., Pnyushkov, A.V., Smirnov, A.V.: Creation of an oceanographic database for climate monitoring in the North European basin of the Arctic. In: Trudy AANII. Issue 447 (2007). 85-108
16. Ashik, I.: Recent Russian Marine Research Activities in the Arctic Ocean. Arctic Science, International Law and Climate Change. Volume 235 (2012), 59-66