**V. B. Mokin, Dc. Sc. (Eng.), Prof.; L. M. Skoryna; Y. M. Kryzhanovskyi, Cand. Sc. (Eng.), Ass. Prof.; M. A. Horash**

# CONSTRUCTION OF GIS-INTEGRATED SYSTEM OF DATA AND MODELS, BASED ON XML FORMALIZATION, FOR SIMULATING PROCESSES, TAKING PLACE IN RIVERS

*The paper presents the conducted analysis of the known formats for formalization of mathematical models and spatial formats which are relevant for processing data and models occurring in rivers, namely, PMML, MathML, SBML, GML, WaterML, according to the system of the following criteria:availability of repositories of the already identified models, integration with programming languages, availability of specific tags for working with spatial data, availability of environments for automated formalization and import/export of XML models. An integral criterion is proposed. Examples of using this criteria system for selection of the optimal XML format for storing data and models are presented for different weights of certain criteria and conditions: if availability of geo-reference is the main criterion, Water ML will the optimal XML language, if it is the possibility to formalize any analytical models, especially hydro-biological ones, then SBML will be the optimal choice and if automation of artificial intelligence algorithms is required, then PMML should be chosen.*

*The following known technologies for storing attributive and spatial data of GIS for their automated processing are characterized: KML, Shapefile, GPX, GeoJSON, SXF, ArcGIS geo-data base, Spatialite (SQLite), MapInfo TAB format. It is noted that by the criterion of the number of references with the word "map" found by Google search system such formats as KML (22 million), GPX (20 million) and Shapefile (11,5 million) are the most popular in the world.*

*GIS integrated system of data and models based on XML formalization is proposed for the first time and its operability is illustrated by the example of predicting average annual water consumption over a multiyear period for 50 % provision in the Dniester river basin in KNIME Analytics Platform environment.*

*The obtained results make it possible to provide fast construction, versatility and broad functional of GIS-integrated system of data and models.*

***Keywords:*** *geoinformation system, mathematical model, database, XML, PMML, MathML, SBML, GML, WaterML, KNIME, simulation of the processes, river.*

## Introduction

Analysis of publications in the journals with high impact-factor in SCOPUS system indicates that considerable number of scholars in the leading universities and research institutions of many countries in the world (USA, Canada, France, Netherlands, Finland, Spain, Norway, Taiwan, etc.) [1 – 9] are engaged in the research relating to the problem of simulation and optimization of the processes occurring in complex cybernetic systems with GIS application. Most research teams are developing narrow-field specialized GIS-integrated bases of data and models (e.g. for simulation of water condition [1] or air pollution over Europe [2]), but such an approach, however, limits their development. Many researchers in Ukraine, including those from the institutes of NAS of Ukraine (O. Y. Stryzhak, M. Y. Zhelezniak, Y. O. Yevdin and others) and abroad, are engaged in the research work, mostly, on the creation of ontological GIS-integrated databases [3, 4] or systems based on the known software platforms (e. g. OpenMI, OMS, TIME, KEPLER, FRAMES, MODCOM [5], distributed wrapper objects [6]). However, this approach complicates both design of such systems and development of their analytical capabilities as well as their application for optimizing parameters of the mathematical models. Instead, scientists dealing with processing of medical and biological information use XML markup languages (both those specific for their field – CellML, SED-ML, SBML and universal languages – MathML, PMML [7, 8]) for formalization of all the model parameters, including formalization by importing the models from the known

computational packages. This enables formation of international repositories of these models. However, in such packages insufficient attention is paid to formalization and storage of spatial information and integration with it. Another approach is formation of object-oriented programming languages (DSL), although the existing analogs or their prototypes do not provide sufficient interaction and interoperability with various formats of storing data and models

To solve the above problems, it is expedient to develop universal approaches to formalizing metadata of spatial data of GIS and methods for integration of these data with mathematical models, formalized in standard markup languages (MathML, CellML, SBML etc.), with data processing methods (DataMining and others), formalized in PMML mark-up language. Such integration of different formats and approaches will provide fast construction, universality and wide functionality of GIS-integrated systems of databases and models.

This paper aims at analyzing known technologies and data formalization formats and using them as a basis for building a single GIS-integrated system of data and models based on XML-formalization.

As general analysis of the entire variety of data and models, based on XML formalization, is a large-scale problem, it will be expedient to limit it to the problems of simulation of the processes occurring in the rivers. This problem is especially relevant for Ukraine that occupies the last place in Europe as to the annual surface water supply per one person.

**Analysis of the known XML markup languages of mathematical models and environments for their automatic processing**

In order to achieve the stated aim, first of all it is necessary to analyze the existing standard XML-markup languages for storing data models, the existing standard formats for storing attributive and spatial data of geoinformation systems (GIS) and to determine the way for their automatic, maximally fast and reliable mutual integration.

The existing standard XML-markup languages were compared according to the following criteria:

– $X_1$ – the number of classes of mathematical models by the example of simulating processes occurring in the rivers: as 1.0, we chose MathML, where any analytical expression can be formalized, and as for the others, the number of standard mathematical models that could be formalized in them was evaluated by the example of the problems related to simulation of the processes occurring in the rivers.

– $X_2$ – availability of repositories of the models (not only templates) that were already formalized: if an adequate repository is available, then $X_2 = 1$, otherwise, depending on the proportion of the number of model classes in repositories from the total quantity of the model classes in the problem of simulating the processes occurring in the rivers, $X_2$ – from 0 to 1.

– $X_3$ – integration with programming languages if sufficiently universal programming languages with special libraries for direct handling the information in the given format exist, which would provide practically unlimited possibilities for data management and their automatic processing. Then $X_3 = 1$, otherwise, $X_3 = 0$;

– $X_4$ – availability of specific tags for working with spatial data (we mean not only the possibility of writing spatial coordinates into separate attributes, but tags that are specific for spatial analysis): if such tags exist, $X_4 = 1$, otherwise $X_4 = 0$;

– $X_5$ – availability of the environments for automatic formalization and import / export of XML models (preferably, open-source ones): if at least one such environment exists, then $X_5 = 1$, otherwise, $X_5 = O$.

We propose to choose the following classic criterion $J_X$ as an integral one:

$$J_X = \sum_{i=1}^{N} w_i X_i , \qquad (1)$$

where $w_i$ – the weight of the $i$-th criterion ($i$ = 1, …, $N$), which is determined experimentally and satisfies the condition:

$$\sum_{i=1}^{N} w_i = 1. \quad (2)$$

We analyzed XML-markup languages, which could be useful for formalization of the processes occurring in the rivers, with available repositories of those models (both universal formats and specific for hydrobiological and hydrological models as well as for models with geoinformational referencing):

– PMML (Predictive Model Markup Language) – main universal markup language for artificial intelligence models (http://dmg.org) [9 – 11];

– MathML (Mathematical Markup Language) – universal markup language for symbols and formulas;

– SBML (Systems Biology Markup Language) – markup language for mathematical models of biological processes (http://sbml.org) [7, 8, 12];

– WaterML ( Water Markup Language) – markup language for representation of hydrological structures (https://en.wikipedia.org/wiki/WaterML) [14, 15];

– GML (Geography Markup Language) – markup language for simulation of geographical (geoinformational) systems (http://www.opengeospatial.org/standards/gml) [13].

Some of them are interconnected, e.g. WaterML uses GML for geo-referencing of the data, and SBML – MathML for formalization of the models .

In general, other XML standards also exist, e.g. for biological models with similar characteristics the following ones are used: CellML and SED-ML (Simulation Experiment Description Markup Language). The following known computational software packages also handle XML: Matlab, Mathcad, MS Excel, etc., but discussion of their capabilities is a subject of a separate study.

The results of comparison of these markup languages according to the above criteria is presented in Table 1.

Table1

**The results of multicriteria analysis of the known XML markup languages for mathematical models**

| Criteria $X_i$ | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $J_X$ |
|---|---|---|---|---|---|---|
| Weights $w_i$ | 0.2 | 0.1 | 0.2 | 0.3 | 0.2 | |
| WaterML | 0.8 – practically any analytical models of the processes occurring in the rivers | 1 – «WaterOneFlow» bank with a large quantity of models [16] | 1 – Java, Python | 1 (due to the use of GML) | 1 – HydroDesktop, HEC-DSSVue, GEOSS Water Services | 0.96 |
| GML | 0.1 – only models for conversion of spatial information | 0.2 – y [15] There are 6 models used as examples, but other models also exist | 1 – Java, C++ | 1 | 1 – FME, Document Object Model, OpenGIS | 0.74 |
| SBML | 1 – any analytical models due to the use of MathML for formalization | 1 – «BioModels» bank of models that as of June 26, 2017 contained 143 070 models [14] | 1 – C, C++, Java, Python | 0 – not provided | 1 – BioSPICE Dashboard, iBioSim, JSim, SynBioSS | 070 |
| MathML | 1 – any analytical | 0.5 – models exist, | 1 – | 0 – not | 1 – MS Word, | 065 |

| | | | | | | |
|---|---|---|---|---|---|---|
| | models could be written | but mostly for simple mathematical operations | JavaScript | provided | Apache OpenOffice, MathMagic, MathType, Maple | |
| PMML | 0.25 – only models of artificial intelligence and intellectual analysis (17 – 20 classes of models depending on the version) | 1 – many models exist, see [12] | 1 – Java, R, Python, Perl, SQL | 0 –not provided | 1 – there is a large quantity of environments: KNIME Analytics Platform, WEKA, KXEN [13] | 0.55 |

Analysis of the information presented Table 1 shows that if main criteria are used in the following order: first, availability of geo-referencing (weight 0.3), then the number of classes of mathematical models, integration with programming languages and availability of the environments for automatic data processing (weight 0.2), then WaterML will be the optimal language. If the main criterion is the possibility of formalization of any analytical models, especially hydrobiological models (weight $X_1$ – 0.4) and availability of geo-referencing is not essential (weight $X_4$ – 0.05), then SBML language will be optimal. At the same time, if automation of the artificial intelligence algorithms is required, especially of those that do not have a single analytical expression and contain a complex algorithm for their application (neural networks and different mathematical Data Mining tools), then PMML turns out to be beyond competition.

**Analysis of the technologies for storing attributive and spatial data of GIS for their automatic processing**

Experience of working with the formats and technologies for storing attributive and spatial data of GIS shows that the following formats are optimal and the most widely used today for the problems of their automatic processing (ordered by descending number of documents in Google in the context with the word "map") [16]:

– KML (22 million) – XML-based markup language for representing three-dimensional geospatial data in the popular service "Google Maps" (www.google.com.ua);

– GPX (20 million) – XML-based textual format for storage and exchange of GPS data;

– Shapefile (11.5 million) – popular spatial format of GIS data, developed by ESRI, which acquired the status of the universal standard for formalization of spatial vector data for both commercial and free-access software packages;

– GeoJSON (2.4 million) – JSON-based format with open code for coding various geo-data structures (www.geojson.org);

– SXF (0.7 million) – open-source format of digital information about the area designed for application in geo-information systems for storing digital information about the area, data exchange between different systems, creation of digital and electronic maps and solving applied problems (https://ru.wikipedia.org/wiki/Sxf);

– geo-database ArcGIS (0.65 references in Google is a set of the arrays of various data types in ArcGIS software package (developed by ESRI – Environmental Systems Research Institute, USA), which are stored in the free-access place of the file system – database (www.esri.com);

– Spatialite (SQLite) (0.3 million) is a spatial extension for SQLite, which provides vector functionality of geo-data (www.sqlite.org). Such extensions also have database management systems PostGIS, Oracle Spatial and SQL Server with spatial extensions;

– MapInfo TAB format (0.28 million) – vector format of the «MapInfo» software package with the use of TAB, DAT, MAP files (www.pbinsight.com).

Analysis of the structure of these formats shows that each of them contains elements that provide storage of spatial data in the form of object coordinates as well as their attributive data. In terms of

the use of GIS-formats for integration with the data and models as well as at the stage of the simulation results visualization, the following functional capabilities are significant [16, 17]:

1) Storage of the simulation results in the form of attributes using software;
2) Association (establishing connection) between simulation results and certain spatial objects;
3) Programmable read-out of both attributive and spatial data;
4) Support of specific geo-information models (geometric networks with information about their topology (for simulation of river systems)) by some of them, TIN models (for models of catchment basins of rivers), interpolation of the surfaces by isolines (for models of the river basin drainage maps), etc.).

Each of the above GIS formats adequately supports these functional capabilities and, therefore, could be used for realization of GIS-integrated systems of data and models. This will make it possible to raise the level of the system universality and automate its identification and use for solving many applied problems. And now we will pass to the issue of automatic integration of all these components into a single whole.

**Building structure of GIS-integrated system of data and models**

We propose structure of GIS-integrated system of data and models, based on XML formalization, in the form presented in Fig. 1.

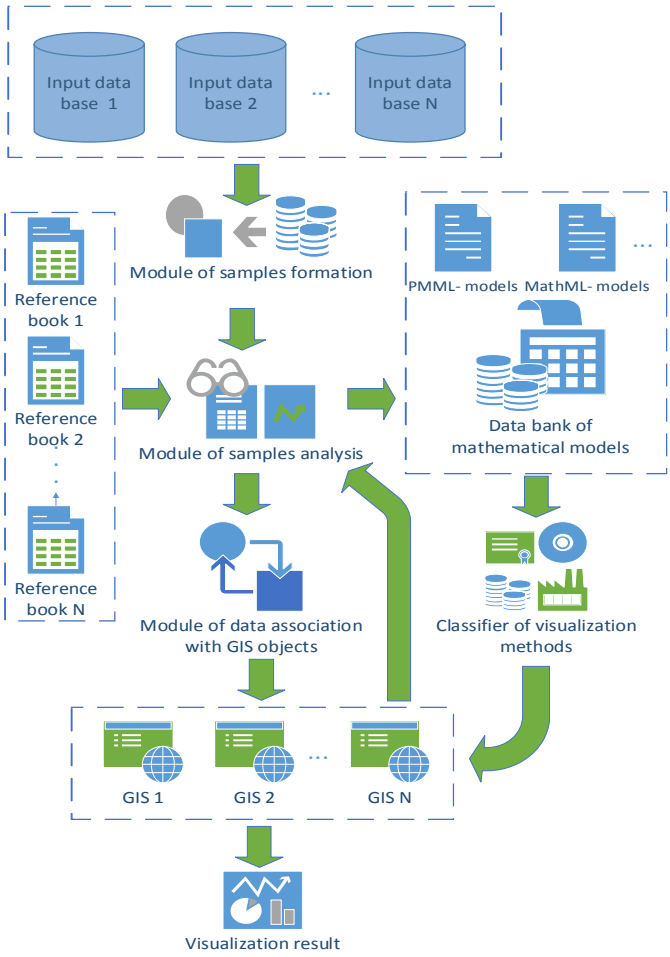


Fig. 1. Structure of GIS-integrated system of data and models based on XML formalization

The structure of GIS-integrated system of data and models, presented in Fig.1, includes the following components: databases that contain input information serving as a basis for simulation and implementation; a module for formation of samples that generates output data arrays based on their

selection from databases of input information according to certain criteria; a module for analyzing samples that provides post-processing of the selected data with the possibility to involve data from various directories taking into account spatial regularities of those data and their specific geo-informational models; data bank of mathematical models that contains classic and specialized models formalized in different formats (see Table 1); a module of the data association with GIS objects that provides data referencing to the objects of GIS layers; a classifier of the data visualization methods that contains a list of GIS visualization methods that could be used for visual representation of the simulation results taking into account specific features of the models; GIS complex that contains spatial and related to it attributive information about objects, which are simulated, in different GIS; visualization results that could be represented in the form of topical

Operability of this structure will be illustrated by the example given below.

**The example of simulation of the processes occurring in the rivers with the application of GIS-integrated system of data and models**

The problem of predicting the average annual water consumption for 50 % provision over a multiyear period at 17 gauging station of water management regions of the Dniester River basin can be used to construct the water balance of this basin. As it is known from research [18] and web-system (http://vb.dniester-basin.org/), created within the framework of OSCE (Organization for Security and Cooperation in Europe) project with participation of the authors of this paper, the following information is required in order to solve this problem:

– coordinates of 17 gauging stations of water management regions in the Dniester River basin;

– models of the regression analysis of the provision curves.

For formalization of regression analysis in KNIME Analytics Platform environment PMML format was used (see Table 1). As it was noted above, it is optimal for statistical analysis of artificial intelligence problems. According to the structure in Fig. 1, GIS-integrated system of data and models was built in the form presented in Fig. 2.

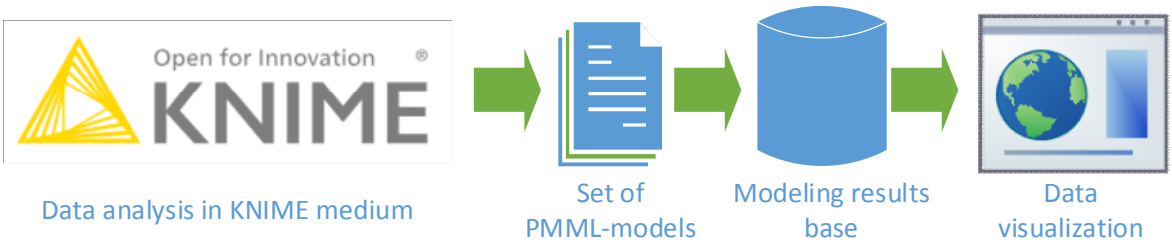


Fig. 2. The structure of GIS-integrated system of data and models for predicting provision of the water flow in the river basin, based on PMML format, in KNIME Analytics Platform environment

Slution of the set problem included the following stages (Fig. 3):

1. Connection of the database about average monthly water consumption over a multiyear period at the gauging stations of the water management regions of the Dniester river basin from MS Excel package (it is the format where the above-mentioned OSCE system saves computation results about the Dniester river).
2. Formation of the data sample about average annual water consumption for all 17 gauging stations for different provision levels and its analysis.
3. Formation of the separate samples of average annual water consumption for different provision levels for each gauging station of the Dniester river.

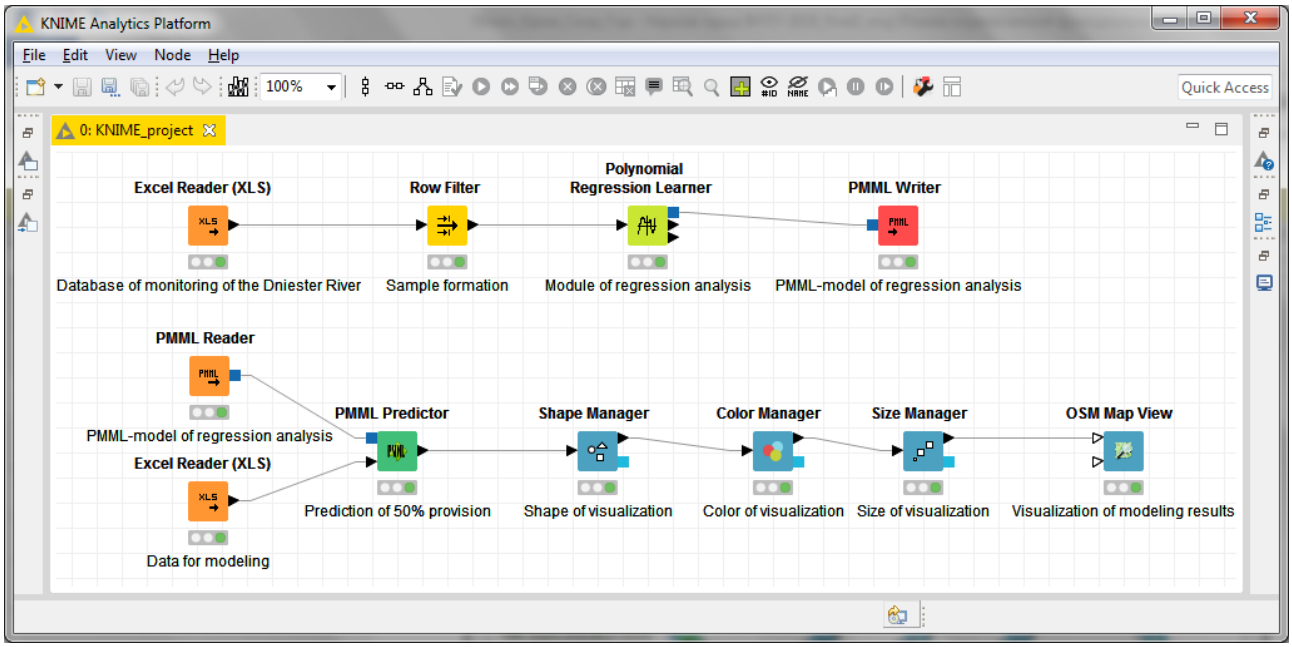


Fig. 3. System implementation in KNME Analytics Platform environment

4. Using the model of regression analysis of the provision curves for each gauging station separately (simulation results by the example of the gauging station “Rozdol» in the Dniester river basin are presented in Fig. 4).

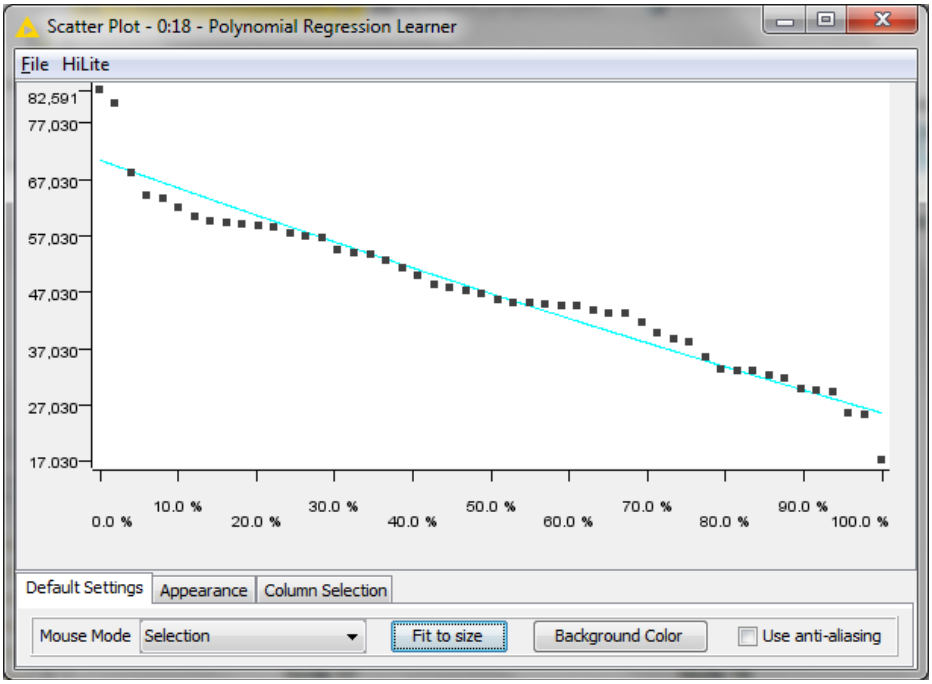


Fig. 4. The regression analysis results: provision curve of the water station “Rozdol”

5. Formalizing the obtained models of the regression analysis of the provision curves with the application of PMML, writing them into 17 separate PMML files and forming database of PMML models.
6. Based on the obtained set of PMML models, prediction of the average annual water consumption for 50 % provision over a multiyear period for 17 gauging stations at water management regions of the Dniester river basin.

7. Displaying simulation results at the OpenStreetMap (Fig. 5).

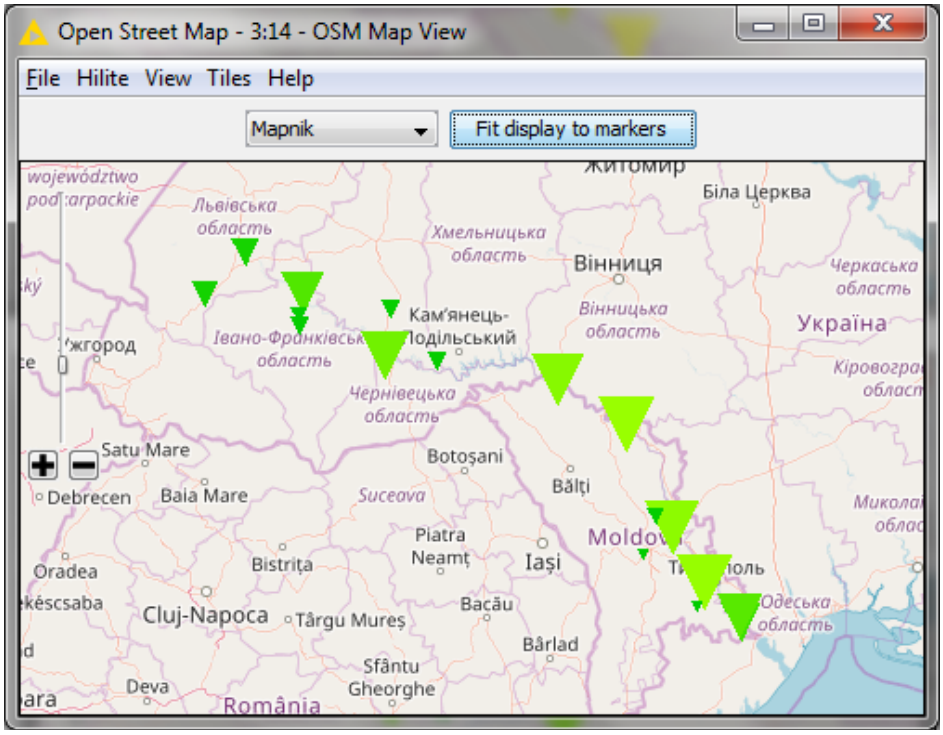


Fig. 5. Results of predicting average annual water consumption for 50 % provision of the Dnister river basin at the OpenStreet Map

The presented example illustrates operation of GIS-integrated system of data and models in KNIME Analytics Platform, proposed in Fig. 1. However, the analysis has shown that a powerful (at first sight) for this problem KNIME Analytics Platform software package has the following disadvantages: 1) it is impossible to use the regression analysis model for all gauging stations simultaneously; 2) it is impossible to write the regression analysis results for all gauging stations into one file or a table and to use them for computation simultaneously; 3) it is impossible to construct a map for several indicators simultaneously; 4) there are significant limitations as to the properties of the visualization methods. However, these problems could be removed by the use of Java, Python, Perl, R, SQL programming languages. The problem will be especially complicated if instead of the data on gauging stations, the basin flow map will be used. However, in this case the balance and the map, similar to those of Fig 5, could be constructed for any river of the basin.

## Conclusions

The paper has presented the conducted analysis of the known formats for formalization of mathematical models and spatial formats, which are relevant for processing data and models of the processes occurring in the rivers. The system of criteria is proposed and the example of its application for choosing an optimal XML format for storing data and models is presented. A structure of GIS-integrated system of data and models based on XML formalization is proposed for the first time and its operability is illustrated by the example of predicting average annual water consumption over a multiyear period for 50 % provision in the Dniester river basin in KNIME Analytics Platform environment. The obtained results make it possible to provide fast construction, versatility and broad functional of GIS-integrated system of data and models.

## REFERENCES

1. An integrated GIS-based tool for aquifer test analysis / R. Criollo, V. Velasco, E. Vázquez-Suñé [et al.] //

Environmental Earth Sciences. – February 2016. – № 75:391. – P. 1 – 11. – DOI:10.1007/s12665-016-5292-3.

2. A regional air quality forecasting system over Europe: the MACC-II daily ensemble production / M. Sofiev, V. Marecal, V.-H. Peuch [et al.] // Geoscientific Model Development. – 2015. – № 8. – P. 2777 – 2813.

3. Stryzhak O. Y. Means of ontological integration and support of distributed spatial and semantic informational resources / O. E. Strizhak // Ecological safety and nature management. – 2013. – № 12. – P. 166 – 177. (Ukr).

4. Chiao-Ling Kuo. Interoperable cross-domain semantic and geospatial framework for automatic change detection / Chiao-Ling Kuo, Jung-Hong Hong // Journal Computers & Geosciences. – January 2016. – Volume 86, Issue C. – P. 109 – 119 . – DOI 10.1016/j.cageo.2015.10.011.

5. Evaluating OpenMI as a model integration platform across disciplines / M. J. R. Knapen, S. J. C. Janssen [et al.] // Environmental Modelling & Software. – 2013. – № 39. – P. 274 – 282.

6. Yevdin Y. O. Development of the cross-platform version of JRODOS decision support system for radiation accidents / Y. O. Evdin, D. M. Tribushnyi, M.J. Zhelezniak // Mathematical Machines and Systems. – 2012. – № 1. – P. 45 – 59. (Ukr).

7. SBML Level 3 package: Hierarchical Model Composition, Version 1 Release 3 / L. P. Smith, M. Hucka [et al.] // Journal of Integrative Bioinformatics. – 2015. – № 12 (2):268. – P. 1 – 56. – DOI:10.2390/biecoll-jib-2015-268.

8. Methodological challenges and analytic opportunities for modeling and interpreting Big Healthcare Data / Ivo D. Dinov // GigaScience. – 25 February 2016. – № 5:12. – P. 1 – 15. – DOI 10.1186/s13742-016-0117-6.

9. Ensembles and PMML in KNIME / A. Fillbrunn, I. Adä, T. R. Gabriel [et al.] // PMML Workshop. Chicago. – Aug 10, 2013. – P. 1 – 6. – ISBN 978-1-4503-2573-8.

10. PMML Sample Files [Electronic resource]. – Mode of access : http://dmg.org/pmml/pmml_examples/index.html.

11. PMML Powered [Electronic resource]. – Mode of access: http://dmg.org/pmml/products.html.

12. BioModels Database [Electronic resource]. – Mode of access: http://biomodels.caltech.edu/.

13. 6 Worked Examples of Application Schemas (Non-Normative) [Electronic resource]. – Mode of access: http://etutorials.org/Mobile+devices/mobile+location+services/Appendix+B.+Geography+Markup+Language/6+Worked+Examples+of+Application+Schemas+Non-Normative/.

14. WaterOneFlow Web Services & WaterML [Electronic resource]. – Mode of access: http://his.cuahsi.org/wofws.html#waterml.

15. Web Based Access to Water Related Data UsingOGC WaterML 2.0 / A. Almoradie, A. Jonoski, I. Popescu [et al.] // International Journal of Advanced Computer Science and Applications, EnviroGRIDS Special Issue on «Building a Regional Observation System in the Black Sea Catchment». – 2013. – P. 83 – 89.

16. System analysis and design of GIS: electronic tutorial / Y. M. Kryzhanovskiy, V. B. Mokin, A. R. Yashcholt, L. M. Skoryna. – Vinnytsia: VNTU, 2015. – 127 p. – Mode of access: http://ir.lib.vntu.edu.ua/bitstream/handle/123456789/8960/Posibnik_2015_3%20(1).pdf?sequence=1. (Ukr).

17. Scientific principles of rational use of water resources of Ukraine on basin principle: monograph / [V. A. Stashuk, V. B. Mokin, V. V. Grebin, O. V. Chunarev]; under editorship of V. A. Stashuk. – Kherson: Grin D. S., 2014. – 320 p. (Ukr).

18. Automation of calculating the water balance of river basin areas / V. B Mokin, Ye. M. Kryzhanovsky, A. R. Yaschalt [and others.] // Water resources of Ukraine. – 2017. – № 3 (129). – P. 25 – 30. (Ukr).

Editorial office received the paper 30.03.2018.
The paper was reviewed 03.04.2018.

***Mokin Vitalii*** – Dc. Sc (Eng.), Prof., Head of the Department of System Analysis, Computer Monitoring and Engineering Graphics.

***Skoryna Liubov*** – Junior lecturer, post-graduate with the Department of System Analysis, Computer Monitoring and Engineering Graphics.

***Kryzhanovskyi Yevhenii*** – Cand. Sc. (Eng.), Ass. Prof. with the Department of System Analysis, Computer Monitoring and Engineering Graphics.
Vinnytsia National Technical University.

***Horash Mykola*** – software engineer.
Soft Generation Ltd., Vinnytsia.