GIS-oriented solutions for advanced clustering analysis of geoscience data using ArcGIS platform

A. A. Soloviev[1,2], J. I. Zharkikh[1], R. I. Krasnoperov[1], B. P. Nikolov[1], and S. M. Agayan[1]

Received 24 November 2016; accepted 1 December 2016; published 13 December 2016.

This paper presents software solutions for integration of geoscience data and data processing algorithms based on the Discrete Mathematical Analysis (DMA) in GIS environment. The DMA algorithms have been adapted and implemented within the ESRI ArcGIS software as geoprocessing tools and combined into a single set of tools named "Clustering". This set can be used along with the standard ArcGIS geoprocessing instruments. The tools of the "Clustering" set have also been published on the GIS-server as geoprocessing services providing powerful analytical functions via the Internet. This paper gives a brief outlook of the geoprocessing tools preparation techniques. The results of DMA-based geoprocessing tools' application to geophysical data are also discussed. *KEYWORDS:* Geoscience data analysis; cluster analysis; geoprocessing tools; geoprocessing services; web-oriented GIS.

Citation: Soloviev, A. A., J. I. Zharkikh, R. I. Krasnoperov, B. P. Nikolov, and S. M. Agayan (2016), GIS-oriented solutions for advanced clustering analysis of geoscience data using ArcGIS platform, *Russ. J. Earth. Sci., 16,* ES6004, doi:10.2205/2016ES000587.

Introduction

Efficient processing and analysis of georeferenced data require adequate instruments, which nowadays are provided by geoinformation systems (GIS), including a great variety of geoprocessing tools. In some cases, specific tasks require implementation of algorithms based on latest advances in data mining techniques. Modern GIS software such as ESRI ArcGIS allow designing new geoprocessing tools based on original mathematical algorithms, using the Python or ModelBuilder program languages. The newly designed tools, which are called "custom tools", along with the standard ones broaden the software geoprocessing functionality. Custom tools can be unified into a single package, which is easily transferrable. They can also be published on a GIS server enabling its further use in web applications, thus becoming accessible to a wide community. The presented results are the part of the ongoing development of the intellectual GIS for geoscience data analysis, carried out at the GC RAS [*Berezko et al.*, 2009a, 2009b, 2010, 2011a, 2011b; *Beriozko et al.*, 2007, 2008, 2011; *Gvishiani et al.*, 2007; *Krasnoperov and Soloviev*, 2015; *Krasnoperov et al.*, 2012; *Nikolov et al.*, 2015; *Soloviev et al.*, 2007]. This paper describes the results of adaptation of clustering algorithms based on the Discrete Mathematical Analysis (DMA) approach for their integration into GIS environment as new geoprocessing tools. The developed tools have been combined in a geoprocessing package "Clustering" implemented as the part of the geoprocessing toolbox set of the ESRI ArcGIS platform. The instruments of the created toolbox have also been published as geoprocessing services on the GIS server. The advantages of this approach are discussed in [*Nikolov et al.*, 2015].

[1]Geophysical Center RAS, Moscow, Russia
[2]Schmidt Institute of Physics of the Earth of the Russian Academy of Sciences (IPE RAS), Moscow, Russia

http://elpub.wdcb.ru/journals/rjes/doi/2016ES000587-res.html

Geoprocessing Algorithms

Mathematical algorithms used for the creation of geoprocessing tools and services are the part of the DMA approach developed at the Geophysical Center of the Russian Academy of Sciences (GC RAS). DMA represents a set of algorithms unified by a single formal basis aimed for studying and analysis of multidimensional arrays and time series. The series of DMA algorithms have been successfully implemented for analysis of various geological and geophysical spatial data [*Agayan and Soloviev*, 2004; *Agayan et al.*, 2010, 2011, 2014a, 2014b, 2016; *Bogoutdinov et al.*, 2010; *Gvishiani et al.*, 2002a, 2002b, 2008, 2010, 2011, 2013, 2014; *Mikhailov et al.*, 2003; *Soloviev et al.*, 2005, 2009, 2012a, 2012b, 2013, *Soloviev et al.,* Estimation of Geomagnetic Activity..., *Annals of Geophysics, 59*(6), in press; *Widiwijayanti et al.*, 2003; *Zelinskiy et al.*, 2014]. In the framework of the current research the following DMA-algorithms have been adapted for implementation as geoprocessing tools: "Discrete Perfect Sets" (DPS), its modification (DPSm), "Monolith" and "Rodin-2".

The "DPS" clustering algorithm was created for separation of dense regions with a certain density level α in a set of point features. Two main parameters that can be adjusted by the user are ω ($\omega < 0$) for determination of closeness radius and $\beta \in [0,1]$ for determination of density level. Determination of dense areas in the modified "DPSm" algorithm is performed generally in the same way as in the conventional DPS. The difference is in the method of determining the density of points [*Agayan et al.*, 2011, 2014a].

The "Monolith" algorithm is aimed at recognition of dense subsets of the elements in metric spaces with accordance to the level of density. The result of the algorithm application may contain isolated points. The algorithm parameters are the level of density $\alpha \in [0,1]$ and a negative parameter $\omega < 0$, which is required for determining the proximity radius.

The "Rodin-2" algorithm was designed for recognition of dense regions in finite metric spaces. It is based on the Kolmogorov mean and "fuzzy comparison" constructions. The "Rodin-2" free parameters include the proximity parameter $p \in \Re$ and the threshold parameter $\alpha \in [-1,1]$ [*Gvishiani et al.*, 2008, 2010; *Nikolov et al.*, 2015].

Integration of Algorithms Into GIS

The ESRI ArcGIS platform has been chosen as the GIS environment for integration of the DMA clustering algorithms as geoprocessing tools. It is an advanced applied geoinformation software platform, which provides the following functions: support of the most of the conventional GIS standards and formats; compatibility with other platforms, databases, development languages and applications; efficient server-based solutions.

The DMA clustering algorithms "DPS", "DPSm", "Monolith", and "Rodin-2" were programmed using Python 2.7, which is one of the main development languages used in ArcGIS. It is a non-proprietary language, which has many additional libraries that allow to expand significantly its functionality [*Zandbergen*, 2013]. Python is a popular language among GIS developers since it provides a large number of well-documented classes and functions, numerous specialized libraries (ArcPy, NumPy, etc.), ease of loading and running of scripts, a great variety of examples. Advantages of the Python language are:

- software quality (the code is easier to read than in other programming languages);
- high speed of development (a smaller size of a code, programs start directly);
- portability of programs (most of Python programs run without any change on all major platforms);
- large support library.

For geodata management a special library ArcPy was used. It includes the majority of GIS functions and facilitates greatly the process of scripts' compilation. ArcPy module allows to convert and analyze georeferenced data and automate mapping functions using Python scripts.

Another Python library that was used is the NumPy library, which is a free and powerful equivalent of MATLAB [*Lutz*, 2010]. NumPy is a fundamental package for scientific computing, which provides a great variety of capabilities such as operations with multi-dimensional array objects; tools for integrating C/C++ and Fortran code; capabilities of linear algebra, Fourier analysis and many others [NumPy: [web site] URL: http://www.numpy.org/].

Initially the clustering algorithms were programmed as Python scripts with certain functionality which were later combined as an ArcGIS set of geoprocessing tools. This allows using the developed tools along with the standard ones. Creation of a geoprocessing tool requires an initial script (in this case, an algorithm programmed in Python), an empty geoprocessing package, and the exact description of the parameters and the order of their execution in the compiled script.

On the next step, it is required to add all the developed scripts into the empty toolbox as separate tools. To do this the paths to each of the script's files are specified as well as the input parameters for each of them. On this stage, it is very important to designate the order in which the parameters are specified and determine their type. The main types, which are used for the input parameters of these algorithms are "Layer" (reference to a source of point data, such as a shapefile or geodatabase feature class), "Double" (any floating point number), "Long" (integer number value) and "Workspace" (geodatabase or folder that will contain the result of the geoprocessing). Finally, the toolbox (*.tbx) file, containing the clustering algorithms, is formed. It can be easily transferred via local computer networks or the Internet and distributed among users. After receiving a toolbox file, a user cannot change the tools' source code or their structure. It is available only for input of the tools' parameters while execution. To change the code of the tools, the user should know the password, set by the author of the toolbox.

To run the geoprocessing tools the user should utilize one of the components of the ArcGIS platform (ArcMap, ArcCatalog, etc.) with the ArcToolbox panel. Once the "Clustering" toolbox is available at the ArcToolbox panel (Figure 1), the user can apply its tools for specified geospatial data processing.

The "Clustering" geoprocessing toolbox has a simple, understandable and user-friendly interface that allows the user to run the required geoprocessing tool quite easily. After selecting the required tool a dialog window pops up. The appearance of the dialog window of the running tool is shown in Figure 2. The window provides fields for input of parameters, required for the tool execution (data layer path, free parameters, output layer path etc.), and a text field with a short description of the tool. The results of the tool execution are automatically stored in the specified location. If the user works in the ArcMap application the results are automatically added to the current map, and the user can view and analyze them along with the initial data.

After the execution of the algorithm a window with the status information and results pops up. It displays the values of the input parameters, data messages, the elapsed execution time, errors and warnings if they occur during the

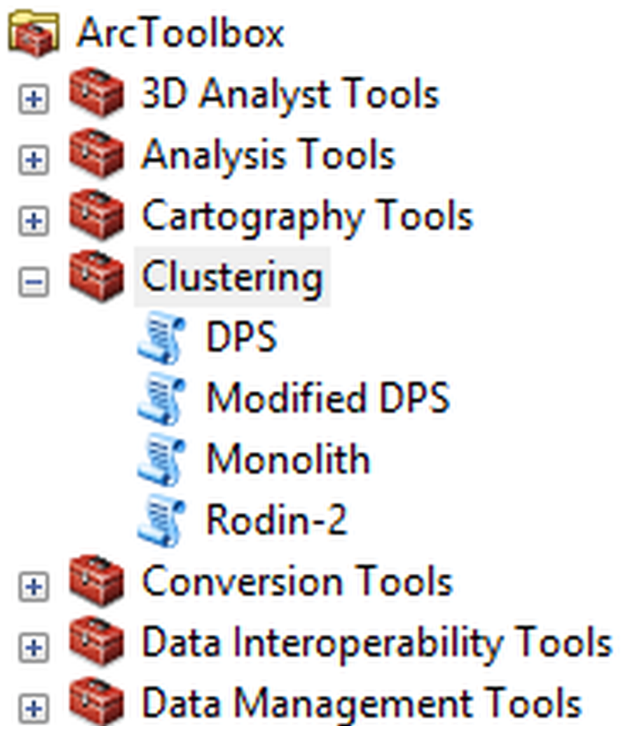

Figure 1. ArcToolbox geoprocessing toolbar with the added "Clustering" toolbox.

process (Figure 3a, Figure 3b).

All the above mentioned clustering algorithms are applicable only to point data, so in the input "Layer" field of the tool dialog window only data layers with point objects should be selected. If the user inputs incorrect data, an error message appears in the execution results window. An error message also appears if the user enters the tool parameter values that do not fit the specified requirements (e.g. free parameter is out of the range). The implemented geoprocessing instruments can be presented not only as toolbox packages for desktop software, but they also can be published on the GIS-server as geoprocessing services, which makes them available to the whole Internet community. The presented geoprocessing instruments were published on the GC RAS GIS-server as the part of the Centralized Catalog of Geodata Processing Algorithms (CCGPA). This subsystem is responsible for access to specific methods of geodata processing performed centrally on the GIS-server.

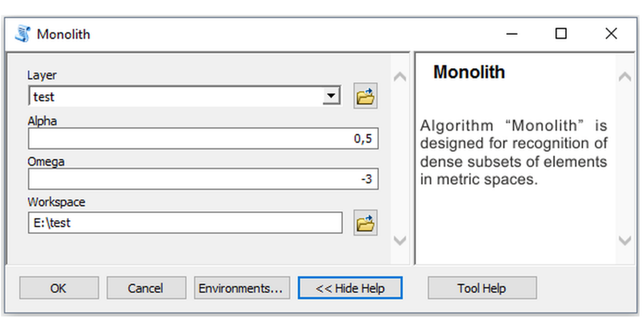

Figure 2. The dialog window of the "Monolith" tool.

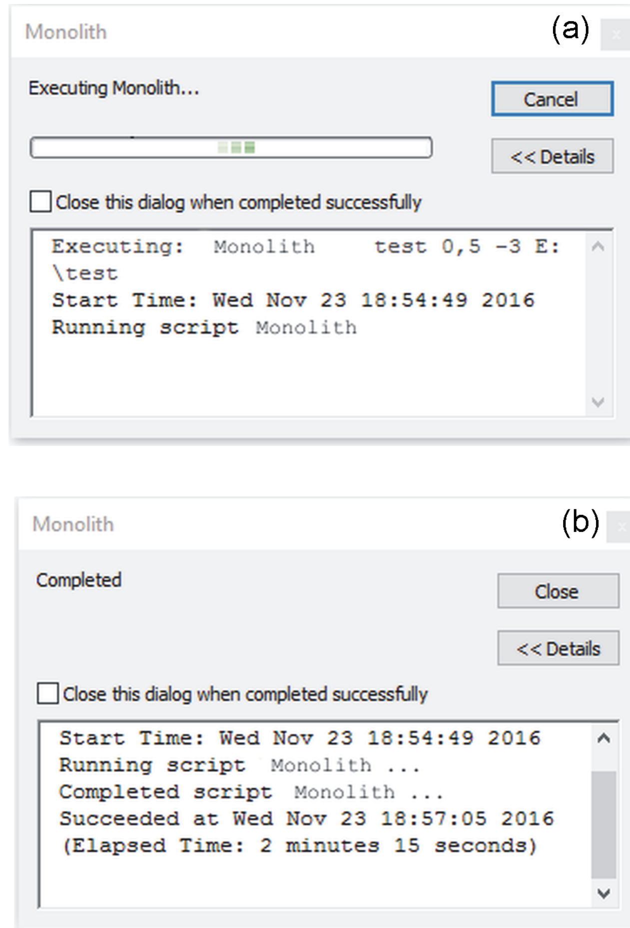

Figure 3. Pop-up window with the geoprocessing tool execution status: a) execution progress; b) successful execution results.

The principal advantages of this approach are [*Lebedev and Beriozko*, 2009]:

- minimum requirements for users' workstations (all calculations are performed on a server, user receives only the results);
- creation of a unified library of geoprocessing methods;
- online access to a comprehensive base of geoprocessing instruments based on the latest advances in data mining;
- facilitation of global exchange of knowledge.

Application, Results, Experiments

The implemented geoprocessing instruments were tested on both synthetic and real spatial datasets. A synthetic

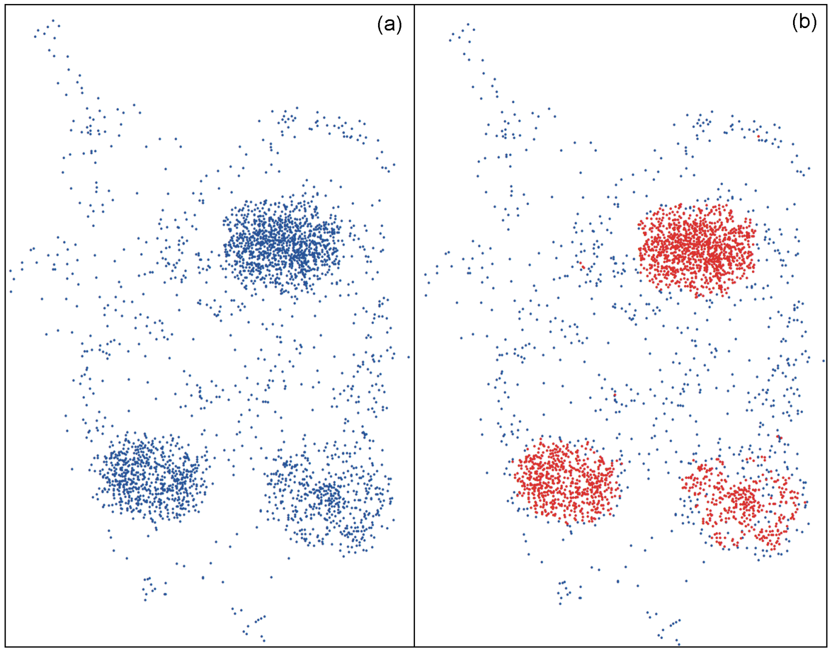

Figure 4. “Monolith” geoprocessing tool testing on a synthetic point layer: a) initial layer (blue dots); b) results of clustering (red dots) obtained with free parameters $\alpha = 0.5$ and $\omega = -3$ and superimposed on the initial data.

point dataset was generated and imported as a layer in the ArcMap software. The instruments of the “Clustering” toolbox were applied to this layer with different free parameters' values. One of the examples of the “Monolith” tool testing is presented in Figure 4. The results of clustering were saved as new layers, which can be used in further data analysis.

The algorithms published as geoprocessing services were also tested in online mode using seismological data from the online GIS on Earth Sciences, maintained by GC RAS (http://gis.gcras.ru/). The results are presented in Figure 5.

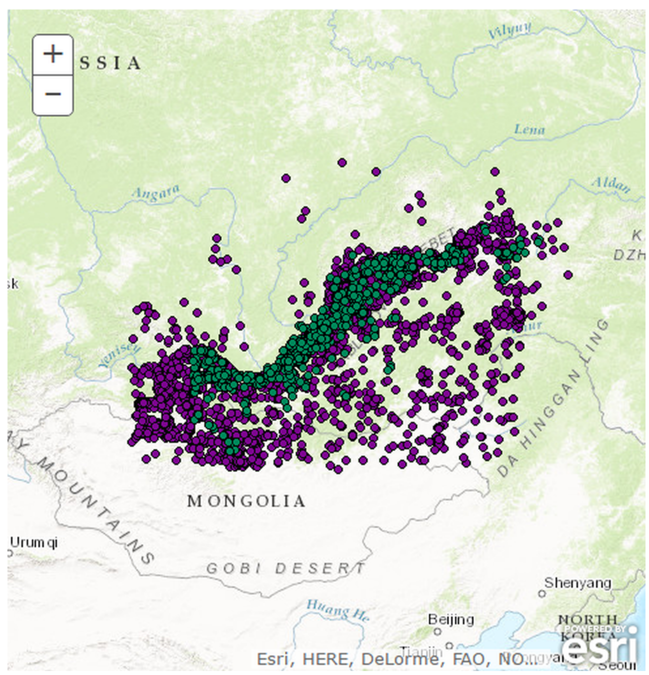

Figure 5. Results of the online clustering (green dots) of earthquake epicenters (purple dots) in the Baikal region using “Rodin-2” algorithm with free parameters $\alpha = 0.7$ and $P = -2.7$.

Conclusion

Adaptation of original clustering algorithms (“DPS”, “DPSm”, “Rodin-2”, and “Monolith”) as geoprocessing tools and services broadens the analysis apparatus for efficient management of geospatial data. The created “Clustering” toolbox can be easily distributed among the GIS community via the Internet or other digital media. The programmed algorithms are also published online as geoprocessing services thus giving an opportunity to analyze geoscience data without using commercial desktop software. Further theoretical and practical research will focus on algorithms, capable to process not only point data, but also linear and polygonal objects.

Acknowledgments. This paper is based on the report presented at the international conference “Data Intensive System Analysis for Geohazard Studies” held on 18–21 July, 2016 in Sochi, Russia. This work was carried out in the framework of the project of the Fundamental Research Program of the RAS Presidium No. 4 “Strategic mineral resources of Russia: innovative approaches to forecasting, assessment and extraction”.

References

Agayan, S. M., A. A. Soloviev (2004), Allocation of dense areas in metric spaces basing on crystallization, *System Research and Information Technologies*, No. 2, 7–23. (in Russian)

Agayan, S. M., Sh. R. Bogoutdinov, M. N. Dobrovolsky (2011), About one algorithm for searching the dense regions and its geophysical applications, *Reports of the 15th All-Russian conference "Mathematical methods for pattern recognition. MMPR"* p. 543–546, Maks Press, Moscow. (in Russian)

Agayan, S. M., Sh. R. Bogoutdinov, M. N. Dobrovolsky (2014a), Discrete Perfect Sets and their application in cluster analysis, *Cybernetics and Systems Analysis*, *50,* No. 2, 176–190.

Agayan, S. M., Sh. R. Bogoutdinov, M. N. Dobrovolsky, A. I. Kagan (2014b), Weighted gravitational time series smoothing, *Russian Journal of Earth Sciences*, *14,* ES3002, doi:10.2205/2014ES000543

Agayan, S., S. Bogoutdinov, A. Soloviev, R. Sidorov (2016), The Study of Time Series Using the DMA Methods and Geophysical Applications, *Data Science Journal*, *15,* 16, doi:10.5334/dsj-2016-016

Agayan, S. M., A. D. Gvishiani, S. R. Bogoutdinov, A. I. Kagan (2010), Sglazhivaniye vremennyh ryadov metodami diskretnogo matematicheskogo analiza (Smoothing of time series by the methods of Discrete Mathematical Analysis), *Russian Journal of Earth Sciences*, *11,* RE4001, (in Russian) doi:10.2205/2009ES000436

Berezko, A., A. Gvishiani, A. Soloviev, R. Krasnoperov, A. Rybkina, A. Lebedev (2010), Intellectual GIS "Earth Science Data for the Territory of Russia", *Problems of protection of population and territories against emergencies.* p. 210–218, EMERCOM, Moscow, Russia. (in Russian)

Berezko, A., A. Gvishiani, A. Soloviev, R. Krasnoperov, A. Rybkina, A. Lebedev (2011a), Multidisciplinary GIS in Earth sciences, *Applied problems in geology, geophysics, and geoecology in connection with modern information technologies. Proceedings of international conference. May 16–20, 2011* p. 37–44, Publishing house "Marinin O.G.", Maikop, Russia.

Berezko, A., A. Gvishiani, A. Soloviev, R. Krasnoperov, A. Lebedev, A. Rybkina (2011b), Geoinformation system for mineralogical research support, *"Mineralogical prospective". Proceedings of international mineralogical workshop. Syktyvkar, Russia, May 17–20, 2011* p. 19–21, Institute of Geology, Komi SC of URAS, Syktyvkar, Russia. (in Russian)

Berezko, A., A. Rybkina, A. Soloviev, R. Krasnoperov (2009a), Intellectual GIS, *Vestn. Otd. nauk Zemle RAN*, No. 1, NZ3002, doi:10.2205/2009NZ000006

Berezko, A., A. Soloviev, R. Krasnoperov, A. Rybkina (2009b), Intellectual analytical geoinformation system "Earth Science Data for the Territory of Russia", *Environment. Technology. Resources: Proceedings of the 7th International Scientific and Practical Conference, Rezekne, June 25–27, 2009.* 1, p. 215–221, Rezeknes Augstskola, Rezekne, RA Izdevnieciba, Rezekne, Latvia. (in Russian) doi:10.17770/etr2009vol1.1122

Beriozko, A., A. Lebedev, A. Soloviev, R. Krasnoperov, A. Rybkina (2011c), Geoinformation system with algorithmic shell as a new tool for Earth sciences, *Russian Journal of Earth Sciences*, *12,* ES1001, doi:10.2205/2011ES000501

Beriozko, A., A. Soloviev, R. Krasnoperov (2007), Representation of geological-geophysical data in a unified integrated GIS environment, *Russian Journal of Earth Sciences*, *9,* ES2001, doi:10.2205/2007ES000245

Beriozko, A. E., A. A. Soloviev, A. D. Gvishiani, E. A. Jalkovski, R. I. Krasnoperov, S. A. Smagin, E. S. Bolotsky (2008), Intellectual geoinformation system "Earth sciences data for the territory of Russia", *Engineering Ecology*, No. 5, 32–40. (in Russian)

Bogoutdinov, Sh. R., A.D. Gvishiani, S. M. Agayan, et al. (2010), Recognition of Disturbances with Specified Morphology in Time Series. Part 1: Spikes on Magnetograms of the Worldwide INTERMAGNET Network, *Izvestiya, Physics of the Solid Earth*, *46,* No. 11, 1004–1016.

Gvishiani, A. D., S. M. Agayan, Sh. R. Bogoutdinov (2002a), Mathematical Methods of Geoinformatics. I. A New Approach to Clusterization, *Cybernetics and Systems Analysis*, *38,* No. 2, 238–254.

Gvishiani, A. D., S. M. Agayan, S. R. Bogoutdinov, A. I. Kagan (2011), Gravitazionnoe sglazhivaniye vremennyh ryadov (Gravitational smoothing of time series), *Tr. IMM UrO RAN*, *17,* No. 2, 62–70. (in Russian)

Gvishiani, A. D., S. M. Agayan, Sh. R. Bogoutdinov, et al. (2010), Discrete mathematical analysis and applications geology and geophysics, *Bulletin of KRAESC. Earth Sciences*, No. 10, 109–125. (in Russian)

Gvishiani, A., S. Agayan, Sh. Bogoutdinov, J. Zlotnicki, J. Bonnin (2008), Mathematical methods of geoinformatics. III. Fuzzy comparisons and recognition of anomalies in time series, *Cybernetics and System Analysis*, *44,* No. 3, 309–323, doi:10.1007/s10559-008-9009-9

Gvishiani, A. D., B. A. Dzeboev, S. M. Agayan (2013), A new approach to recognition of the strong earthquake-prone areas in the Caucasus, *Izvestiya. Physics of the Solid Earth*, *49,* No. 6, 747–766, doi:10.1134/S1069351313060049

Gvishiani, A., R. Lukianova, A. Soloviev, A. Khokhlov (2014), Survey of Geomagnetic Observations Made in the Northern Sector of Russia and New Methods for Analysing Them, *Surveys in Geophysics*, *35,* No. 5, 1123–1154, doi:10.1007/s10712-014-9297-8

Gvishiani, A. D., V. O. Mikhailov, S. M. Agayan, et al. (2002b), Artificial intelligence algorithms for magnetic anomaly clustering, *Izvestiya. Physics of the Solid Earth*, *38,* No. 7, 545–559.

Gvishiani, A., A. Soloviev, A. Beriozko (2007), Development and creation of integral geoinformation analytical system "Earth Science Data for the Territory of Russia", *IST4Balt News Journal*, No. 3, 38–40.

Krasnoperov, R., A. Lebedev, O. Pyatygina, A. Rybkina, A. Shibaeva (2012), Multidisciplinary analytical GIS for processing and visualization of remote sensing data, *Contemporary problems of space remote sensing of the Earth*, *9,* No. 3, 50–54. (in Russian)

Krasnoperov, R. I., A. A. Soloviev (2015), Analytical geoinformation system for integrated geological-geophysical research in the territory of Russia, *Gornyi Zhurnal (Mining Journal)*, *10,* 89–93, (in Russian) doi:10.17580/gzh.2015.10.16

Lebedev, A. Yu., A. E. Beriozko (2009), Development of centralized catalog of geophysical data processing algorithms, *Russian Journal of Earth Sciences*, *11,* RE2002, (in Russian) doi:10.2205/2009ES000399

Lutz, M. (2010), *Learning Python, Fourth Edition*, 1216 pp., O'Reilly Media, Inc., California, USA.

Mikhailov, V., A. Galdeano, M. Diament, A. Gvishiani, S. Agayan, Sh. Bogoutdinov, E. Graeva, P. Sailhac (2003), Application of artificial intelligence for Euler solutions clustering, *Geophysics*, *68,* No. 1, 168–180, doi:10.1190/1.1543204

Nikolov, B. P., J. I. Zharkikh, A. A. Soloviev, R. I. Krasnoperov, S. M. Agayan (2015), Integration of data mining methods for Earth science data analysis in GIS environment, *Russian Journal of Earth Sciences*, *15,* ES4004, doi:10.2205/2015ES000559

Soloviev, A. A., S. M. Agayan, A. D. Gvishiani, et al. (2012a), Recognition of Disturbances with Specified Morphology in Time Series: Part 2. Spikes on 1-s Magnetograms, *Izvestiya. Physics of the Solid Earth*, *48,* No. 5, 395–409.

Soloviev, A. A., A. E. Berezko, A. D. Gvishiani, E. A. Zhalkovsky, S. M. Agayan (2007), Creation of an integral geoinformation analytical system "Earth sciences data for the territory of Russia", *"Problems of sustainable of natural and technogenic resources of the Barents region in the technology of construction and technical materials". Proceedings of the III International Conference* p. 247–249, Institute of Geology, Komi SC of URAS, Syktyvkar. (in Russian)

Soloviev, A. A., Sh. R. Bogoutdinov, S. M. Agayan, et al. (2009), Detection of hardware failures at INTERMAGNET observatories: application of artificial intelligence techniques to geomagnetic records study, *Russ. J. Earth Sci.*, *11,* ES2006, doi:10.2205/2009ES000387

Soloviev, A., S. Bogoutdinov, A. Gvishiani, R. Kulchinskiy, J. Zlotnicki (2013), Mathematical Tools for Geomagnetic Data Monitoring and the INTERMAGNET Russian Segment, *Data Science Journal*, *12,* WDS114–WDS119, doi:10.2481/dsj.WDS-019

Soloviev, A., A. Chulliat, S. Bogoutdinov, et al. (2012b), Automated recognition of spikes in 1 Hz data recorded at the Easter Island magnetic observatory, *Earth Planets Space*, *64,* No. 9, 743–752, doi:10.5047/eps.2012.03.004

Solov'ev, A. A., D. Yu. Shur, A. D. Gvishiani, V. O. Mikhailov, S. A. Tikhotskii (2005), Determination of the magnetic moment vector using cluster analysis of the local linear pseudoinversion of Delta T anomalies, *Doklady Earth Sciences*, *404,* No. 7, 1068–1071.

Widiwijayanti, C., V. Mikhailov, M. Diament, et al. (2003), Structure and evolution of the Molucca Sea area: constraints based on interpretation of a combined sea-surface and satellite gravity dataset, *Earth and Planetary Science Letters*, *215,* 135–150, doi:10.1016/S0012-821X(03)00416-3

Zandbergen, P. A. (2013), *Python scripting for ArcGIS*, 353 pp., Esri Press, California, USA. (ISBN 9781589482821)

Zelinskiy, N. R., N. G. Kleimenova, O. V. Kozyreva, et al. (2014), Algorithm for recognizing Pc3 geomagnetic pulsations in 1-s data from INTERMAGNET equatorial observatories, *Izvestiya. Physics of the Solid Earth*, *50,* No. 2, 240–248, doi:10.1134/S106935131402013X

S. M. Agayan, R. I. Krasnoperov, B. P. Nikolov, A. A. Soloviev, J. I. Zharkikh, Geophysical Center RAS, Molodezhnaya St. 3, 119296, Moscow, Russia. (a.soloviev@gcras.ru)