

information. The semistructured information is understood as information in which a certain structure can be identified, but this structure is completely unknown in advance or may change with time. With a long period of the enterprise's work, a large number of electronic documentation with semistructured content is accumulated, as a rule, represented by MS Word documents and minimization of its storage volume is an actual task. The described method of storing electronic documents is to minimize the amount of storage of electronic documents with semistructured content by allocating a single template for processing documents, extracting factual data from documents, and then compressing the template and factographic data. The restoring the document is done by extracting from the archive the factual data of the form template and specified content of the document, and applying the form template to the newly received document content. The method provides compression of electronic documents with semistructured content up to 22 times more efficient than standard archivers.

Keywords: semistructured data; information compression; electronic documents processing

DOI: 10.17587/it.25.53-58

References

1. **Palej D.** *Modelirovanie kvazistrukturirovannykh dannykh* (Simulation of quasi-structured data), *Otkrytye Sistemy*, 2002, no. 9, pp. 57–64 (in Russian).
2. **Chernikov B. V.** *Tehnologiya hranenija slaboformalizuemyykh dokumentov na osnove leksikologicheskogo sinteza* (Storage technology of poorly formalized documents based on lexicological synthesis), *Informatika i ee Primenenie*, 2009, vol. 3, iss. 4, pp. 64–75 (in Russian).
3. **Chernikov B. V.** *Sposob preobrazovaniya slaboformalizuemyykh dokumentov dlya minimizatsii ih obzema pri hranenii* (The method of converting poorly formalized documents to minimize their volume during storage), Patent na izobretenie, reg. № 2413985 ot 10.03.2011, Moscow, Rospatent, 2011 (in Russian).
4. **Polishhuk Ju. V., Chernyh T. A.** *Sintez kvazistrukturirovannykh modelej informatsionnogo napolneniya elektronnykh dokumentov* (Synthesis of quasi-structured models of informational filling of electronic), *Vestnik Komp'yuternykh i Informatsionnykh Tekhnologiy*, 2012, no. 6, pp. 20–27 (in Russian).
5. **Polishhuk Ju. V.** *Sposob formirovaniya kvazistrukturirovannykh modelej faktogra-ficheskogo informatsionnogo napolneniya dokumentov* (The method of forming quasi-structured models of factual content of documents), Patent na izobretenie, reg. № 2517428 ot 28.03.2014, Moscow, Rospatent, 2014 (in Russian).
6. **Polishhuk Ju. V., Larin A. V.** *Avtomatizatsiya processa razmetki upravlencheskogo kontenta jelek-tronnykh dokumentov s kvazistrukturirovannym informatsionnym napolneniem na osnove patternov* (Automating the process of marking management content of electronic documents with quasi-structured content based on patterns), *Vestnik komp'yuternykh i informatsionnykh tekhnologiy*, 2013, no. 3, pp. 55–60 (in Russian).
7. **Vouter V. V.** *Open XML — Kratko i dostupno* (Open XML — Brief and Available), Open XML Technical Evangelist, Microsoft, 2007, 101 p. (in Russian).
8. **Polishhuk Ju. V., Polishhuk P. V.** *Sposob preobrazovaniya dokumentov dlya minimizatsii ih obzema pri hranenii elektronnykh dokumentov s kvazistrukturirovannym informatsionnym napolneniem* (The method of converting documents to minimize their volume when storing electronic documents with quasi-structured information content), Patent na izobretenie, reg. № 2625611 ot 17.07.2017, Moscow, Rospatent, 2017 (in Russian).

УДК 004.421

DOI: 10.17587/it.25.58-63

Т. О. Перемитина, канд. техн. наук, доц., e-mail: peremitinat@mail.ru,

И. Г. Яценко, канд. геол.-минер. наук, e-mail: sric@ipc.tsc.ru,

Федеральное государственное бюджетное учреждение науки Институт химии нефти СО РАН, Томск

Алгоритм комплексного анализа многомерных данных об экологическом состоянии окружающей среды

Рассматриваются методические вопросы реализации и практического применения комплексного подхода к анализу многомерных данных об экологическом состоянии пространственно-распределенных объектов исследования. Подход основан на сочетании метода главных компонент и метода пространственного анализа с применением геоинформационных систем. Метод главных компонент применяется для статистической обработки и анализа данных, метод пространственного анализа используется для учета пространственных свойств исследуемых объектов. В статье приведены результаты анализа многомерных данных о состоянии природной среды и здоровья населения территорий Сибирского и Дальневосточного федеральных округов.

Ключевые слова: многомерные данные, метод главных компонент, геоинформационные системы, окружающая среда, экология

Введение

Мониторинг состояния окружающей среды сопряжен с обработкой постоянно увеличивающегося массива разрозненной совокупной информации об объектах исследования. Аналитические процессы предполагают обработку

вающегося массива разрозненной совокупной информации об объектах исследования. Аналитические процессы предполагают обработку

огромного количества материалов и включают такие трудоемкие процессы, как систематизация, выявление каких-либо закономерностей и преобразование данных.

Широко известен и используется для анализа многомерных данных метод главных компонент (МГК) [1–3]. Однако для анализа объектов исследования, имеющих выраженные пространственные свойства, применение МГК оказывается недостаточным, так как он не позволяет учитывать их пространственные характеристики. В связи с этим возникла необходимость разработки нового комплексного подхода к анализу многомерных данных, направленного на учет пространственных свойств объектов исследования.

Анализ экологического состояния окружающей природной среды, выявление причин загрязнения и прогнозирование дальнейшего состояния являются актуальными и важными задачами, поскольку именно качество природных объектов во многом определяет состояние здоровья людей, которые являются лишь частью природных экосистем [4].

В данной работе обсуждается комплексный анализ многомерных данных о состоянии природной среды и здоровья населения территории Сибирского и Дальневосточного Федеральных округов (ФО) и его практическое применение.

Методические вопросы реализации комплексного анализа

Основная идея комплексного анализа заключается в сочетании МГК и метода пространственного анализа с применением геоинформационных систем (ГИС). МГК применяется для статистической обработки различных количественных характеристик, а методом пространственного анализа исследуются пространственные свойства объектов.

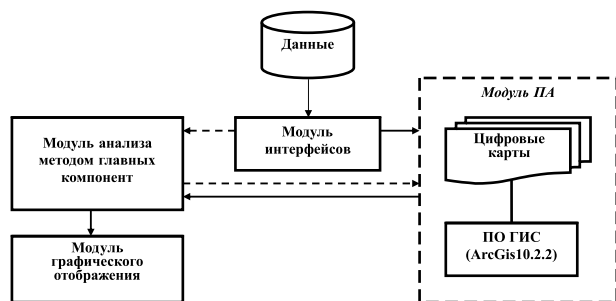


Рис. 1. Схема комплексного анализа многомерных данных на основе сочетания методов главных компонент и пространственного анализа

На рис. 1 приведена обобщенная схема проведения комплексного анализа многомерных данных о пространственно-распределенных объектах.

Модуль анализа МГК включает (рис. 1):

- процедуры нормирования и стандартизации исходного массива данных;
- вычисление ковариационной матрицы;
- нахождение собственных значений и собственных векторов ковариационной матрицы методом вращения Якоби;
- вычисление главных компонент;
- вычисление вклада каждой главной компоненты в общую дисперсию;
- анализ вкладов главных компонент в общую дисперсию и отбор главных компонент для проведения дальнейшего анализа;
- вычисление проекций исходных наблюдений (нагрузок) на отобранные главные компоненты;
- табличное представление нагрузок на отобранные главные компоненты.

Модуль пространственного анализа (ПА) данных включает:

- подготовку атрибутивной таблицы — создание массива исследуемых данных средствами ArcGis 10.2.2;
- отображение объектов на цифровых картах — создание цифрового слоя исследуемых объектов;
- выявление пространственных группирований объектов — в атрибутивной таблице каждому объекту приписывается номер класса, к которому он отнесен.

Модуль графического отображения включает:

- графическое отображение объектов исследования в пространстве двух главных компонент;
- графическое представление нагрузок на две главные компоненты;
- векторное представление нагрузок в пространстве двух главных компонент;
- графическое представление результатов анализа с отображением средних значений и доверительных областей.

Модуль интерфейсов предназначен для осуществления взаимодействия между такими разнородными частями системы, как модуль пространственного анализа (ПО ГИС) и модуль МГК-анализа (среда разработки Delphi).

Результаты комплексного анализа

В данном исследовании был проведен комплексный анализ состояния природной среды

Перечень субъектов РФ, включенных в анализ

Федеральный округ	№	Классификатор объектов административно-территориального деления
Сибирский федеральный округ	1	Алтайский край
	2	Забайкальский край
	3	Иркутская область
	4	Кемеровская область
	5	Красноярский край
	6	Новосибирская область
	7	Омская область
	8	Республика Алтай
	9	Республика Бурятия
	10	Республика Тыва
	11	Республика Хакасия
	12	Томская область
Дальневосточный федеральный округ	13	Амурская область
	14	Еврейская автономная область
	15	Камчатский край
	16	Магаданская область
	17	Приморский край
	18	Республика Саха (Якутия)
	19	Сахалинская область
	20	Хабаровский край
	21	Чукотский автономный округ

и здоровья населения территорий Сибири и Дальнего Востока. В таблице приведен перечень субъектов Российской Федерации, включенных в анализ. Открытые литературные источники и официальный сайт Федеральной службы государственной статистики [5] позволили собрать следующие количественные показатели о состоянии природной среды и здоровья населения исследуемых субъектов РФ в 2015 г. (в скобках указаны единицы измерения):

- 1) нарушено земель в связи с несельскохозяйственной деятельностью (тыс. га);
- 2) нелесные земли, пройденные пожарами (га);
- 3) лесные земли, пройденные пожарами (га);
- 4) территория с погибшим молодняком (га);
- 5) территория со сгоревшим лесом на корню (м^3);
- 6) объем сброса загрязненных сточных вод по бассейнам рек и озер (млрд м^3);
- 7) объем загрязненных вод из сброшенных сточных вод (млн м^3);
- 8) количество образованных отходов производства и потребления за отчетный год (тыс. т);
- 9) количество выброшенных в атмосферу загрязняющих веществ, отходящих от стационарных источников (тыс. т);

10) количество выброшенных без очистки загрязняющих веществ за отчетный год организованными источниками загрязнения (тыс. т);

11) количество выброшенных загрязняющих веществ в атмосферу (тыс. т);

12) количество выброшенных без очистки загрязняющих веществ (тыс. т);

13) заболеваемость населения злокачественными новообразованиями на 100 тысяч населения (человек);

14) заболеваемость населения на 100 тысяч населения, с диагнозом, установленным впервые (человек);

15) заболеваемость населения, взятого под наблюдение в отчетном году с впервые установленным диагнозом на 100 тысяч населения (человек).

Возможности сайта Федеральной службы государственной статистики не позволили наглядно отобразить значения количественных показателей для всех исследуемых субъектов. В связи с отсутствием возможности отобразить в виде графика, гистограммы или картографического отображения по 15 количественным показателям о состоянии природной среды и здоровья населения 21 субъекта РФ была использована одна из функций сайта — выгрузка (экспорт) многомерных данных (рис. 2). Выбранные данные были выгружены (экспортированы) в виде таблицы "объект—свойство", где представлены объекты исследования — субъекты РФ (21 строка) и их свойства — количественные характеристики природных ресурсов и здоровья населения (15 столбцов).

На первом шаге проведения комплексного анализа многомерных данных проводился МГК-анализ статистических данных. Основ-

Территория	Выборы в атмосферу загрязняющих веществ, отходящих от стационарных источников (тыс. тонн)	Выборы без очистки наиболее распространенных загрязняющих веществ за отчетный год (тыс. тонн)
Алтайский край	255.159	25.829
Амурская область	91.752	16.083
Иркутская область	504.93	34.422
Камчатский край	46.821	9.072
Кемеровская область	1225.182	54.017
Красноярский край	2589.226	29.354
Магаданская область	36.817	5.976
Приморский край	351.936	16.399
Новосибирская область	198.049	18.572
Омская область	238.349	13.877
Сахалинская область	100.003	12.014
Томская область	254.598	24.747

Рис. 2. Возможности сайта Федеральной службы государственной статистики

ная идея МГК состоит в переходе к новой системе координат и оценке важности, информативности каждой из новых переменных — главных компонент. При проведении МГК-анализа вычисляются ортогональные векторы (главные компоненты), учитывающие максимальную долю дисперсии исходного распределения. Поскольку исходными данными для нахождения главных компонент являются выборочные коэффициенты корреляции, то очевидно, что необходимо проверить значимость, достоверность полученных результатов. Разработанный программный комплекс анализа многомерных данных (см. рис. 1) включает процедуру оценки информативности главных компонент, на основании чего можно говорить о степени целесообразности перехода к новым системам координат (в пространство главных компонент) и проведения дальнейшего МГК-анализа.

Для сформированной выборки результаты МГК-анализа показали, что первая (F1) и вторая (F2) главные компоненты описывают более 90 % общей дисперсии, что позволяет перейти к двумерному пространству главных компонент. Для изучения особенностей исследуемых объектов необходимо проанализировать значения нагрузок на главные компоненты.

Матрица нагрузок — это матрица перехода из исходного пространства переменных (15-мерного) в пространство главных компонент (двумерное). График нагрузок применяется для исследования роли переменных, их информативности и значимости [6]. Графическое представление нагрузок на главные компоненты позволило оценить влияние каждого количественного показателя анализируемого субъекта РФ на их распределение в пространстве главных компонент и определить набор наиболее информативных параметров. Анализ нагрузок на первую главную компоненту показал, что для данного набора показателей наиболее информативными (вариабельными) являются всего пять из пятнадцати количественных показателей состояния природной среды и здоровья населения субъектов РФ. Таким образом, в дальнейший анализ необходимо включать только эти пять характеристик, остальные можно не принимать в рассмотрение в силу их малой вариабельности от объекта к объекту. Алгоритм комплексного анализа позволил исключить из анализа незначимые характеристики, и для анализа состояния природной среды и здоровья населения Сибири, Дальнего Востока были отобраны пять наиболее значимых количественных показателей:

1) количество выброшенных в атмосферу загрязняющих веществ, отходящих от стационарных источников (тыс. т);

2) количество выброшенных без очистки загрязняющих веществ (тыс. т);

3) количество выброшенных без очистки загрязняющих веществ за отчетный год организованными источниками загрязнения (тыс. т);

4) объем загрязненных вод из сброшенных сточных вод (млн м³);

5) заболеваемость населения злокачественными новообразованиями на 100 тысяч населения (человек).

На рис. 3 показано взаимное распределение всех исследуемых субъектов РФ в двумерном пространстве главных компонент F1 и F2.

Из рис. 3 видно, что максимальными значениями первой главной компоненты обладают Красноярский край (5) и Кемеровская область (4), что говорит о высоких значениях всех количественных показателей состояния природной среды и состояния здоровья населения для данных субъектов.

На рис. 4 представлены значения нагрузок пяти количественных показателей для 21 субъекта РФ, где (по горизонтальной оси) 1 — количество выброшенных в атмосферу загрязняющих веществ, отходящих от стационарных источников (тыс. т); 2 — количество

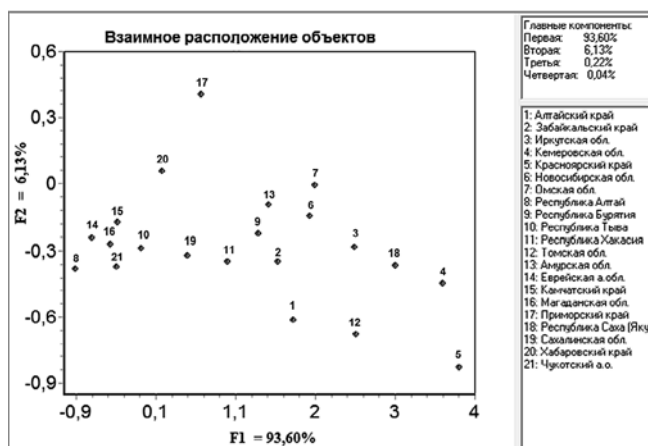


Рис. 3. Результаты МГК-анализа

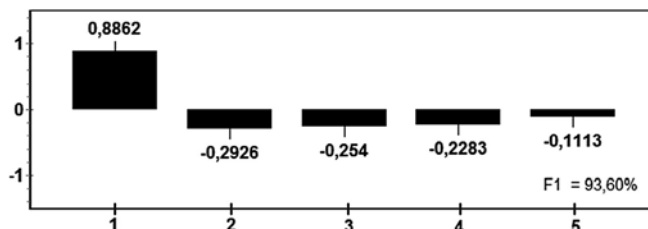


Рис. 4. Нагрузки на первую главную компоненту

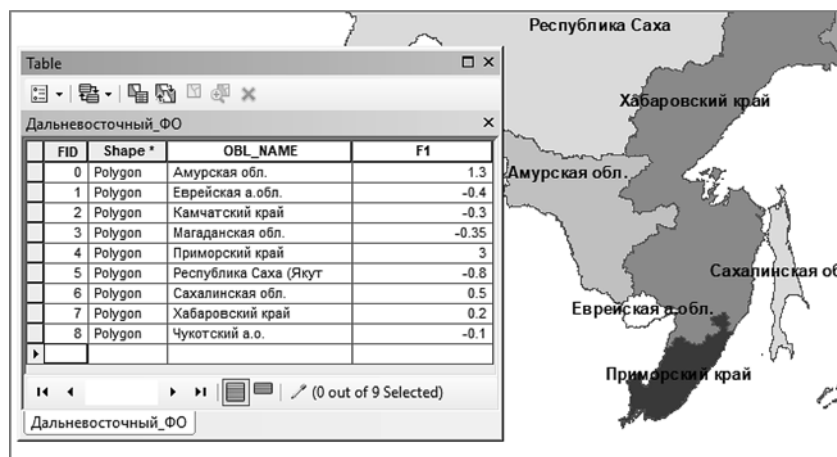


Рис. 5. Картирование результатов МГК-анализа данных о природной среде и состоянии здоровья населения субъектов РФ

выброшенных без очистки загрязняющих веществ (тыс. т); 3 — количество выброшенных без очистки загрязняющих веществ за отчетный год организованными источниками загрязнения (тыс. т); 4 — из сброшенных сточных вод — объем загрязненных (млн м³); 5 — заболеваемость населения злокачественными новообразованиями на 100 тысяч населения (человек).

Из рис. 4 видно, что наиболее значимым параметром для 21 субъекта РФ является характеристика — "количество выброшенных в атмосферу загрязняющих веществ, отходя-

щих от стационарных источников (тыс. т)" с положительным значением 0,8862. Нагрузки для 2-, 3- и 4-го показателей практически одинаковы и имеют отрицательные значения от -0,2926 до -0,2283. Самая малая нагрузка с отрицательным значением -0,1113 имеет характеристика "заболеваемость населения злокачественными новообразованиями на 100 тысяч населения (человек)".

Средствами геоинформационной системы ArcGis 10.2.2 результаты МГК-анализа были добавлены (экспортированы) в атрибутивную таблицу [7] исследуемых субъектов РФ (рис. 5).

На рис. 6 показано картирование результатов МГК-анализа и ранжирование субъектов РФ Сибирского ФО по значению первой главной компоненты, которая характеризует пять количественных параметров состояния природной среды и здоровья населения. На рис. 6 наглядно отображено, что максимальные значения первой главной компоненты имеют субъекты Красноярский край и Кемеровская область, минимальные значения — Республика Алтай и Республика Тыва.

Пространственный анализ показал, что на каждой исследуемой территории (Сибирский ФО и Дальневосточный ФО) есть субъекты, отличающиеся максимальными значениями первой главной компоненты. Наиболее загрязненными является территории Красноярского края, Кемеровской области и Приморского края.

Заключение

При описании природных объектов в задачах мониторинга состояния окружающей среды и природопользования, как правило, используется большой объем пространственной информации. Традиционный метод главных компонент не позволяет учитывать пространственные свойства объектов, что требует разработки методических, алгоритмических и программных средств его использования в сочетании с методами пространственного анализа, реализуемыми с помощью ГИС-технологий.

Применение разработанного подхода к обработке многомерных данных о состоянии природной среды и здоровья населения Сибирского и Дальневосточного ФО позволило проанализи-

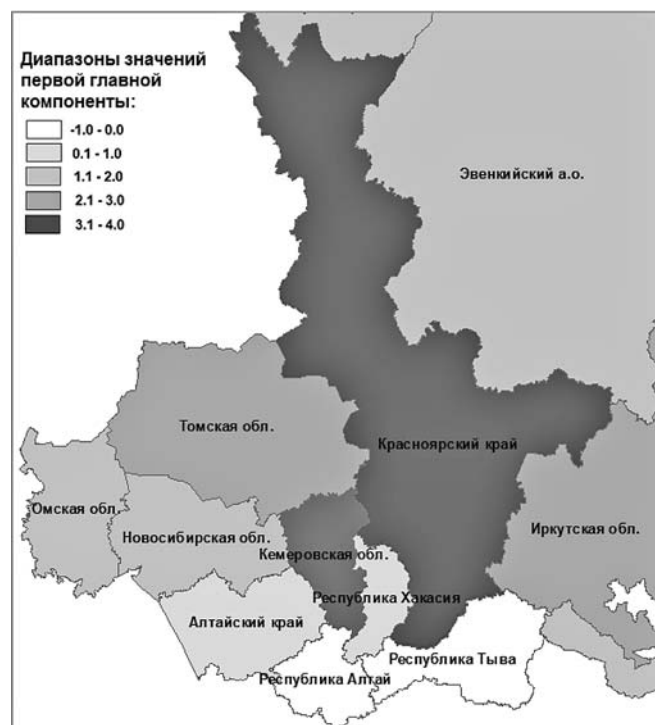


Рис. 6. Картирование результатов МГК-анализа

зировать особенности исследуемых территорий. Применение МГК позволило выявить наиболее значимые характеристики для территорий, а пространственный анализ дал возможность наглядно отобразить результаты МГК-анализа и выявить экологически благополучные и неблагополучные регионы.

Список литературы

1. Abson D. J., Dougill A. J., Stringer L. C. Using Principal Component Analysis for information-rich socio-ecological vulnerability mapping in Southern Africa // *Applied Geography*. 2012. Vol. 35, Iss. 1—2. P. 515—524. DOI: <https://doi.org/10.1016/j.envpol.2005.10.028>
2. Андрукович П. Ф. Применение метода главных компонент в практических исследованиях. М.: Изд-во МГУ, 1973. Вып. 36. 124 с.
3. Zhang C. Using multivariate analyses and GIS to identify pollutants and their spatial patterns in urban soils in Gal-

way, Ireland // *Environmental Pollution*. 2006. Vol. 142, Iss. 3. P. 501—511. DOI: <https://doi.org/10.1016/j.apgeog.2012.08.004>

4. Суслов М. И., Булкина К. А., Бубнов А. Г., Буймова С. А., Царев Ю. В. Статистический анализ данных экологического мониторинга родниковых вод урбанизированных территорий // *Южно-Сибирский научный вестник*. 2016. № 3(15). С. 15—19.

5. Федеральная служба государственной статистики: Окружающая среда. URL: http://www.gks.ru/wps/wcm/connect/rosstat_main/rosstat/ru/statistics/environment/ (дата обращения: 22.06.2018).

6. Лучкова С. В., Перемитина Т. О., Ященко И. Г. Использование нечеткого моделирования для повышения репрезентативности информации на примере анализа характеристик нефти // *Информационные технологии*. 2015. Т. 21, № 5. С. 336—340.

7. Brewer C. A. Basic Mapping Principles for Visualizing Cancer Data Using Geographic Information Systems (GIS) // *American Journal of Preventive Medicine*. 2006. Vol. 30, Iss. 2. P. S25—S36. DOI: <https://doi.org/10.1016/j.amepre.2005.09.007>

T. O. Peremitina, Ph.D. in Technical Science, e-mal: peremitinat@mail.ru,
I. G. Yashchenko, Ph.D. in Geology and Mineralogy, e-mal: sric@ipc.tsc.ru,
Institute of Petroleum Chemistry, Siberian Branch of the Russian Academy of Sciences

Algorithm for Complex Analysis of Environment Ecological State

There is considered methodical questions of realization and practical application of the complex approach to the analysis of multidimensional data on the ecological state of spatially distributed objects of research. The approach is based on a combination of the method of principal components and the method of spatial analysis using geoinformation systems. The principal components analysis (PCA) is used for statistical processing and data analysis, the method of spatial analysis is used to take into account the spatial properties of the objects under study. The article presents the results of the analysis of multidimensional data on the state of the natural environment and the health of the population of the territories of the Siberian and Far Eastern Federal Districts.

Keywords: multidimensional data, principal component analysis, geographic information systems, environment, ecology

DOI: 10.17587/it.25.58-63

References

1. Abson D. J., Dougill A. J., Stringer L. C. Using Principal Component Analysis for information-rich socio-ecological vulnerability mapping in Southern Africa, *Applied Geography*, 2012, vol. 35, iss. 1—2, pp. 515—524, DOI: <https://doi.org/10.1016/j.envpol.2005.10.028>
2. Andrukovich P. F. *Primenenie metoda glavnnykh komponent v prakticheskikh issledovaniyakh* (Application of principal components analysis (PCA) in practical research), Moscow, Publishing house of MGU, 1973, iss. 36, 124 p. (in Russian).
3. Zhang C. Using multivariate analyses and GIS to identify pollutants and their spatial patterns in urban soils in Galway, Ireland, *Environmental Pollution*, 2006, vol. 142, iss. 3, pp. 501—511, DOI: <https://doi.org/10.1016/j.apgeog.2012.08.004>
4. Suslov M. I., Bulkina K. A., Bubnov A. G., Bujmova S. A., Caryov Yu. V. *Statisticheskij anali z dannyh ehkologicheskogo monitoringa rodnikovykh vod urbanizirovannykh territorij* (Statistical

analysis of data of environmental monitoring of spring water in urban areas), *YUzhno-Sibirskij nauchnyj vestnik*, 2016, no. 3(15), pp. 15—19 (in Russian).

5. *Federal'naya sluzhba gosudarstvennoj statistiki: Okruzhayushchaya sreda* (Federal state statistics service), available at: http://www.gks.ru/wps/wcm/connect/rosstat_main/rosstat/ru/statistics/environment/ (data obrashcheniya: 22.06.2018) (in Russian).

6. Luchkova S. V., Peremitina T. O., Yashchenko I. G. *Ispol'zovanie nechetkogo modelirovaniya dlya povysheniya reprezentativnosti informacii na primere analiza harakteristik nefii* (Usage fuzzy modelling for raising a sample representative on the example the analysis of characteristics oils), *Informacionnye Tekhnologii*, 2015, vol. 21, no. 5, pp. 336—340 (in Russian).

7. Brewer C. A. Basic Mapping Principles for Visualizing Cancer Data Using Geographic Information Systems (GIS), *American Journal of Preventive Medicine*, 2006, vol. 30, iss. 2, pp. S25—S36. DOI: <https://doi.org/10.1016/j.amepre.2005.09.007>