УДК 348

А.Ю. Трусова, А.И. Ильина*

ВИЗУАЛИЗАЦИЯ И КЛАССИФИКАЦИЯ ПОКАЗАТЕЛЕЙ ЭКОЛОГИИ САМАРСКОГО РЕГИОНА

В работе средствами многомерного статистического анализа изучены показатели экологии Самарского региона. Визуализация данных проводилась с помощью компонентного анализа. Классификация проводилась методом к-средних.

Ключевые слова: многомерный статистический анализ, факторный и компонентный анализы, кластерный анализ, метод к-средних.

В настоящее время показатели экологии изучаются и анализируются многопланово различными методами химии, физики, биологии. Особое место в анализе данных занимают методы визуализации и классификации показателей. Широкий математический инструментарий в сочетании с информационными технологиями позволяет комплексно рассматривать проблемы экологии в их связи с техническими и экономическими проблемами.

Самарская область исторически является зоной промышленного производства. В Самарской области находится значительное количество предприятий, которые оказывают сильное экологическое воздействие на окружающую атмосферу. Экономическое развитие Самарской области предполагает развитие нефтеперерабатывающей отрасли, которое также способствует ухудшению экологической обстановки в регионе. Комплексное решение проблемы экономического развития региона и решения экологических проблем, связанных с развитием нефтеперерабатывающей отрасли является актуальным и практически значимым**.** Для решения перечисленных проблем необходимо сочетание и комплексное применение наук: биологии, химии, физики, математики, экономики и других для поддержания стабильности экологической ситуации в Самарском регионе. В этой связи, в данной работе рассматриваются математические и информационные подходы к изучению проблемы анализа существующих в настоящее время показателей, описывающих экологическую ситуацию в регионе. В работе средствами многомерного анализа изучены показатели, характеризующие количество и качество выбросов в атмосферу. Министерством лесного хозяйства, охраны окружающей среды и природопользования Самарской области ведется контроль за экологической ситуацией в Самарском регионе. Исходные данные для анализа представлены на сайте данного министерства.

Многообразие многомерных статистических методов позволяет, в первую очередь, их визуализировать и классифицировать. Для визуализации данных в работе используются методы факторного анализа и многомерного шкалирования, классификация проводилась методами кластерного анализа. Математический аппарат данных методов широко представлен в научной литературе. Ниже представлен краткий обзор используемых методов.

Многомерное шкалирование (МШ) позволяет решать различные проблемы в научных исследованиях самого широкого спектра. Независимо от типа решаемой задачи МШ используется как инструмент наглядного представления (визуализации) исходных данных. Поиск координатного пространства в МШ осуществляется не по значениям самих характеризующих объекты признаков, а по данным, представляющим различия или сходство этих объектов. Анализ индивидуальных различий является мощным математическим инструментом среди разнообразных методов многомерного шкалирования.

В работе методом МШ изучается модель индивидуальных различий. Основополагающим является предположение, что полученные в ходе подгонки модели оценки ее параметров хорошо воспроизводят скалярные произведения:

Трусова Алла Юрьевна (a_yu_ssu@mail.ru), *Ильина Алла Ивановна* (iai.62@mail.ru), кафедра математики и бизнес-информатики, Самарский национальный исследовательский университет имени академика С.П. Королева, 443086, Российская Федерация, г. Самара, Московское шоссе, 34.

$$\delta_{ijs}^{*} = \sum_{k} x_{ik} x_{jk} \omega_{ks}^{2} = \sum_{k} x_{iks} x_{jks}$$

или в матричном виде: $\Delta_{s}^{*} = X W_{s}^{2} X^{T}$.

Стартовая конфигурация матрицы координат стимулов формируется методом главных компонент, который является частью факторного анализа.

В современной трактовке факторный анализ – это совокупность методов, в которых на основе реально существующих связей признаков, осуществляется выявление неявных обобщающих характеристик. С помощью факторного анализа возможно выявление скрытых переменных факторов, отвечающих за наличие линейных статистических корреляций между наблюдаемыми переменными. Таким образом, факторный анализ позволяет определить взаимосвязи между переменными и сократить число переменных, необходимых для описания данных.

В факторном анализе латентные факторы объединяют тесно связанные между собой переменные. В результате перераспределения дисперсии между компонентами получается максимально простая и наглядная структура факторов. В целом факторный анализ позволяет выделить из всей совокупности переменных небольшое число латентных независимых друг от друга группировок, внутри которых переменные связаны сильнее, чем переменные, относящиеся к разным группировкам. В частности, метод главных компонент – один из основных способов уменьшить размерность данных, потеряв наименьшее количество информации. Вычисление главных компонент сводится к вычислению собственных векторов и собственных значений корреляционной матрицы исходных данных. Формирование однородных групп осуществляется в работе средствами кластерного анализа, а именно методом к-средних.

Министерством лесного хозяйства, охраны окружающей среды и природопользования Самарской области ведется комплексное изучение хозяйственной деятельности предприятий на экологию региона, регулярно осуществляются измерения показателей, данная деятельность находит свое отражение в данных федеральной службы государственной статистики. Исходный массив для анализа представлен в таблице 1.

Таблица 1

Исходный массив данных

t	X_1	$\mathbf{X_2}$	$\mathbf{X_3}$	X_4	X_5	Y_1	Y_2
2004	739	31997	23027	853	342	4021	372
2005	716	31707	22697	807	310	5019	679
2006	725	33573	22943	906	322	5648	986
2007	751	34787	23657	782	311	224	1426
2008	830	38044	26081	599	296	5330	2300
2009	878	39901	26898	595	278	5312	1249
2010	879	40809	27195	675	298	5902	1230
2011	909	41821	28251	747	283	7294	2298
2012	965	43055	28861	724	265	7249	2851
2013	1197	46682	30923	695	252	8307	3551
2014	1249	52048	31101	769	257	8796	5916

В качестве показателей в анализе выбраны следующие: X_1 – количество объектов, имеющих выбросы загрязняющих веществ (единиц); X_2 – количество источников выбросов загрязняющих ве-

ществ, всего; X_3 – количество организованных источников выбросов загрязняющих веществ; X_4 – количество загрязняющих веществ, отходящих от всех источников выделения (Выбросы и улавливание загрязняющих атмосферу веществ, отходящих от стационарных [1)] (тысяч тонн); X_5 – количество загрязняющих веществ, отходящих от всех источников выделения без очистки; Y_1 – всего текущих затрат на охрану окружающей природы (в фактически действовавших ценах; миллионов рублей); Y_2 – инвестиции в основной капитал, направленные на охрану окружающей среды. Все данные указаны за период с 2004-2014 года. Матрица корреляций представлена в таблице 2.

Таблица 2

Матрица корреляций

R	x_1	x_2	x_3	x_4	x_5	y_1	y_2
x_1	1	0,966	0,9489	-0,313	-0,88	0,754	0,917335
x_2	0,966	1	0,9751	-0,401	-0,91	0,747	0,919611
x_3	0,949	0,975	1	-0,491	-0,93	0,773	0,849888
x_4	-0,31	-0,4	-0,491	1	0,541	-0,18	-0,20384
x_5	-0,88	-0,91	-0,934	0,5406	1	-0,7	-0,80848
y_1	0,754	0,747	0,7729	-0,184	-0,7	1	0,662494
y_2	0,917	0,92	0,8499	-0,204	-0,81	0,662	1

Используя пакет SPSSStatistika, мы провели факторный анализ, многомерное шкалирование и кластерный анализ. В результате использования метода главных компонент были выделены два главных фактора, методом варимаксного вращения были получены улучшенные компоненты матрицы факторного отображения, представленные в таблице 3.

Таблица 3

Метод выделения: Анализ методом главных компонент,
метод вращения: Варимакс с нормализацией Кайзера

Матрицы	Матрица компонент		Матрица повернутых компонент		
	Компонента		Компонента		
R	F_1	F_2	*R*	F_1	F_2
X_1	0,968	0,139	X_1	0,959	-0,188
X_2	0,983	0,047	X_2	0,944	-0,28
X_3	0,985	-0,057	X_3	0,911	-0,379
X_4	-0,456	0,881	X_4	-0,14	0,982
X_5	-0,947	0,152	X_5	-0,844	0,456
Y_1	0,805	0,252	Y_1	0,843	-0,027
Y_2	0,905	0,241	Y_2	0,934	-0,071

В таблице 4 представлена статистика меры адекватности выделения двух компонент, которая свидетельствует о достаточности выделенных двух главных компонент. В таблице 5 представлены общности выделенных компонент и полная объясненная дисперсия.

Таблица 4

Мера адекватности и критерий Бартлетта

Наименование	Мера выборочной адекватности Кайзера-Мейера-Олкина	0,835
Критерий сферичности Бартлетта	Прибл. хи-квадрат	82,816
	Ст. св.	21
	Знач.	0,000

Таблица 5

Общности выделенных двух компонент. Полная объясненная дисперсия

Общности			Полная объясненная дисперсия			
R	Начальные	Извлеченные	Компонента	Итого	% дисперсии	Кумулятивный %
X_1	1	0,956	F_1	5,444	77,775	_
X_2	1	0,969	F_2	0,946	13,511	_
X_3	1	0,973	F_3	0,381	5,436	96,723
X_4	1	0,984	F_4	0,104	1,491	98,213
X_5	1	0,921	F_5	0,082	1,173	99,386
Y_1	1	0,711	F_6	0,033	0,474	99,86
Y_2	1	0,877	F_7	0,01	0,14	100

На рис. 1 представлены изучаемые показатели в пространстве двух главных компонент после варимаксного вращения.

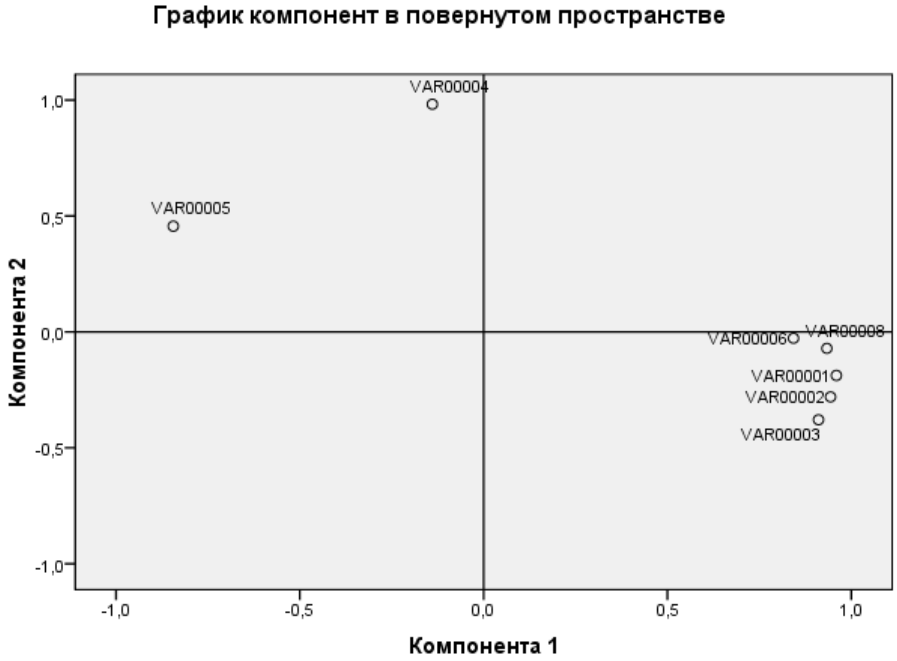

Рис. 1. Изучаемые показатели в пространстве латентных факторов

На рис. 2 представлены временные периоды в пространстве латентных факторов без вращения.

Таким образом, средствами факторного анализа многомерные данные представлены в двумерном пространстве латентных факторов. Данное представление позволяет глубже проследить изменение в экологических показателях Самарского региона.

Классификация данных осуществлялась методом к-средних кластерного анализа. Временной промежуток от 2004 до 2009 года характеризуется схожестью показателей. В таблице 6 представлены данные о принадлежности к кластерам изучаемые временные промежутки.

Следующий период выделяется по однородным показателям с 2009 по 2013 год. В этот период наблюдается изменение показателей экологии в сторону их улучшения. Особо выделяется 2014 год, который можно представить как отдельный кластер. Центрами классов являются 2005 и 2013 годы, показатели этих периодов можно рассматривать в качестве основных для принятия взвешенных решений. В таблицах 7 и 8 представлены характеристики кластеров и расстояние между кластерами.

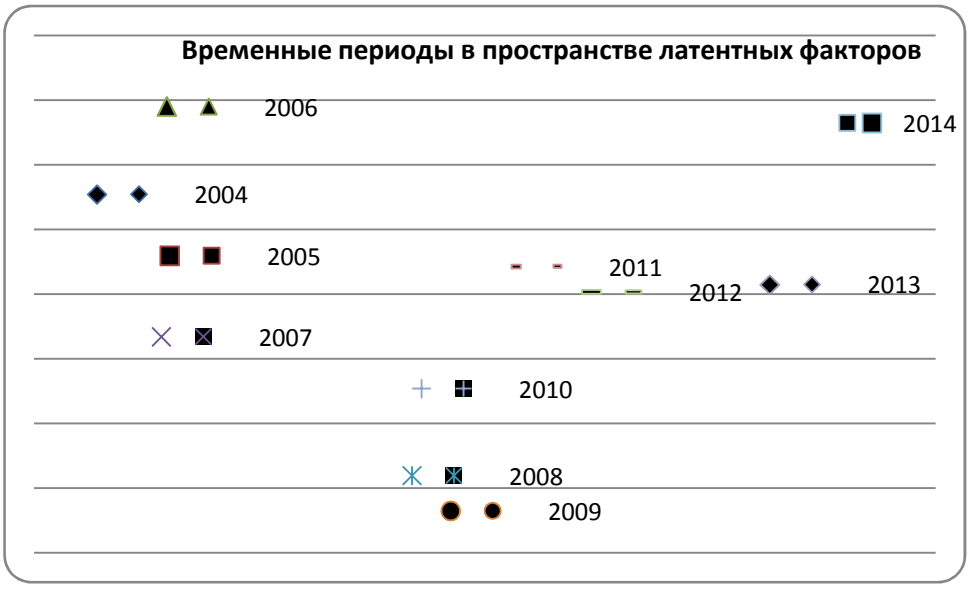

Рис. 2. Временные периоды в пространстве латентных факторов без вращения

Таблица 6

Принадлежность к кластерам

Год	Кластер	Расстояние
2004	2	1134,906
2005	2	0
2006	2	2020,791
2007	2	5835,024
2008	2	7383,974
2009	3	8755,173
2010	3	7738,405
2011	3	5812,256
2012	3	4399,826
2013	3	0
2014	1	0

Таблица 7

Конечные центры кластеров

R	Кластер		
	1	2	3
X_1	1249	752,2	965,6
X_2	52048	34022	42454
X_3	31101	23681	28426
X_4	769	789,4	687,2
X_5	257	316,2	275,2
X_6	8796	4048,4	6812,8
Y_1	1070	1587,6	1312,8
Y_2	5916	1152,6	2235,8

Таблица 8

Расстояния между конечными центрами кластеров

Кластер	1	2	3
1	0	20634	10809
2	20634	0	10127
3	10809	10127	0

Как видно из таблиц, характеризующих параметры кластеров и расстояния между ними, изучаемые временные промежутки можно рассматривать как однородные структурные объекты.

В результате визуализации средствами многомерного шкалирования получен график расположения годов с 2004 по 2014 год в двумерном шкальном пространстве. Номер на рисунке 3 соответствует номеру года: 1 – 2004, 2 – 2005 и т. д. Умеренное распределение показателей экологии в двумерном пространстве латентных факторов позволяет сделать вывод об определенной стабильности этих показателей или их незначительное изменение.

Рис. 3. Временные периоды в двумерном шкальном пространстве

Таким образом, в результате визуализации многомерных данных средствами компонентного анализа и многомерного шкалирования имеется возможность более детального изучения показателей экологии. Кластеризация временных периодов позволяется глубже анализировать однородные по структуре показатели экологии.

Библиографический список

1. Дубров А.М., Мхитарян В.С. Многомерные статистические методы. М.: Финансы и статистика, 1998. 338 с.
2. Наследов А.Д. SPSS 15: профессиональный статистический анализ данных. СПб: Питер, 2008. 320 с.
3. Сошникова Л. А., Тамашевич П. А. Многомерный статистический анализ в экономике. М.: Юнити, 1999. 320 с.
4. Трусова А.Ю., Сизова, Орлова И.С. Факторный анализ как средство визуализации многомерных данных // Вычислительные системы и информационные технологии: межвуз. сб. Самара. 2009, С. 60–65.

5. Сошникова Л.А., Тимашевич В.Н., Уебе Г., Шефер М. Многомерный статистический анализ в экономике: учеб. пособие для вузов. М.: ЮНИТИ-ДАНА, 1999.

6. Айвазян С.А., Мхитарян В.С. Прикладная статистика. Основы эконометрики: в 2 т.: учебник для вузов. М.: ЮНИТИ-ДАНА, 2001.

7. Трусова А.Ю., Тетерин А.Е. Сжатие социологической информации средствами факторного анализа // Труды Второй Всероссийской ФАМ'2003 конференции (Красноярск, 28 февраля – 2 марта, 2003 г.). Красноярск, 2003. С. 230–233.

8. Трусова А.Ю., Макарова И.С. Математическое моделирование социальных процессов // Образовательные технологии: межвуз. сб. науч. тр. Вып. 10. Воронеж, 2003. С. 87–91.

References

1. Dubrov A.M., Mkhitaryan V.S. Mnogomernyye statisticheskiye metody [Multivariate statistical methods]. M.: Finansy i statistika, 1998, 338 p.

2. Nasledov A.D. SPSS 15: professional'nyy statisticheskiy analiz dannykh [SPSS 15: professional statistical analysis of data]. SPb: Piter, 2008, 320 p.

3. Soshnikova L. A., Tamashevich P. A. Mnogomernyy statisticheskiy analiz v ekonomike [Multidimensional statistical analysis in economics]. M.: Yuniti, 1999. 320 p.

4. Trusova A.Yu., Sizova, Orlova I.S. Faktornyy analiz kak sredstvo vizualizatsii mnogomernykh dannykh [Factor analysis as a means of visualization of multidimensional data]. In: Vychislitel'nyye sistemy i informatsionnyye tekhnologii: mezhvuz. sb. [Computational systems and information technologies]. Samara. 2009, pp. 60–65.

5. Soshnikova L.A., Timashevich V.N., Uyebe G., Shefer M. Mnogomernyy statisticheskiy analiz v ekonomike: ucheb. posobiye dlya vuzov [Multivariate statistical analysis in economics: textbook. manual for universities]. M.: YUNITI-DANA, 1999.

6. Ayvazyan S.A., Mkhitaryan V.S. Prikladnaya statistika. Osnovy ekonometriki: v 2 t.: uchebnik dlya vuzov [Applied statistics. Fundamentals of econometrics: in 2 t: a textbook for universities]. M.: YUNITI-DANA, 2001.

7. Trusova A.Yu., Teterin A.E. Szhatiye sotsiologicheskoy informatsii sredstvami faktornogo analiza [Compression of sociological information by means of factor analysis]. In: Trudy Vtoroy Vserossiyskoy FAM'2003 konferentsii (Krasnoyarsk, 28 fevralya – 2 marta, 2003 g.) [Proceedings of the Second All-Russian FAM'2003 conference (Krasnoyarsk, February 28 – March 2, 2003)]. Krasnoyarsk, 2003, pp. 230–233.

8. Trusova A.Yu., Makarova I.S. Matematicheskoye modelirovaniye sotsial'nykh protsessov [Mathematical modeling of social processes]. In: Obrazovatel'nyye tekhnologii: mezhvuz. sb. nauch. tr. [Educational technologies: interuniversity. Sat. sci. tr.]. Issue. 10. Voronezh, 2003, pp. 87–91.

A.Yu. Trusova, A.I. Ilyina*

VISUALIZATION AND CLASSIFICATION OF INDICATORSECOLOGY OF THE SAMARA REGION

In the work of multidimensional statistical analysis, the environmental indicators of the Samara region were studied. Data visualization was carried out using component analysis. Classification was carried out by the method of k-means.

Key words: multidimensional statistical analysis, factor and component analysis, cluster analysis, k-means method.

Статья поступила в редакцию 14/IX/2016.
The article received 14/IX/2016.

* *Trusova Alla Yuriyevna* (a_yu_ssu@mail.ru), *Ilyina Elena Alekseevna* (elenaalex.ilyina@yandex.ru), Department of Mathematics and Business Informatics, Samara National Research University, 34, Moskovskoye shosse, Samara, 443086, Russian Federation.