УДК 311

В.Б. Яковлев

Снижение размерности данных в региональной статистике российского образования

В статье рассматривается методика применения факторного анализа для снижения размерности данных в региональной статистике российского образования с помощью статистического пакета SPSS 14.0 for Windows на примере обеспеченности высшими образовательными учреждениями Приволжского федерального округа Российской Федерации.

Ключевые слова: статистика российского образования; методы многомерного анализа; факторный анализ.

При анализе и прогнозировании показателей развития образования исследователь довольно часто сталкивается с многомерностью их описания.

В многомерном статистическом анализе каждый объект описывается с помощью вектора произвольного размера. При этом можно легко проанализировать данные, расположенные на плоскости (двумерный вектор). Анализ же данных в трехмерном пространстве уже вызывает затруднение, а в пространстве более высокого порядка он просто невозможен. Поэтому при исследовании данных необходим переход от многомерной выборки к выборке меньшей размерности.

Чаще всего для снижения размерности используют факторный анализ. С помощью данного статистического метода исходные переменные сводятся к меньшему количеству независимых величин, которые называют факторами. Принцип метода основан на том, что в один фактор объединяются переменные, сильно коррелирующие между собой. В результате дисперсия перераспределяется между факторами и получается максимально простая и наглядная их структура.

Порядок выполнения факторного анализа следующий. Вначале осуществляют стандартизацию исходных значений переменных и между рассматриваемыми переменными рассчитывают коэффициенты корреляции Пирсона. Затем выполняют тест на целесообразность проведения факторного анализа. После на основе корреляционной матрицы определяют собственные значения

и соответствующие им собственные векторы. Далее собственные значения сортируют в порядке убывания. При этом обычно отбирают факторы, имеющие собственные значения, которые по своей величине превосходят единицу. Собственные векторы, соответствующие этим собственным значениям, образуют факторы, элементы которых называют факторными нагрузками. Они представляют собой коэффициенты корреляции между соответствующими переменными и факторами. После того как факторы найдены и истолкованы, на последнем шаге факторного анализа, отдельным наблюдениям присваивают значения этих факторов (факторные значения). В результате для каждого наблюдения значения большого количества переменных переводят в значения небольшого количества факторов.

Для решения такого рода задач разработаны многочисленные методы факторного анализа, доступные во всех профессиональных статистических пакетах обработки данных: SPSS, SAS, R, Statistica и др. Так, в статистическом пакете SPSS 14.0 for Windows реализованы следующие методы:

- анализ главных компонент;
- метод невзвешенных наименьших квадратов;
- обобщенный метод наименьших квадратов;
- метод максимального правдоподобия;
- факторизация главных осей;
- альфа;
- анализ образов.

Из них чаще всего применяют метод анализа главных компонент, считающийся наиболее простым и универсальным методом.

Метод главных компонент является методом выделения факторов (компонент), который используют для формирования некоррелированных линейных комбинаций наблюдаемых переменных. При этом первая компонента имеет максимальную дисперсию. Другие же получаемые компоненты объясняют все меньшие доли дисперсии. Причем выделенные компоненты не коррелируют между собой. Обычно метод главных компонент используют для получения начального факторного решения.

Метод главных компонент был применен для анализа обеспеченности регионов Приволжского федерального округа РФ высшими образовательными учреждениями в 2014 г. (данные Федеральной службы государственной статистики URL: http://www.gks.ru/wps/wcm/connect/rosstat_main/rosstat/ru/statistics/publications/catalog/doc_1135087342078). Были отобраны восемь показателей (табл. 1). Расчеты проведены с помощью статистического пакета SPSS 14.0 for Windows.

Таблица 1

Обеспеченность регионов Приволжского федерального округа РФ высшими образовательными учреждениями в 2014 г.

Регион	Число образовательных организаций (2014/2015 уч. г.), x_1	Число филиалов образовательных организаций (2014/2015 уч. г.), x_2	Численность студентов, тыс. чел. (2014/2015 уч. г.), x_3	Прием в студенты, тыс. чел. (2014 г.), x_4	Выпуск студентов, тыс. чел. (2014 г.), x_5	Численность студентов на 10 000 жителей, чел. (2014/2015 уч. г.), x_6	Численность профессорско-преподавательского персонала, чел. (2014/2015 уч. г.), x_7	Численность профессорско-преподавательского персонала на 1000 студентов, чел. (2014/2015 уч. г.), x_8
Республика Башкортостан	11	34	126,7	27,7	28,4	311	6464	51
Республика Марий Эл	3	5	20,4	4,9	5,4	297	1091	53
Республика Мордовия	3	7	31,9	6,9	7,2	395	1997	63
Республика Татарстан	25	45	170,1	40,6	38,3	441	9143	54
Удмуртская Республика	7	14	52,5	13,8	13,4	346	2474	47
Чувашская Республика	5	16	42,6	9,9	12,0	344	2210	52
Пермский край	13	25	72,4	17,1	16,1	275	4089	56
Кировская область	6	15	38,5	8,4	9,4	295	1633	42
Нижегородская область	13	40	111,1	25,3	29,0	340	6289	57
Оренбургская область	6	22	62,3	12,9	13,6	312	3051	49
Пензенская область	4	12	42,9	8,8	9,3	317	2291	53
Самарская область	26	22	119,1	27,1	28,0	371	6727	56
Саратовская область	7	20	90,8	21,1	20,3	364	5574	61
Ульяновская область	5	10	42,5	9,5	10,0	336	2226	52

При проверке целесообразности выполнения факторного анализа были использованы два критерия: адекватности выборки Кайзера – Мейера – Олкина (КМО) и сферичности Бартлетта (табл. 2). Первый критерий позволяет проверить, насколько корреляция между парами переменных объясняется другими переменными (факторами), второй проверяет нулевую гипотезу об отсутствии корреляций между переменными в генеральной совокупности.

Таблица 2

KMO и критерий Бартлетта

Мера адекватности выборки Кайзера – Майера – Олкина (КМО)		0,797
Критерий сферичности Бартлетта	Примерная Хи-квадрат	187,918
	Степень свободы	28
	Значимость	1,294E-25

Результаты проведения теста КМО позволили сделать вывод о пригодности имеющихся данных для факторного анализа, поскольку значение меры адекватности превышает 0,5. Это подтверждает и критерий сферичности Бартлетта. Его значимость практически равна нулю, что свидетельствует о том, что между переменными исходных данных существуют корреляционные связи и поэтому возможна их группировка.

В результате дальнейших расчетов была получена матрица компонент, преобразованная с помощью ее вращения по критерию «варимакс». Данный метод позволяет улучшить факторную структуру матрицы, то есть в максимально возможной мере увеличить факторные нагрузки по одним показателям за счет уменьшения нагрузок по другим. Были выделены две компоненты, имеющие собственное число более единицы. На долю этих компонент приходится 89,2 % суммарной дисперсии (табл. 3).

Таблица 3

Матрица компонент

Показатели	**Компонента после вращения методом «варимакс»**	
	U_1	U_2
Число образовательных организаций (2014/2015 уч. г.), x_1	**0,871**	0,191
Число филиалов образовательных организаций (2014/2015 уч. г.), x_2	**0,942**	–0,008
Численность студентов, тыс. чел. (2014/2015 уч. г.), x_3	**0,975**	0,196
Прием в студенты, тыс. чел. (2014 г.), x_4	**0,973**	0,210

Показатели	Компонента после вращения методом «варимакс»	
	U_1	U_2
Выпуск студентов, тыс. чел. (2014 г.), x_5	**0,977**	0,186
Численность студентов на 10 000 жителей, чел. (2014/2015 уч. г.), x_6	0,387	**0,726**
Численность профессорско-преподавательского персонала, чел. (2014/2015 уч. г.), x_7	**0,947**	0,294
Численность профессорско-преподавательского персонала на 1000 студентов, чел. (2014/2015 уч. г.), x_8	–0,005	**0,907**
Суммарная дисперсия	5,928	1,205
% суммарной дисперсии	74,1	15,1

Из таблицы 3 видно, что первая компонента U_1 наиболее тесно связана с абсолютными показателями, такими как число образовательных организаций, число филиалов образовательных организаций, численность студентов, прием в студенты, выпуск студентов, численность профессорско-преподавательского персонала, поэтому первую компоненту целесообразно назвать «уровень абсолютной обеспеченности высшими образовательными учреждениями». Вторую компоненту U_2 можно определить как «уровень относительной обеспеченности высшими образовательными учреждениями», поскольку она имеет наибольшую положительную нагрузку на показатели численности студентов на 10 000 жителей и численность профессорско-преподавательского персонала на 1000 студентов.

Полученные результаты анализа можно представить графически (см. рис. 1). Здесь особенно наглядно видна близость анализируемых показателей и их принадлежность к выделенным компонентам.

На основе матрицы корреляций между исходными показателями и матрицы компонент были определены численные значения выделенных компонент для каждой из исследуемых областей (см. табл. 4).

Обычными методами данные компоненты не поддаются численному измерению. С помощью же полученных данных можно легко определить, какое место занимают исследуемые регионы Приволжского федерального округа РФ по уровню обеспеченности высшими учебными заведениями (см. табл. 5).

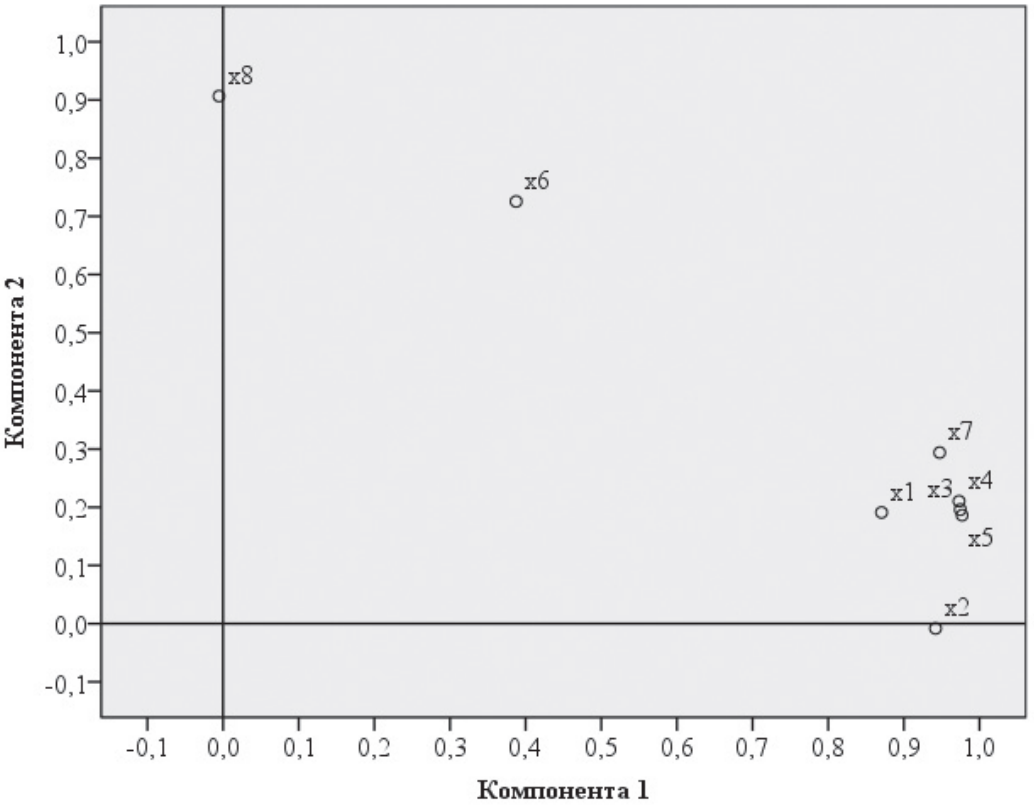

Рис. 1. Диаграмма компонент во вращаемом пространстве

Таблица 4

Расчетные значения компонент

Регион	Компоненты	
	уровень абсолютной обеспеченности высшими образовательными учреждениями	уровень относительной обеспеченности высшими образовательными учреждениями
	U_1	U_2
Республика Башкортостан	1,14899	–,85242
Республика Марий Эл	–1,19109	–,21324
Республика Мордовия	–1,36686	2,07537
Республика Татарстан	2,17651	,78333
Удмуртская Республика	–,27905	–,66528
Чувашская Республика	–,58321	–,07515

Регион	Компоненты	
	уровень абсолютной обеспеченности высшими образовательными учреждениями	уровень относительной обеспеченности высшими образовательными учреждениями
	U_1	U_2
Пермский край	,11526	–,37571
Кировская область	–,36264	–1,86899
Нижегородская область	,98691	,13012
Оренбургская область	–,11716	–,90939
Пензенская область	–,75639	–,10437
Самарская область	,99335	,71697
Саратовская область	–,02598	1,36226
Ульяновская область	–,73863	–,00351

Таблица 5

Ранжирование регионов по уровню обеспеченности высшими образовательными учреждениями

Регион	Компоненты	
	уровень абсолютной обеспеченности высшими образовательными учреждениями	уровень относительной обеспеченности высшими образовательными учреждениями
	U_1	U_2
Республика Башкортостан	2	12
Республика Марий Эл	13	9
Республика Мордовия	14	1
Республика Татарстан	1	3
Удмуртская Республика	8	11
Чувашская Республика	10	7
Пермский край	5	10
Кировская область	9	14
Нижегородская область	4	5
Оренбургская область	7	13

Регион	Компоненты	
	уровень абсолютной обеспеченности высшими образовательными учреждениями	уровень относительной обеспеченности высшими образовательными учреждениями
	U_1	U_2
Пензенская область	12	8
Самарская область	3	4
Саратовская область	6	2
Ульяновская область	11	6

Видно, что по уровню абсолютной обеспеченности высшими образовательными учреждениями первое место занимает Республика Татарстан, а последнее место — Республика Мордовия. Анализ исходных данных показывает (см. табл. 1), что указанные регионы по числу образовательных организаций занимают соответственно 2 и 14 места, по числу филиалов образовательных организаций, по численности студентов, по приему в студенты и выпуску студентов — 1 и 13 места, по численности профессорско-преподавательского персонала — 1 и 12 места. По уровню относительной обеспеченности высшими образовательными учреждениями первое место занимает Республика Мордовия, последнее место — Кировская область. Соответственно по численности студентов на 10 000 жителей у них 2 и 3 места, а по численности профессорско-преподавательского персонала на 1000 студентов — 13 и 2 места.

Проведенный анализ позволяет сделать вывод, что применение факторного анализа предпочтительнее по сравнению с традиционными методами, в которых сравнение производится по отдельным показателям. В рассмотренном примере 8 исходных показателей были сведены к 2 компонентам, которые полно и объективно характеризуют основные стороны обеспеченности регионов высшими учебными заведениями.

Литература

1. *Орлова И.В., Концевая Н.В., Турундаевский В.Б., Уродовских В.Н., Филонова Е.С.* Многомерный статистический анализ в экономических задачах: компьютерное моделирование в SPSS: учебное пособие. М.: Вузовский учебник, 2009. 320 с.

2. Российский статистический ежегодник. 2015: Статистический сборник. М.: Росстат, 2015. 728 с.

3. *Яковлев В.Б.* Факторный анализ подготовки и трудоустройства молодых специалистов в сельском хозяйстве // Вестник Российского государственного аграрного заочного университета. 2007. № 3 (8). С. 140–150.

Literatura

1. *Orlova I.V., Koncevaya N.V., Turundaevskij V.B., Urodovskix V.N., Filonova E.S.* Mnogomerny'j statisticheskij analiz v e'konomicheskix zadachax: komp'yuternoe modelirovanie v SPSS: uchebnoe posobie. M.: Vuzovskij uchebnik, 2009. 320 s.
2. Rossijskij statisticheskij ezhegodnik. 2015: Statisticheskij sbornik. M.: Rosstat, 2015. 728 s.
3. *Yakovlev V.B.* Faktorny'j analiz podgotovki i trudoustrojstva molody'x specialistov v sel'skom hozyajstve // Vestnik Rossijskogo gosudarstvennogo agrarnogo zaochnogo universiteta. 2007. № 3 (8). S. 140–150.

V.B. Yakovlev

Reduction of Dimensionality of Data in Regional Statistics of Russian Education

The article discusses the methods of application of factor analysis to reduce the data dimensionality in the regional statistics of Russian education with the help of statistical package SPSS 14.0 for Windows on the example of provision of higher educational institutions of the Volga Federal District of the Russian Federation.

Keywords: statistics of education in Russia; methods of multidimensional analysis; factor analysis.