УДК 519.6:004.622

Исследование уровня социально-экономического развития регионов Российской Федерации методами многомерного анализа данных

А.Ю. Юдинцев, Г.Н. Трошкина

Алтайский государственный университет (Барнаул, Россия)

Multidimensional Data Analysis of Social and Economic Development in Russian Federation Regions

A.Yu. Yudintsev, G.N. Troshkina

Altai State University (Barnaul, Russia)

Статья посвящена исследованию уровня социально-экономического развития субъектов Российской Федерации за 2021 г. на основе данных Федеральной службы государственной статистики для мониторинга социально-экономического положения регионов. Использована методика многомерного анализа данных, представляющая собой: редукцию исходного множества неортогональных переменных при помощи факторного анализа к малоразмерному ортогональному факторному пространству; определение оптимального количества кластеров при помощи древовидной классификации; выполнение кластерного анализа в факторном пространстве. Вычислено положение регионов Российской Федерации в факторном пространстве. Определены расположение, состав, статистические характеристики кластеров. Рассчитаны объемы и плотности кластеров, выделены наиболее плотные кластеры с близкими и регионы с существенно отличающимися от среднего уровнями социально-экономического развития. Исследования выполнены машинными и графическими методами вычислительной математики, а результаты имеют теоретическое и прикладное значение.

Ключевые слова: многомерный анализ данных, факторный анализ, кластерный анализ.

In this paper, we examine the social and economic development level of the Russian Federation regions in 2021 using data from the Federal State Statistics Service. We employ multidimensional data analysis techniques to reduce the non-orthogonal variables via factor analysis to a small, orthogonal factor space, determine the optimal number of clusters via tree classification, and perform cluster analysis in the factor space. We calculate the position of the regions in the factor space, determine the location, composition, and statistical characteristics of clusters, and compute the volumes and densities of clusters. We identify the densest clusters, including those with close proximity and regions with significant variations from the average level of social and economic development. The study utilizes machine and graphical methods of computational mathematics and has both theoretical and practical implications.

Key words: multidimensional data analysis, factor analysis, cluster analysis.

DOI: 10.14258/izvasu(2023)1-24

Введение

Статья посвящена анализу и интерпретации статистических данных мониторинга социально-экономического положения регионов Российской Федерации. Как правило, исходные массивы данных имеют большую размерность, избыточны, неортогональны — состоят из переменных, сильно коррелирующих между собой, вследствие чего нельзя непосредственно применять классификацию методами кластерного анализа с евклидовой метрикой. Методы многомерного анализа данных — статистический факторный анализ [1–3], кластерный анализ [4–5] позволяют уменьшить размерность исходного пространства показателей и сформировать ортогональное факторное пространство с евклидовой метрикой, в котором можно выполнять дальнейшую кластеризацию. Методы мно-

гомерного анализа данных успешно применяются в решении задач социально-экономического анализа регионов [6–10].

В данной работе выполнено исследование уровня социально-экономического развития субъектов Российской Федерации за 2021 г. на основе данных Федеральной службы государственной статистики для мониторинга социально-экономического положения регионов [11]. Методами многомерного анализа данных выполнен переход к ортогональному двухмерному факторному пространству, исходные 82 регионов Российской Федерации разделены на 10 кластеров. Определены расположение, состав, статистические характеристики кластеров, выделены регионы с близкими и регионы с существенно отличающимися от среднего уровнями социально-экономического развития.

Постановка задачи и формирование массива данных для анализа

Для анализа социально-экономического положения субъектов РФ были выбраны следующие показатели за 2021 г. на основе данных статистического бюллетеня «Информация для ведения мониторинга социально-экономического положения субъектов Российской Федерации в январе — декабре 2021 г.» Федеральной службы государственной статистики:

V_1 — среднемесячная номинальная начисленная заработная плата работников (руб.) за период январь — декабрь 2021 г. V_2 — среднедушевые денежные доходы населения (руб.) в среднем за период январь — декабрь 2021 г.; V_3 — численность рабочей силы (тыс. чел.) в среднем за период январь — декабрь 2021 г.; V_4 — численность безработных (тыс. чел.) в среднем за период январь — декабрь 2021 г.; V_5 — объем инвестиций в основной капитал (млн руб.) за январь — декабрь 2021г.; V_6 — оборот розничной торговли (млн руб.) за 2021 г.; V_7 — отгружено товаров собственного производства, выполнено работ и услуг собственным силами за 2021 г. (без НДС, акцизов и аналогичных обязательных платежей), млн руб.

Каждый показатель представляет собой набор 82 значений (для каждого региона Российской Федерации), имеющих разные размерности, разные масштабы, и неортогональны, что затрудняет непосредственное использование методов многомерного анализа данных. Для дальнейшего анализа данных перейдем к стандартизованным переменным и уменьшим размерность исходного массива данных посредством факторного анализа. Использована формула стандартизации: $\mathrm{X}_{\mathrm{i}}^{\mathrm{j}} = \dfrac{(\mathrm{V}_{\mathrm{i}}^{\mathrm{j}} - \overline{\mathrm{V}}_{\mathrm{1}})}{\mathrm{S}_{\mathrm{i}}}$, где V_i^j — элементы исходных выборок V_i, i = 1,2, ... 7 — номер выборки, j = 1,2, ... 82 — номер региона для каждой выборки, $\overline{V}_i$, S_i — среднее и стандартное отклонение по выборке i.

В результате выполнения процедуры стандартизации осуществлен переход к новым, стандартизованным наборам данных: $X_1 - X_7$, которые содержат отклонения исходных значений от средних величин каждой выборки, деленных на стандартное отклонение. Таким образом, получаем безразмерные величины, которые имеют нулевое среднее значение и единичное стандартное отклонение.

Результаты факторного анализа

Множество векторов $X_1 - X_1$ неортогонально, большинство элементов корреляционной матрицы $R_{n,m} = corr(X_n, X_m)$, n = 1,2, ... 7, m = 1,2, ... 7, имеют значения значительно больше нуля, а часть векторов сильно коррелированы: $R_{1,2}$ = 0,94, $R_{3,4}$ = 0,78, $R_{3,5}$ = 0,85, $R_{3,6}$ = 0,98, $R_{3,7}$ = 0,88, $R_{4,6}$ = 0,72, $R_{5,6}$ = 0,85, $R_{5,7}$ = 0,99, $R_{6,7}$ = 0,91. Наличие сильно коррелированных векторов в исходных данных свидетельствует о возможности уменьшения размерности исходного множества, например, при помощи факторного анализа. Также проанализируем данные при помощи тестов Кайзера — Мейера — Олкина (КМО) и Бартлетта. В результате получаем значение КМО = 0,71 и, по тесту Бартлетта, χ^2 = 1011, p<0,0001. Таким образом, данные пригодны для выполнения факторного анализа.

В соответствии с правилом Кайзера при выполнении редукции исходного множества при помощи факторного анализа достаточно использовать факторы, соответствующие собственным значениям корреляционной матрицы больше единицы [1, 2]. В нашем случае получаем, что можно уменьшить размерность исходного пространства до двух факторов: F_1 и F_2 с величинами собственных значений 4,8 и 1,6 соответственно. В совокупности эти два фактора описывают 91,1 % общей дисперсии исходного множества.

Для проведения процедуры факторного анализа будем использовать ортогональное вращение quartimax с методом principal [1–3]. Ортогональное вращение приводит к построению ортогональных факторов, что в дальнейшем позволит в полученном ортогональном факторном пространстве использовать евклидово расстояние при проведении кластерного анализа и определении расстояний между регионами, центрами кластеров, определять относительное положение регионов внутри кластера. В результате сформировались два ортогональных фактора: F_1 и F_2. Причем фактор F_1 имеет сильную корреляцию (факторные нагрузки) с переменными $X_3 - X_7$ соответственно: 0,977; 0,828; 0,967; 0,914, а фактор F_1 коррелирует с X_1, X_2 соответственно: 0,949; 0,897. Таким образом, значение фактора F_1 пропорционально численности рабочей силы, численности безработных, объему инвестиций в основной капитал, обороту розничной торговли и объему отгруженных товаров и выполненных работ и услуг — соответствует уровню общего экономического развития региона и отра-

жает его вклад в экономику Российской Федерации. Фактор F_2 коррелирует с переменными X_1, X_2 и соответствует совокупному доходу населения региона. Для обоих факторов справедливо утверждение, что чем больше значение фактора, тем лучше социально-экономическое положение региона.

Результаты кластерного анализа

Для анализа положения регионов в факторном пространстве будем использовать кластерный анализ. На первом этапе выполним древовидную кластеризацию с евклидовым расстоянием. Анализ матрицы слияния показал, что при расстоянии слияния от 4,5 до 1,5 выделяются два устойчивых кластера: г. Москва и остальные регионы. Следующие сравнительно устойчивые конфигурации существуют при расстоянии слияния в диапазоне от 1,2 до 1,0 — пять кластеров и десять кластеров при расстоянии слияния от 0,7 до 0,5.

На рисунке 1 отражено расположение регионов Российской Федерации в факторном пространстве при разбиении на десять кластеров. Разбиение выполнялось методом k-средних для десяти кластеров. Кластеры обозначены C1, C2, … C10 соответственно. Легко заметить, что большая часть регионов Российской Федерации занимает положение в пределах одной дисперсии от центра координат — в квадранте $-1 < F_1 < 1$ и $-1 < F_2 < 1$. Далеко вправо на графике расположен регион г. Москва (6,809; 2,085), устойчивое положение в правом верхнем квадранте занимают Тюменская область (1,701; 1,621), г. Санкт-Петербург (1,460; 1,188) и высокое значение F_1 при положительном F_2 имеет Московская область (2,706; 0,016).

Высокие номинальные доходы F_1 при небольших значениях F_2 имеют регионы Дальнего Востока: Чукотский автономный округ (-1,229; 4,838), Магаданская область (-1,041; 3,425), Сахалинская область (-0,705; 2,577), Камчатский край (-0,909; 2,457), также значительные положительные значения общих доходов при низком значении F_2 имеют Республика Саха (Якутия) (-0,348; 1,622) и Мурманская область (-0,505; 1,530). Центральная область факторного пространства $-1 < F_1 < 1$ и $-1 < F_2 < 1$ приведена на рисунке 2.

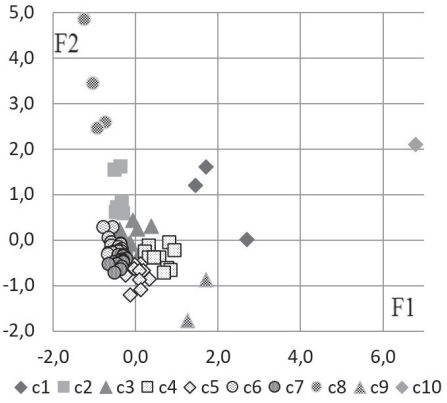

Рис. 1. Распределение регионов РФ в факторном пространстве (F_1, F_2)

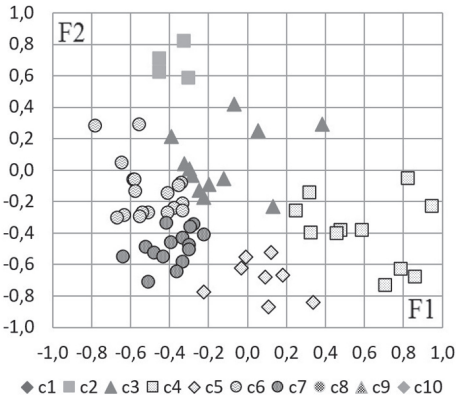

Рис. 2. Центральная часть распределения регионов РФ в факторном пространстве (F_1, F_2)

Распределение регионов по кластерам приведено в таблице 1, здесь С — номер кластера, F_1, F_2 — значения факторов.

Таблица 1

Распределение регионов РФ в факторном пространстве (F_1, F_2)

	С	F_1	F_2
УФО-3	1	1,701	1,621
СЗФО-10	1	1,460	1,188
ЦФО-10	1	2,706	0,016
ДФО-2	2	-0,348	1,622
СЗФО-7	2	-0,505	1,530
ДФО-6	2	-0,328	0,821
ДФО-7	2	-0,447	0,711
СЗФО-2	2	-0,451	0,622
СЗФО-3	2	-0,303	0,589
ДФО-5	3	-0,070	0,420

	С	F_1	F_2
СКФО-2	5	-0,118	-1,222
СЗФО-1	6	-0,552	0,286
ДФО-10	6	-0,776	0,278
ЮФО-8	6	-0,644	0,047
СФО-3	6	-0,586	-0,060
ЮФО-1	6	-0,581	-0,061
ЦФО-8	6	-0,339	-0,084
СЗФО-5	6	-0,350	-0,098
СЗФО-8	6	-0,571	-0,138
ЦФО-12	6	-0,411	-0,149

Окончание таблицы 1

	C	F_1	F_2
СФО-5	3	0,387	0,293
СЗФО-6	3	0,058	0,252
ЦФО-6	3	-0,388	0,214
СФО-10	3	-0,324	0,038
СЗФО-4	3	-0,299	0,010
ЦФО-9	3	-0,286	-0,031
ЦФО-1	3	-0,122	-0,057
ЦФО-16	3	-0,194	-0,093
ДФО-3	3	-0,250	-0,126
ЦФО-17	3	-0,228	-0,173
ЦФО-4	3	0,134	-0,232
ПФО-4	4	0,828	-0,058
СФО-6	4	0,319	-0,146
УФО-2	4	0,951	-0,228
ПФО-7	4	0,252	-0,263
СФО-8	4	0,479	-0,383
ПФО-9	4	0,588	-0,385
СФО-7	4	0,328	-0,396
ПФО-12	4	0,460	-0,401
ПФО-1	4	0,790	-0,632
ЮФО-7	4	0,859	-0,682
УФО-4	4	0,709	-0,735
СФО-9	5	0,122	-0,521
ПФО-10	5	-0,006	-0,550
ЮФО-3	5	-0,035	-0,625
ЮФО-6	5	0,182	-0,670
ПФО-13	5	0,089	-0,681
СКФО-3	5	-0,226	-0,774
СКФО-7	5	0,340	-0,841
СФО-4	5	0,107	-0,870
СКФО-6	5	0,120	-1,074
ЦФО-15	6	-0,332	-0,213
ЦФО-2	6	-0,372	-0,244
ДФО-1	6	-0,331	-0,257
ЦФО-11	6	-0,510	-0,272
ЦФО-13	6	-0,408	-0,273
СЗФО-9	6	-0,537	-0,277
СФО-2	6	-0,632	-0,288
ЦФО-7	6	-0,549	-0,298
СФО-1	6	-0,669	-0,305
ЦФО-14	7	-0,415	-0,336
ЦФО-3	7	-0,271	-0,350
ЮФО-5	7	-0,290	-0,359
ПФО-5	7	-0,218	-0,412
ПФО-14	7	-0,330	-0,434
ЦФО-5	7	-0,392	-0,462
ПФО-11	7	-0,297	-0,477
ПФО-2	7	-0,519	-0,489
ПФО-8	7	-0,297	-0,505
ПФО-3	7	-0,476	-0,532
УФО-1	7	-0,433	-0,549
ЮФО-2	7	-0,635	-0,553
ПФО-6	7	-0,334	-0,587
СКФО-5	7	-0,361	-0,647
СКФО-4	7	-0,505	-0,714
ДФО-11	8	-1,229	4,838
ДФО-8	8	-1,041	3,425
ДФО-9	8	-0,705	2,577
ДФО-4	8	-0,909	2,457
ЮФО-4	9	1,710	-0,881
СКФО-1	9	1,248	-1,787
ЦФО-18	10	6,809	2,085

Положение центров кластеров (CF_1, CF_2), размеры кластеров D_1, D_2, объем V и плотность P приведены в таблице 2. Здесь C — номер кластера, N — количество регионов в кластере. Размеры кластеров вычисляются как разница между максимальным и минимальным значениями соответствующих факторов для данного кластера: $D_1 = \max(F_1) - \min(F_1)$, $D_2 = \max(F_2) - \min(F_2)$. Объем кластера будем оценивать как произведение его размеров: $V = D_1 x D_2$. Для оценки степени близости регионов, входящих в кластер, введем понятие плотности кластера — количества регионов, входящих в кластер, приходящихся на единицу объема кластера: $P = N/V$.

Таблица 2

Сводная информация по кластерам

C	N	CF_1	CF_2	D_1	D_2	V	P
1	3	1,956	0,942	1,246	1,605	2,001	1,5
2	6	-0,397	0,982	0,202	1,034	0,209	28,7
3	12	-0,132	0,043	0,775	0,651	0,505	23,8
4	11	0,597	-0,392	0,699	0,677	0,473	23,3
5	10	0,058	-0,783	0,565	0,702	0,397	25,2
6	18	-0,508	-0,134	0,445	0,591	0,263	68,5
7	15	-0,385	-0,494	0,417	0,378	0,158	95,1
8	4	-0,971	3,324	0,524	1,097	0,575	7,0
9	2	1,479	-1,334	0,462	0,906	0,419	4,8
10	1	6,809	2,085	0,000	0,000	0,000	

Можно выделить кластеры, состоящие из небольшого количества регионов: С1, С2, С8, С9, С10. Все эти кластеры находятся на значительном расстоянии от центра и, соответственно, содержат регионы, существенно отличающиеся по уровню социально-экономического развития от кластеров: С3, С4, С5, С6, С7. Кластер С8: Чукотский автономный округ, Магаданская область, Сахалинская область и Камчатский край — имеет самые высокие значения фактора уровня доходов при самых низких значениях фактора F_2. Кластер С2: Республика Саха (Якутия), Хабаровский край, Амурская область, Мурманская область, Республика Коми, Архангельская область — также характеризуется весьма значительными величинами F_2 и несколько большими, чем С8, но весьма небольшими значениями фактора F_1. Кластер С9, состоящий из двух регионов: Краснодарский край и Республика Дагестан, напротив, имеет самые низкие показатели по уровню доходов населения при значительно выше средних значениях производственного фактора F_2. Лидерами как по уровню жизни, так и по уровню общего экономического развития являются кластеры С10: г. Москва и С1: Московская обл., Тюменская обл., г. Санкт-Петербург. Регионы, входящие в эти кластеры, отличаются высоким уровнем общего дохода населения и очень высоким уровнем экономического развития.

Рассмотрим центральную часть факторного пространства (см. рис. 2), содержащую остальные 66 регионов РФ, со сравнительно близким уровнем социально-экономического развития: кластеры С3, С4, С5, С6, С7. Наиболее плотными кластерами центральной части факторного пространства оказались кластеры со значительным количеством регионов и с небольшими объемами: С6 — 18 регионов и С7 — 15 регионов. Центроиды кластеров С6, С7 располагаются в левом нижнем квадранте факторного пространства, но часть регионов кластера С6 (Респ. Карелия, Еврейская авт. обл. и г. Севастополь) имеют значения общего уровня доходов населения выше среднего по РФ. При этом кластер С6 имеет по сравнению с С7 более низкое значение уровня экономического развития — F_2. Регионы кластера С5, примыкающего к группе С6, С7, имеют еще более низкие значения общего дохода при более высоком уровне экономического развития, соответствующего средним значениям по РФ. Лидерами по уровню социально-экономического развития среди регионов РФ, занимающих среднюю часть факторного пространства, являются регионы, входящие в кластеры С3, С4. Кластер С3 имеет более высокое положительное значение фактора F_2, чем кластер С4, причем регионы кластера С4 по уровню общих доходов существенно уступают и кластеру С6, при этом кластер С4 значительно превосходит С3 по уровню экономического развития.

Заключение

Представлены результаты исследования уровня социально-экономического развития субъектов Российской Федерации за 2021 г. методами многомерного анализа данных. Из исходных данных сформировано двухмерное факторное пространство, выполнен кластерный анализ, определены расположение, состав, статистические характеристики кластеров, выделены регионы с близкими и с существенно отличающимися от среднего уровнями социально-экономического развития.

Библиографический список

1. Факторный, дискриминантный и кластерный анализ / пер. с англ. Дж.-О. Ким, Ч.У. Мьюллер, У.Р. Клекка и др.; под ред. И.С. Енюкова. М., 1989.

2. Brown Timothy A. Confirmatory factor analysis for applied research. Guilford Press, 2006.

3. API documentation — factor_analyzer 0.4.0 documentation (factor-analyzer.readthedocs.io). https://factor-analyzer.readthedocs.io/en/latest/factor_analyzer.html.

4. Мандель И.Д. Кластерный анализ. М., 1988.

5. The complete guide to clustering analysis: k-means and hierarchical clustering by hand and in R. https://statsandr.com/.

6. Зубаревич Н.В. Социальная дифференциация регионов и городов России. http://gtmarket.ru/laboratory/expertize/5278.

7. Латышева М.А. Статистическое исследование дифференциации российских регионов по уровню социально-экономического развития // Вестник Волгоградского ун-та. Серия 3: Экономика. Экология. 2010. № 1.

8. Псарев В.И., Юдинцев А.Ю., Трошкина Г.Н. Исследование социально-экономических различий субъектов Сибирского федерального округа методом кластерного анализа // Известия Алт. гос. ун-та. 2015. Т. 1. № 2 (86).

9. Трошкина Г.Н., Юдинцев А.Ю., Межов С.И. Исследование динамики уровня экономической безопасности регионов Сибирского федерального округа Российской Федерации за период 2014–2017 год методами многомерного анализа данных // Российский экономический интернет-журнал. 2019. № 4.

10. Юдинцев А.Ю., Трошкина Г.Н. Формирование пространства показателей для анализа динамики уровня экономической безопасности регионов Российской Федерации за период 2014–2017 год // Российский экономический интернет-журнал. 2019. № 4.

11. Информация для ведения мониторинга социально-экономического положения субъектов Российской Федерации в январе — сентябре 2022 г. Федеральной службы государственной статистики. https://rosstat.gov.ru/storage/mediabank/info-stat-09-2022.rar.