

Применение нейросетей в решении задачи формирования кластеров регионов РФ

©2010 г. Н.С. Лукашевич, Ю.В. Гордополов*

Сущность управления социально-экономическим развитием региона состоит в целенаправленном воздействии органов управления на все региональные субъекты хозяйствования при соблюдении принципа баланса интересов, организованных преимущественно на договорной основе, при условии полной экономической ответственности партнеров, с целью обеспечения повышения качества жизни населения [1]. Основой такого управления являются оценка, мониторинг и сравнительный анализ уровня социально-экономического развития регионов.

Асимметричность развития регионов проявляется в их дифференциации по уровню социально-экономического развития. Многомерность процессов усложняет проблему обобщающей количественной оценки региональных различий. Кроме того, принятие управленческих решений в области социально-экономической политики, которые должны учитывать дифференциацию развития регионов, требует тщательной их проработки всеми уровнями власти, решения ряда правовых, организационных, информационных и аналитических задач. В этом смысле приоритетными являются такие связанные задачи, как оценка уровня социально-экономического развития регионов; выявление кластеров регионов, сходных по уровню развития; сравнительный анализ таких кластеров с целью определения тенденций межрегиональной дифференциации развития.

Поэтому перед исследователем возникает задача выявления характерных регионов с целью их сравнения по уровню социально-экономического развития и разработки специальной рефлексивной политики с учетом специфики кластеров развития регионов. Теоретическая значимость исследования повышается расширением теории принятия управленческих решений через апробацию нейросетевого подхода к кластеризации регионов по уровню социально-экономического развития.

* Н.С. Лукашевич – к. э. н., доцент кафедры «Предпринимательство и коммерция» Санкт-Петербургского государственного политехнического университета.

Ю.В. Гордополов – к. э. н., доцент кафедры «Предпринимательство и коммерция» Санкт-Петербургского государственного политехнического университета.

Сравнение реализуется на основе выявления обобщающего показателя (рейтинга). Для определения рейтингов используется достаточно большое количество методов, например экспертные методы, метод балльных оценок, метод многомерной средней, метод стандартизированных показателей, методы теории нечетких множеств. Результатом применения данных методов является нахождение качественного или количественного значения рейтинга, на основании которого осуществляются интерпретация уровня развития объектов и их сравнение. Недостатками применения традиционных методов являются: во-первых, сложность интерпретации рейтинга с точки зрения сравнения объектов оценки при их значительном количестве; во-вторых, сложность интерпретации рейтинга при большом количестве показателей, которые обуславливают высокий или низкий уровень.

При анализе и прогнозировании социально-экономических явлений исследователь довольно часто сталкивается с проблемой многомерности их описания [2]. Методы многомерного анализа – наиболее действенные количественные инструменты исследования социально-экономических процессов, описываемых большим числом показателей. Кластеризация как метод многомерного анализа успешно применяется в решении задач региональной экономики и оценки дифференциации регионов.

В контексте решаемой задачи результатом кластерного анализа станут кластеры схожих по уровню социально-экономического развития регионов. Каждый выделенный кластер получает рейтинг и качественную экономическую интерпретацию с точки зрения своего развития.

Сформулируем постановку задачи в математическом виде. Пусть X – множество объектов, Y – множество номеров (меток) кластеров, C – множество показателей. Имеется конечная обучающая выборка из m объектов $X_m = (x_1, \dots, x_m) \in X$. Каждому объекту x_m ставится в соответствие набор значений n показателей $C_m = (c_{m1}, \dots, c_{mn}) \in C$. Задана функция расстояния (метрика) между объектами $\beta(X_m; C_m)$. Требуется разбить выборку на k непересекающихся подмножеств $Y = (y_1, \dots, y_k)$, называемых кластерами, так, чтобы каждый кластер состоял из объектов, близких по метрике β , а объекты разных кластеров существенно отличались. При этом каждому объекту x_m ставится в соответствие номер кластера y_k согласно некоторому алгоритму (методу) кластеризации $f: X \rightarrow Y$. Множество Y заранее неизвестно.

На основе вышесказанного необходимо, во-первых, сформировать множество объектов исследования; во-вторых, сформировать набор значений предварительно отобранных показателей; в-третьих, обоснованно выбрать алгоритм кластеризации для решения задачи.

На первом этапе в качестве объектов исследования рассмотрены все субъекты Российской Федерации. На втором этапе отбора показателей возникают следующие проблемы: 1) большое количество показателей, характеризующих социально-экономическое развитие, что усложняет утилизацию результатов, следовательно, возникает проблема отбора основных показателей, которые в большей степени показывают дифференциацию развития; 2) необходимость учесть мультиколлинеарность показателей, которая может исказить результаты кластеризации; 3) отсутствие традиционного набора социально-экономических показателей, т.е. их отбор базируется на субъективном предпочтении исследователей, что во многом обуславливает субъективность результатов.

Поскольку регионы сильно дифференцированы по численности населения, площади территории, числу зарегистрированных организаций и другим показателям, выбирались преимущественно относительные и среднедушевые показатели, что позволило произвести более адекватное сравнение регионов. Основные показатели, в большей степени отражающие уровень социально-экономического развития регионов, публикуются в ежегодном статистическом сборнике [3]. Из этого справочника для примера были отобраны следующие показатели: C_1 – коэффициенты миграционного прироста (на 10 000 человек населения); C_2 – уровень экономической активности населения, %; C_3 – уровень безработицы, %; C_4 – среднедушевые денежные доходы населения, руб.; C_5 – численность населения с денежными доходами ниже величины прожиточного

минимума, %; C_6 – потребительские расходы в среднем на душу населения, руб.; C_7 – число собственных легковых автомобилей на 1 000 человек населения, шт.; C_8 – общая площадь жилых помещений, приходящаяся в среднем на одного жителя, m^2 ; C_9 – удельный вес расходов домашних хозяйств на оплату жилищно-коммунальных услуг, %; C_{10} – численность студентов образовательных учреждений высшего профессионального образования на 10 000 человек населения, чел.; C_{11} – число зарегистрированных преступлений на 100 000 человек населения; C_{12} – валовой региональный продукт на душу населения, руб.; C_{13} – степень износа основных фондов, %; C_{14} – оборот малых предприятий, млрд руб.; C_{15} – производство электроэнергии, млрд кВт·ч; C_{16} – ввод в действие жилых домов на 1 000 человек населения, квадратных метров общей площади; C_{17} – затраты на информационные и коммуникационные технологии, млн руб.; C_{18} – оборот розничной торговли на душу населения, руб.; C_{19} – объем платных услуг на душу населения, руб.; C_{20} – инновационная активность организаций, %; C_{21} – вклады (депозиты) физических лиц в рублях, привлеченные кредитными организациями, млн руб.; C_{22} – задолженность по кредитам в рублях, предоставленным кредитными организациями физическим лицам, млн руб.; C_{23} – удельный вес убыточных организаций, %; C_{24} – просроченная задолженность по заработной плате в расчете на одного работника, перед которым имеется просроченная задолженность, руб.; C_{25} – инвестиции в основной капитал на душу населения, руб.; C_{26} – стоимость фиксированного набора потребительских товаров и услуг, руб; C_{27} – экспорт, млн долл. США; C_{28} – импорт, млн долл. США; C_{29} – иностранные инвестиции, тыс. долл. США.

Фрагмент значений десяти показателей по Северо-Западному федеральному округу представлен в табл. 1.

Регион	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}
Республика Карелия	4,00	69,60	8,40	12 229,00	17,10	8 061,00	235,20	23,60	7,80	362,00
Республика Коми	–94,00	73,10	7,30	18 636,00	15,90	12 728,00	188,30	23,30	8,70	368,00
Архангельская область	–51,00	67,70	7,00	14 824,00	15,10	9 495,00	170,10	24,30	7,60	367,00
Вологодская область	5,00	69,00	5,90	12 194,00	16,00	7 131,00	221,10	25,00	8,20	422,00
Калининградская область	40,00	71,40	8,70	12 922,00	14,00	9 256,00	278,80	22,20	7,70	445,00
Ленинградская область	83,00	70,00	6,00	12 014,00	13,80	7 854,00	245,60	25,90	7,40	112,00
Мурманская область	–88,00	73,90	7,00	18 773,00	14,80	12 237,00	185,30	23,00	10,30	440,00
Новгородская область	3,00	67,10	4,90	11 646,00	18,00	8 341,00	196,70	26,30	9,30	344,00
Псковская область	–11,00	67,30	6,80	10 291,00	16,10	7 896,00	202,40	27,20	7,70	367,00
г. Санкт-Петербург	72,00	73,50	2,00	17 649,00	11,00	15 530,00	285,10	23,50	8,00	1 001,00

Использование большого числа показателей при кластеризации приводит к тому, что выделенные кластеры могут и не иметь четкой структуры [2], что порождает необходимость построения обобщенных критериев, каждый из которых содержит информацию сразу о нескольких показателях. Кроме того, кластерный анализ предъявляет следующие требования к данным [2, 4]: во-первых, показатели не должны коррелировать между собой; во-вторых, показатели должны быть безразмерными; в-третьих, их распределение должно быть близко к нормальному. Для решения этих задач целесообразно применять факторный анализ.

При анализе в один фактор объединяются сильно коррелирующие между собой переменные, как следствие – происходит перераспределение дисперсии между компонентами и получается максимально простая и наглядная структура факторов. После объединения корреляция компонент внутри каждого фактора между собой будет выше, чем их корреляция с компонентами из других факторов.

В качестве метода факторного анализа выбран традиционно используемый метод главных компонент, подробно описанный Г.О. Читай [5]. Метод варимакс с нормализацией Кайзера, выбранный в качестве метода вращения факторов, максимизирует разброс квадратов нагрузок для каждого фактора, что приводит к увеличению больших и уменьшению малых значений факторных нагрузок [4]. В качестве критерия определения числа факторов выбран критерий Кайзера, согласно которому отбираются только факторы с собственными значениями, равными или большими единицы. Расчеты проведены в ста-

тистическом пакете «SPSS 11.5». Высокое значение критерия адекватности выборки Кайзера-Мейера-Олкина и значение p -уровня критерия сферичности Барлетта свидетельствуют о высокой адекватности результатов факторного анализа. Рассчитанные факторные нагрузки, процент объясненной дисперсии и факторные коэффициенты представлены в **табл. 2**. При этом исключены показатели, которые не удалось однозначно отнести к тому или иному фактору, а также показатели, факторные нагрузки которых меньше заданного числа, т.е. являются незначимыми.

Полученные результаты позволили провести интерпретацию факторов F_1 – F_5 . Фактор F_5 включает один показатель, характеризующий объем производства электроэнергии. Фактор F_4 содержит показатели, отражающие инновационно-технологический уровень организаций. Фактор F_3 отражает уровень благосостояния (бедности) населения и организаций. Фактор F_2 содержит основные среднелюдские экономические показатели, отражающие уровень доходов и расходов (экономическую активность) населения. Фактор F_1 однозначно интерпретировать трудно, поскольку он содержит много разнородных показателей, но имеющих большие значения факторных нагрузок: показатели экспортно-импортных сделок; показатели деятельности кредитных организаций; показатель деятельности субъектов малого бизнеса; показатель уровня затрат субъектов предпринимательства. Предлагается интерпретировать фактор F_1 как критерий, характеризующий основные показатели деятельности субъектов предпринимательства и уровня деловой активности региона.

Таблица 2																										
Результаты факторного анализа																										
Показатели	Факторные нагрузки					Факторные коэффициенты																				
	F_1	F_2	F_3	F_4	F_5	F_1	F_2	F_3	F_4	F_5																
C_{21}	0,97	–	–	–	–	0,07	–	–	–	–																
C_{28}	0,97					0,14																				
C_{14}	0,95					0,15																				
C_{29}	0,95					0,16																				
C_{17}	0,94					0,17																				
C_{27}	0,91					0,17																				
C_{22}	0,75	–	–	–	–	0,18	–	–	–																	
C_{26}	–					0,87				–	–	–	0,16													
C_4						0,85							–	–	–	0,19										
C_{25}						0,84										–	–	–	0,22							
C_{12}						0,79													–	–	–	0,25				
C_{19}						0,69																–	–	–	0,28	
C_6		0,61	–	–	–	0,11																				
C_{16}	–	0,77				–	–	–	–	0,30																
C_3		–0,76								–	–	–	–	–0,34												
C_7		0,68												–	–	–	–	0,27								
C_{23}		–0,64																–	–	–	–				–0,23	
C_5		–0,61																				–	–	–	–	–0,23
C_{13}		–	0,80	–	–																					–
C_{20}	0,79		–			–	–	–	0,53																	
C_{15}	–								0,79	–	–	–	–													
Основные статистические показатели																										
Дисперсия, %	34,911	20,389							14,129					7,259	6,516	Критерий КМО		0,83								
Кумулята дисперсии, %	34,911	55,300							69,428					76,687	83,203	Критерий Барлетта		2202,97								

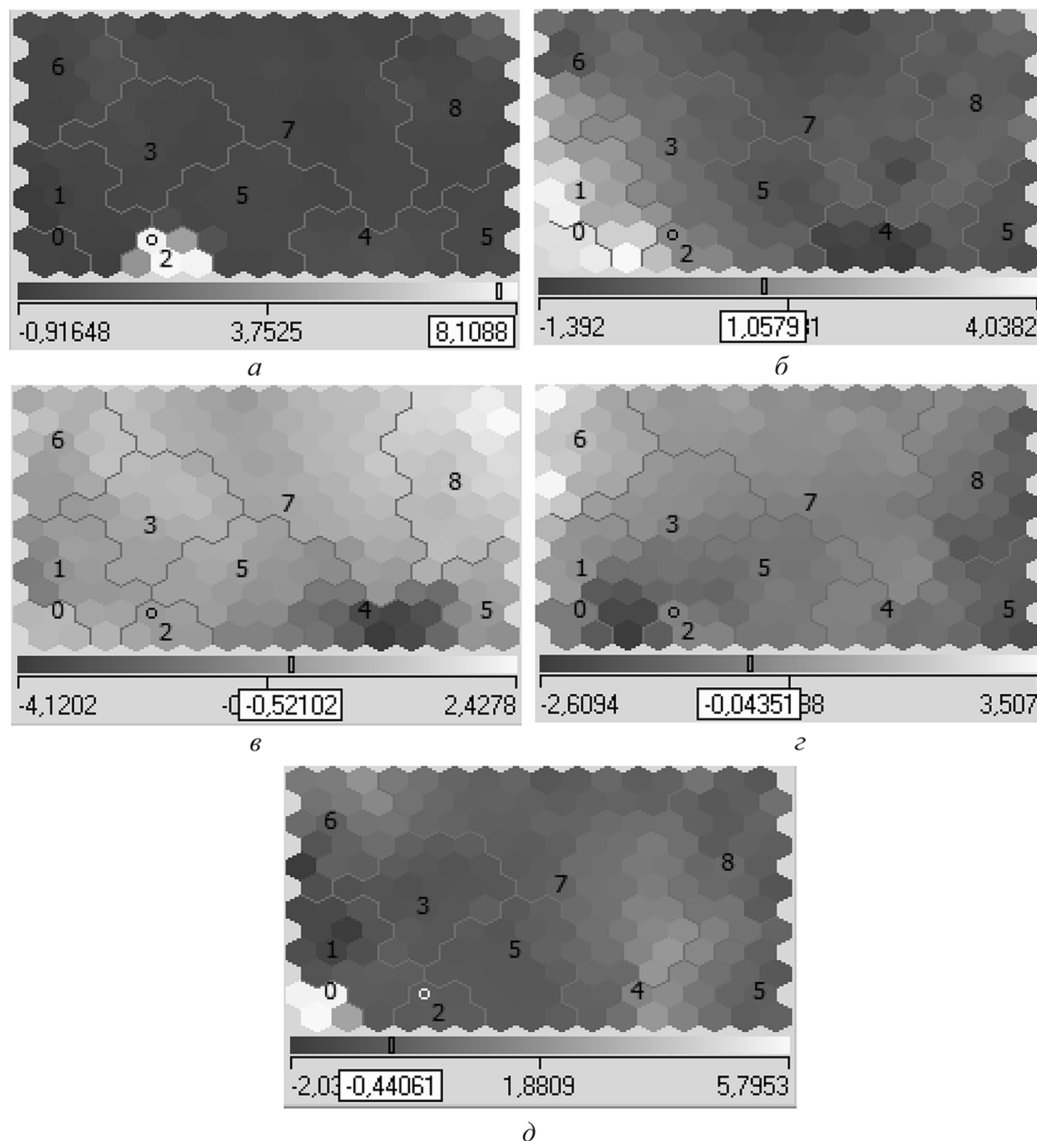


Рис. 1. Карты входных критериев: а – критерий F_1 ; б – критерий F_2 ; в – критерий F_3 ; г – критерий F_4 ; д – критерий F_5

Для каждого фактора F_1-F_5 рассчитаны значения в разрезе регионов, которые являются исходными данными для кластеризации.

На третьем этапе необходимо выбрать метод кластеризации. В литературе существует достаточное количество методов кластеризации, из которых был выбран метод, основанный на нейронных сетях, а именно на самоорганизующихся картах Кохонена.

Процесс соотнесения друг с другом объектов управления с увеличением оценочных характеристик наталкивается на проблему взаимопоглощения отклонений показателей. Поскольку более чем трехмерное пространство не подходит для визуализации группировки близких многомерных моделей, то для этого применяют нейросетевые модели данных. На это и направлено построение карт Кохонена с целью снизить размерность входных (исходных) параметров таким образом (обучение), чтобы возврат к исходной размерности искажал данные с минимальной ошибкой обобщения. Такая

соревновательная сеть спонтанно, без вмешательства исследователя обучается размещать нейроны относительно их соседей и с учетом топографии. На примере субъектов РФ были получены интересные результаты.

Для проведения кластеризации на основе самоорганизующихся карт Кохонена использован пакет анализа данных «Deductor Academic 5.2». Для инициации кластеризации заданы параметры нормализации значений обобщающих критериев, скорость и радиус обучения сети, функция соседства. В качестве способа начальной инициализации карты выбрана инициализация примерами из обучающегося множества, поскольку объем выборки небольшой и необходимо снизить вероятность появления после обучения пустых ячеек [6].

Рассмотрим основные результаты кластеризации регионов. На рис. 1 представлены карты входных критериев, позволяющие интерпретировать кластеры с точки зрения критериев кластеризации.

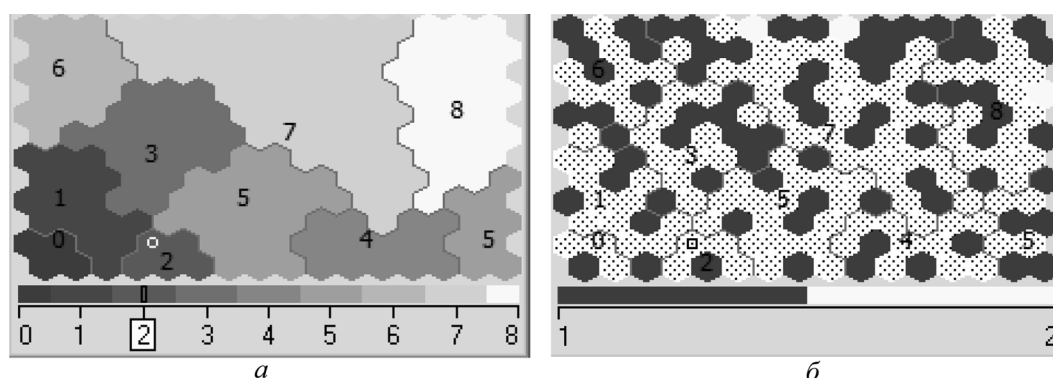


Рис. 2. Кластеры регионов (а) и матрица плотности попадания (б)

На рис. 2 представлены полученные кластеры регионов и матрица плотности попадания регионов в кластеры. В табл. 3 представлены профили полученных кластеров с точки зрения средних значений основных социально-экономических показателей.

На последнем этапе кластеризации дана интерпретация кластерам регионов с точки зрения уровня социально-экономического развития. В целом решение дало девять компактных кластеров.

Кластер 0 (Тюменская область) является аномальным по уровню экономического развития. Данный регион характеризуется очень высокими среднедушевыми показателями, низким уровнем безработицы, высокими темпами строительства жилых домов, высокими объемами производства электроэнергии, высоким благосостоянием жителей, но выделяется высокой степенью износа основных фондов и низкой инновационной активностью.

Профили полученных кластеров										Таблица 3	
Показатель	Кластеры										
	0	1	2	3	4	5	6	7	8		
Число объектов	1	5	1	10	4	12	8	26	13		
C_3	6,5	7,5	0,9	6,8	31,4	10,8	6,9	6,8	5,5		
C_4	27 612,0	22 653,8	34 207,0	12 608,1	6 345,0	10 169,7	15 548,0	10 643,9	13 557,4		
C_5	10,4	15,3	11,8	16,2	33,0	18,7	15,6	16,7	13,1		
C_6	18019,0	11842,4	26732,0	8698,5	2777,3	6536,9	10700,9	7659,2	10161,6		
C_7	252,7	179,4	278,7	201,8	108,1	161,9	213,7	187,2	234,3		
C_{12}	829 155,0	330 375,8	643 733,1	135 884,6	47 709,6	104 868,0	179 663,7	109 807,1	167 785,1		
C_{13}	53,2	41,9	36,1	46,8	44,5	41,9	55,4	49,3	42,5		
C_{14}	363,3	48,0	4522,8	111,5	4,8	38,5	318,0	147,0	433,8		
C_{15}	90,5	6,1	52,8	5,3	–	4,6	25,1	14,8	13,4		
C_{16}	656,0	164,2	311,0	296,1	129,5	265,0	304,6	385,3	659,1		
C_{17}	22 063,0	2 074,4	95 417,0	2 518,4	89,2	1 017,2	6 636,0	2 680,2	6 489,3		
C_{19}	36 170,0	41 248,8	84 581,0	25 810,0	8 816,3	16 129,8	28 000,6	18 651,8	25 551,2		
C_{20}	6,5	6,0	14,9	9,2	–	5,7	16,5	9,4	8,5		
C_{21}	151 229,8	18 059,2	1 228 068,0	28 177,5	1 540,2	11 004,9	57 423,6	34 990,6	82 064,9		
C_{22}	189 710,7	12 972,7	349 270,7	26 760,4	3 381,6	13 873,3	67 256,6	39 292,8	70 534,1		
C_{23}	26,9	36,3	29,6	33,2	40,1	33,4	29,0	27,6	26,2		

Данные значения показателей можно объяснить нефтегазовой специализацией региона.

Кластер 1 (Мурманская область, Республика Саха (Якутия), Камчатский край, Сахалинская область, Чукотский автономный округ) характеризуется средним уровнем социально-экономического развития. Кластер отличают высокие доходы и расходы населения, низкая деловая активность, в частности кредитных организаций, высокая доля убыточных организаций, но высокая инвестиционная привлекательность.

Кластер 2 (Москва) является вторым аномальным кластером. Во многом уровень развития соответствует Тюменской области, но для региона характерны более высокие среднелюдские показатели, высокая деловая активность кредитных организаций, самый низкий уровень безработицы, большие суммы экспортно-импортных сделок и притока инвестиций.

Кластер 7 самый многочисленный и состоит преимущественно из регионов Приволжского и Центрального федеральных округов. По сравнению с *Кластером 6* и *Кластером 8* регионы данного кластера незначительно уступают по показателям и характеризуются более низким (средним) уровнем социально-экономического развития.

Кластер 3 имеет смысл объединить с *Кластером 7*. Регионы *Кластера 3* уступают по сравнению с *Кластером 7* по показателям деятельности кредитных организаций, темпам строительства жилых помещений.

Кластер 4 (Республика Ингушетия, Республика Калмыкия, Чеченская Республика, Республика Тыва) характеризуется очень низким уровнем развития. Регионы данного кластера имеют наименьшие среднелюдские показатели, характеризуются высоким уровнем безработицы и бедности жителей, низкой инвестиционной привлекательностью и наибольшей долей убыточных организаций.

Кластер 5 (Владимирская область, Ивановская область, Республика Дагестан, Кабардино-Балкарская республика, Карачаево-Черкесская республика, Республика Северная Осетия – Алания, Республика Алтай, Республика Бурятия, Республика Хакасия, Забайкальский край, Амурская область, Еврейская автономная область), состоящий преимущественно из регионов Южного, Сибирского и Дальневосточного федеральных округов, характеризуется низким, но более высоким по сравнению с *Кластером 4* уровнем социально-экономического развития. Среднелюдские показатели и инвестиционная привлекательность выше, уровень безработицы ниже.

Кластер 6 и *Кластер 8*, преимущественно регионы Центрального и Северо-Западного федеральных округов, в том числе Санкт-Петербург и Ленинградская область, целесообразно укрупнить, поскольку различие по основным показателям незначительное. Принципиальное отличие кластеров состоит в степени износа основных фондов, притока инвестиций и инновационной активности организаций. Оба кластера характеризуются высоким уровнем социально-экономического развития.

Обобщая выводы, можно утверждать следующее. Регионы достаточно дифференцированы по уровню социально-экономического развития, об этом свиде-

тельствует большой разброс средних по кластерам значений показателей. Выделяются два аномальных региона с высоким уровнем развития. В государственной поддержке нуждаются прежде всего регионы Южного, Сибирского и Дальневосточного федеральных округов. Состав кластеров с точки зрения принадлежности регионов к федеральным округам различен. Предположение о том, что федеральный округ – это кластер регионов, схожих по уровню социально-экономического развития, не нашло подтверждения.

Все это доказывает возможность эффективно-го применения нейросетей к задаче кластеризации регионов по уровню социально-экономического развития с точки зрения наглядности и интерпретируемости результатов, несмотря на неоднозначное отношение к отдельным методам.

Применительно к самоорганизующимся картам Кохонена кластеризация подобна квантизации или разбивке на регионы Вороного. Она производится после построения карты, поскольку сама по себе полученная схема не дает анализа или интерпретации результатов. Внутри каждого кластера задача точности может быть решена отдельно через учет локальных и уникальных зависимостей и факторов. Карта лишь помогает выдвинуть гипотезу, требующую проверки. При большом количестве первичных данных часть из них может быть отброшена в начале обучения, а затем применима для подтверждения правильности гипотезы или оценки ошибки обобщения. Это особенно актуально в экономических исследованиях для сравнительного анализа, например товаров, предприятий, отраслей или регионов – с целью определения конкурентоспособности, цикла развития, инвестиционной привлекательности и пр., поскольку невозможно увидеть на карте наличие и степень влияния отдельных факторов. В связи с этим для удобства дальнейшей работы может использоваться один из вполне однозначных показателей (рентабельность, оборот).

Библиографический список

1. Донченко Ю.В. Сравнительная оценка социально-экономической эффективности развития регионов Центрального федерального округа Российской Федерации // Вопросы статистики. 2004. № 12. С. 48–52.
2. Ким Дж.-О., Мьюллер Ч.У. Факторный анализ: статистические методы и практические вопросы / Сб. работ «Факторный, дискриминантный и кластерный анализ». Пер. с англ. / Под. ред. И.С. Енюкова. – М.: Финансы и статистика, 1989. – 215 с.
3. Регионы России. Социально-экономические показатели: Стат. сб. / Росстат. – М., 2009. – 990 с.
4. Бююль А. SPSS: искусство обработки информации: Анализ статистических данных и восстановление скрытых закономерностей. Пер. с нем. – СПб.: ДиаСофтЮП, 2005. – 608 с.
5. Чистая Г.О. Факторный анализ промышленного развития макрорегионов России // Вопросы статистики, 2006. № 2. С. 19–28.
6. Саймон Хайкин. Нейронные сети: полный курс – Neural Networks: A Comprehensive Foundation. – М.: «Вильямс», 2006. – 1104 с.