М. А. Исакин

Модификация метода *k*-средних с неизвестным числом классов

В работе осуществляется построение и исследование модификации метода k-средних с неизвестным числом классов, общая схема которой была предложена профессором С. А. Айвазяном. В процедуре используется адаптивная метрика Махаланобиса и динамически рассчитываемые меры аномальности наблюдения и однородности классов. С помощью предложенного метода осуществляется классификация регионов РФ с целью определения кластеров, однородных по уровню благосостояния, качеству населения и социальной сферы.

При проведении различных социально-экономических исследований нередко возникает задача разбиения рассматриваемой совокупности регионов на однородные[1] классы. В частности, результаты классификации могут использоваться для нахождения групп регионов со схожим уровнем при определении приоритетов социально-экономического развития на основе анализа интегрального индикатора качества жизни населения.

Этот механизм определения приоритетов [Айвазян, Исакин (2006)] основывается на выявлении так называемых проблемных областей общественной жизни, а также наиболее значимо влияющих на качество жизни населения региона. Таким образом, каждый фактор (показатель) качества жизни населения должен быть рассмотрен и с точки зрения его проблемности, и значимости. При этом приоритетными в социально-экономической политике региона являются направления, характеризующиеся одновременно проблемностью и высокой значимостью показателей. Определение проблемных областей общественной жизни исследуемого региона основывается на анализе, с одной стороны, динамики социально-экономических показателей области, а с другой — положения региона относительно других субъектов федерации. В то же время следует отметить, что ранжирование регионов Российской Федерации по интегральным индикаторам качества жизни может не дать адекватной оценки в связи с существенными различиями в социально-экономическом положении регионов. Поэтому ранжирование следует производить, в первую очередь, в рамках группы регионов, схожих с анализируемым по социально-экономическим, географическим и другим базовым признакам.

В связи с тем, что обучающая выборка отсутствует, определить такие группы возможно с помощью процедуры кластерного анализа. Одним из его наиболее известных методов является процедура последовательной кластеризации — метод *k*-средних при неизвестном числе классов [MacQueen (1967)]. В этой процедуре выделяют несколько очевидных недостатков, таких как игнорирование различий в масштабе шкал и корреляционных зависимо-

[1] Вид типологической однородности зависит от решаемой задачи.

стей между анализируемыми признаками, что обусловливает использование только евклидовой метрики в соответствующем признаковом пространстве[2], а также чисто эвристический характер выбора значений свободных параметров процедуры — так называемых *мер точности и грубости*. И если проблема различия масштабов шкал признаков традиционно решается с помощью подходящего унифицирующего преобразования, приводящего все признаки к единой шкале измерений, то два других недостатка существенно снижают эффективность процедуры: если мы работаем с наблюдениями, извлеченными, например, из многомерных нормальных совокупностей, и различающимися только своими средними значениями, то обычная евклидова метрика должна быть заменена на так называемую *махаланобисскую* [Айвазян, Мхитарян (2001)]; в то же время априорное фиксирование мер точности и грубости, с одной стороны, связано с определенными трудностями выбора численных значений этих параметров, с другой — некорректно, так как эти величины могут меняться в процессе классификации.

В данной работе осуществляется построение и исследование модификации метода *k*-средних при неизвестном числе классов, общая схема которой предложена С. А. Айвазяном. Этот метод не обладает описанными выше недостатками, присущими «классическому» методу, потому что, во-первых, в модифицированном методе используется адаптивная метрика Махаланобиса, во-вторых, меры точности и грубости (аномальности наблюдения и однородности классов) рассчитываются в алгоритме динамически[3]. Модифицированный метод апробируется на выборках, сгенерированных методом Монте-Карло. С помощью предложенного подхода осуществляется классификация регионов РФ на основе трех систем показателей, отражающих интегральные категории качества жизни населения в модели интегральных индикаторов качества жизни: уровень благосостояния населения, качество населения и качество социальной сферы [Айвазян (2003)]. Результаты, представленные в данной статье использовались в научно-исследовательской работе «Внедрение системы целеполагания и оценки деятельности администрации и органов исполнительной власти Пермской области», которая проводилась по заказу аппарата администрации области в 2004–2005 годах.

Описание метода

В рассматриваемой задаче предполагается, что генеральная совокупность, из которой извлечены объекты, представляет собой смесь многомерных нормальных распределений с различными математическими ожиданиями и одинаковыми ковариационными матрицами. В процедуре кластеризации используется адаптивная метрика Махаланобиса.

Классы, которые образуются в результате выполнения процедуры, обозначаются: $\{S_i\}$, $i = 1, 2, \ldots, k$, где k — число классов.

Класс S_i характеризуется двумя величинами: центром (вектор размерности p) $e_i = (e_{i,1}, e_{i,2}, \ldots, e_{i,p})^T$ и весом w_i (скаляр). Центр класса характеризует «центр тяжести» наблюдений, отнесенных к соответствующему кластеру, а вес класса определяет количество таких

[2] Этот недостаток характерен для многих других методов кластерного анализа, использующих расстояние Евклида.

[3] Использование метрической функции с изменяющимися параметрами в некотором смысле опасно (как и вообще игра с изменяющимися в процессе правилами), однако в данном случае продуктивно.

наблюдений. Алгоритм кластеризации описывается следующей последовательностью действий.

1. Задается некое число классов k_0. Из выборочной совокупности извлекается k_0 точек[4], и эти точки объявляются центрами классов. Вес каждого класса устанавливается равным единице:

$$\begin{cases} e_i = X_i \\ w_i = 1, \ i = 1,2,\ldots,k. \end{cases}$$

2. Выполняется процедура проверки однородности классов. Для этого рассчитываются расстояния между каждой парой классов по формуле:

$$d^2(S_i, S_j) = \frac{n_i n_j}{n_i + n_j}(e_i - e_j)^T \Sigma^{-1}(i,j)(e_i - e_j), \quad (1)$$

где n_k, e_k и $\Sigma(i, j)$ — соответственно количество точек, попавших в класс S_k, центр тяжести этого класса и совместная ковариационная матрица классов S_i и S_j. Величины $\{n_k\}$ и матрицы $\{\Sigma(i, j)\}$ рассчитываются на основе *минимального дистанционного разбиения*.

После этого расстояние между двумя ближайшими классами (в смысле указанной метрики) сравнивается с мерой однородности классов:

$$c_0(S_i, S_j) = \frac{(n_i + n_j - 2)p}{n_i + n_j - p - 1} F_\alpha(p; n_i + n_j - p - 1), \quad (2)$$

где $F_\alpha(\nu_1, \nu_2)$ — 100α-процентная точка F-распределения с ν_1- и ν_2-степенями свободы числителя и знаменателя соответственно.

Если расстояние между ближайшими классами оказывается меньше $c_0(S_i, S_j)$, то соответствующие классы объединяются, т. е. эти два класса S_i и S_j удаляются, и формируется новый класс S_k с характеристиками:

$$\begin{cases} e_k = \dfrac{w_i e_i + w_j e_j}{w_i + w_j} \\ w_k = w_i + w_j. \end{cases}$$

Процедура объединения классов повторяется до тех пор, пока расстояние между любыми двумя классами не будет превышать меру однородности классов $c_0(S_i, S_j)$. И в результате мы имеем $k_0' < k_0$ классов.

3. Из выборочной совокупности извлекается очередная точка X и вычисляется расстояние от этой точки до каждого из классов по формуле:

$$d^2(X, S_j) = \frac{n_j}{n_j + 1}(X - e_j)^T \Sigma^{-1}(j)(X - e_j) \quad (3)$$

и определяется ближайший до рассматриваемой точки, в смысле указанной метрики, класс S_i. Потом расстояние до ближайшего класса сравнивается с мерой аномальности наблюдения:

[4] При этом могут быть извлечены первые k_0 точек выборки, либо из выборки случайным образом могут быть извлечены k_0 точек.

$$c_1(S_i) = \frac{(n_i - 1)p}{n_i - p} F_\alpha(p; n_i - p). \quad (4)$$

Если расстояние до ближайшего класса превышает величину $c_1(S_i)$, то рассматриваемая точка объявляется центром нового класса с весом, равным единице. Если же минимальное расстояние не превышает меру аномальности, то рассматриваемая точка присоединяется к ближайшему классу S_i, при этом характеристики этого класса пересчитываются по следующим формулам:

$$\begin{cases} e_i = \dfrac{1}{w_i + 1}(w_i e_i + X) \\ w_i = w_i + 1. \end{cases}$$

Параметры других классов при этом не изменяются.

Далее алгоритм повторяется с шага 2, т. е. производится процедура объединения близких классов с использованием меры однородности классов. Объекты поочередно рассматриваются до тех пор пока не закончится выборка.

Для определения величин $\{n_k\}$ и матриц $\{\Sigma(i)\}$, используемых при расчете расстояний между парами классов, классом и точкой, а также при вычислении мер однородности классов и аномальности наблюдения, осуществляется процедура минимального дистанционного разбиения, которая заключается в разбиении всей выборочной совокупности наблюдений на классы с центрами на момент выполнения процедуры. При этом подсчитываются количества наблюдений n_i, попавших в класс S_i, определяются выборочные средние наблюдений, отнесенных к классу S_i:

$$\bar{X}(i) = \frac{1}{n_i} \sum_{X_k \in S_i} X_k$$

и ковариационная матрица наблюдений, попавшая в класс S_i:

$$\Sigma(i) = \frac{1}{n_i - 1} \sum_{X_k \in S_j} \left(X_k - \bar{X}(i)\right)\left(X_k - \bar{X}(i)\right)^T.$$

При вычислении меры однородности классов S_i и S_j используется совместная ковариационная матрица этих классов:

$$\Sigma = \frac{1}{n_i + n_j - 2}\left[\sum_{X_k \in S_i} \left(X_k - \bar{X}(i)\right)\left(X_k - \bar{X}(i)\right)^T + \sum_{X_k \in S_j} \left(X_k - \bar{X}(j)\right)\left(X_k - \bar{X}(j)\right)^T\right].$$

В процедуре минимального дистанционного разбиения используется метрика (3) с ковариационной матрицей, определенной в предыдущем минимальном дистанционном разбиении. При первом выполнении процедуры в качестве ковариационных матриц классов могут использоваться единичные.

При недостатке информации, для расчета выборочной ковариационной матрицы $\Sigma(i)$ (например, когда к классу S_i отнесено лишь одно наблюдение) при вычислении расстояний может использоваться априорно заданная ковариационная матрица. Аналогично, когда в классах недостаточно наблюдений для того, чтобы использовать формулы (2) и (4) для вычисления мер грубости и точности, используются априорно заданные характеристики.

Ввиду того что выборочная совокупность может иметь небольшой объем, возможна ситуация, при которой не достигается устойчивого разбиения на классы к моменту, когда закончится выборка. В этом случае после последнего наблюдения выборочной совокупности снова извлекается первое, второе наблюдения и т. д., таким образом, выборочная совокупность зацикливается. В качестве критерия устойчивости разбиения на классы (и, следовательно, остановки алгоритма) предлагается следующее естественное условие. Процесс классификации останавливают, когда расстояния последних перемещений центров каждого из классов не превышают априорно заданную величину, т. е. стабилизировались[5]. В этом случае считается, что удалось получить устойчивое разбиение выборочной совокупности на классы и задача кластеризации решена.

Полученное разбиение зависит от уровня значимости, используемого для расчета мер однородности классов и аномальности наблюдения, априорно заданного начального числа классов, а также от выбора первых k_0-точек выборки, которые объявляются начальными центрами классов. Уровень значимости, как правило, задается в промежутке от 1% до 10%. Это значение определяет меру однородности классов и меру аномальности наблюдений (как доверительный интервал *F*-распределения), при этом чем меньше уровень значимости, тем больше указанные меры, и, вероятно, меньшее число классов получится в результате классификации. Определенное влияние на результаты классификации и на скорость сходимости алгоритма оказывает начальный выбор ковариационной матрицы. Если отсутствует информация о дисперсии признаков внутри классов, имеет смысл определить диагональные элементы ковариационной матрицы в зависимости от общего разброса значений признаков, по которым осуществляется классификация априорно заданного начального числа классов k_0 и предполагаемой структуры размещения классов. В большинстве случаев отсутствует априорная информация, позволяющая определить коэффициенты ковариации между признаками внутри классов, в следствие чего первоначальная ковариационная матрица определяется как диагональная. В общем случае первоначальная ковариационная матрица должна определятся исследователем как априорный параметр метода.

Предложенный алгоритм классификации апробирован на выборках, сгенерированных методом Монте-Карло из различных смесей нормальных распределений с коррелированными признаками (корреляционные матрицы признаков всех классов равны по предположению предлагаемого метода) и расстоянием Махаланобиса между классами от 7 до 10 [Апраушева (1981)]. При классификации использовалось начальное число классов, близкое к истинному количеству компонентов смеси (в частности, для пяти классов начальное число классов выбиралось от 3 до 7). Таким образом, получена правильная классификация, результаты которой совпадают для уровней значимости 1, 5 и 10% и не зависят от порядка выбора наблюдений из выборки. Классификация с помощью «классического» метода *k*-средних приводит к неверным результатам (формируются «лишние» классы, либо сливаются априорно различные классы — в зависимости от выбора мер точности и грубости), в связи с тем, что в алгоритме используется «сферическое» евклидово расстояние. На рис. 1 отражены результаты классификации классическим методом *k*-средних выборки, сгенерированной методом

[5] Число учитываемых последних итераций задается до начала процедуры классификации. При этом предполагается, что на этих последних итерациях не происходило образования новых классов или объединения уже существующих.

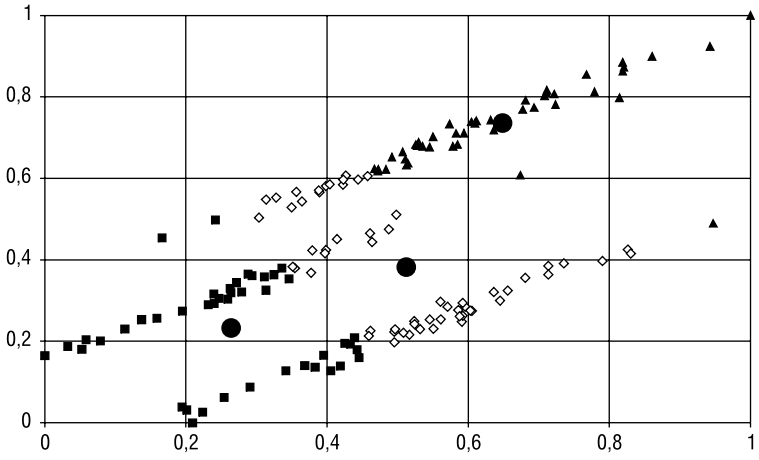

Рис. 1. Результаты классификации методом *k*-средних

Монте-Карло с тремя классами и высокой степенью корреляции между факторами: большие точки — центры полученных классов.

Классификация регионов

С помощью предложенного метода получены три классификационных разбиения регионов РФ по показателям интегральных категорий качества жизни населения: уровень благосостояния населения, качество населения и качество социальной сферы [Айвазян (2003)]. В табл. 1 отражены три соответствующих набора показателей, определяющих пространства классификации. Для проведения исследования использовались данные Госкомстата [Регионы России (2005)] за 2004 год.

Таблица 1

Центры классов полученных разбиений

Уровень благосостояния населения	Класс 1	Класс 2	Класс 3	Класс 4
ВРП на душу населения, *тыс. руб.*	315,67	48,65	51,125	97,9
Покупательная способность среднедушевых денежных доходов по отношению к наборам прожиточного минимума, %	312,5	159,21	176,45	247,22
Доля численности населения с денежными доходами ниже величины прожиточного минимума	20,533	34,363	30,507	21,346
Отношение совокупных доходов 20% самых богатых и 20% самых бедных	18,633	10,452	10,367	13,646
Обеспеченность собственными легковыми автомобилями на 1000 населения	21,033	18,333	21,435	20,537
Доля семей, состоящих на учете на получение жилья	9,4333	7,5479	8,5531	10,145
Приходится общей площади жилищного фонда на 10 жителей, кв. м	160,73	146,77	131,49	154,6
Доля ветхого и аварийного жилья	234,37	87,181	138,84	127,43
Плотность автомобильных дорог общего пользования на 1000 населения	3,2667	6,2042	2,7227	4,075

Окончание табл. 1

Качество населения	Класс 1	Класс 2	Класс 3	Класс 4
Ожидаемая продолжительность жизни при рождении: все население — оба пола	61,892	64,35	66,175	63,545
Число умерших детей в возрасте до 1 года на 1000 населения	15,831	16,975	12,231	12,061
Коэффициент естественного прироста	–7,1979	0,8125	–3,694	–8,4818
Число умерших на 100 000 населения:				
от инфекционных и паразитарных болезней и туберкулеза,	62,723	28,562	46,149	52,109
от новообразований,	198,02	162,02	181,89	203,06
болезней системы кровообращения,	1034,9	549,05	815,01	965,83
болезней органов дыхания,	83,379	52,625	61,885	98,376
болезней органов пищеварения,	61,292	41,075	51,282	69,727
от несчастных случаев, травм и отравлений	326,99	239,86	192,07	289,16
Число инвалидов на 1000 населения	78,183	53,263	66,66	80,745
Зарегистрировано случаев заболевания врожденными аномалиями на 1000 населения	1,8188	1,4125	1,7104	1,6364
Доля специалистов с высшим образованием среди занятых в экономике	20,254	26,787	23,21	20,124
Приведенная производительность труда (ВРП на среднегодовую численность занятых в экономике), *тыс. руб.*	110,64	550,01	137,96	130,98
Доля учащихся средних специальных и высших учебных заведений среди лиц, не достигших 23 лет	18,877	17,437	19,612	19,615
Качество социальной сферы				
Уровень безработицы	8,625	10,247	7,4595	13,073
Доля работников, занятых при вредных и опасных условиях труда в среднегодовой численности занятых в экономике	27,85	28,065	27,576	24,414
Численность пострадавших на производстве со смертельным исходом или с утратой трудоспособности на 1 рабочий день и более, в расчете на 1000 работающих	0,26071	0,3551	0,35952	0,25405
Коэффициент миграционного прироста на 10 000 населения	12,546	–22,386	–12,105	–10,319
Качество социальной сферы	**Класс 1**	**Класс 2**	**Класс 3**	**Класс 4**
Число зарегистрированных на 100 000 населения:				
умышленных убийств и покушений на убийство,	31,836	26,733	21,181	18,549
фактов умышленного причинения тяжкого вреда здоровью,	76,671	48,786	44,586	31,397
изнасилований и покушений на изнасилование	9,2536	8,7102	5,1643	4,8054
разбоев, грабежей, краж из квартир граждан,	406,4	361,19	358,45	269,18
незаконных присвоений или растрат	33,043	50,751	31,514	28,624
Больные, состоящие на учете с диагнозом:				
наркомания и токсикомания,	524,06	210,92	190,15	459,79
алкоголизм	2110,3	2070,7	3156,6	1768,2
Число самоубийств на 100 000 населения	44,914	60,294	40,2	29,738
Число больных, инфицированных ВИЧ	428,3	62,671	63,752	83,095

Классификационная процедура имеет ряд априорно задаваемых параметров, таких как начальное число классов, уровень значимости, начальная ковариационная матрица, начальные центры классов, порядок извлечения объектов выборки и др. В практике кластерного анализа

существуют инструменты, позволяющие обоснованно осуществить выбор каждого их перечисленных параметров. При реализации алгоритма были использованы методы, автоматизирующие выбор начальной ковариационной матрицы и начальных центров классов. В то же время начальное число классов и уровень значимости определяется экспертным путем. Порядок извлечения объектов выборки не оказывает существенного влияния на результаты. Классификация регионов осуществлялась на уровне значимости 5%, при этом начальное число классов определялось в интервале от 2 до 15. При классификации регионов с различными начальными параметрами получаются, как правило, 2–3 различных разбиения, устойчивых к небольшим изменениям начального количества классов. Наиболее устойчивыми разбиениями, при проведении классификации с различным начальным числом классов, по каждой из категорий, являются классификации с четырьмя классами.

Результаты классификаций по показателям категорий «Уровень благосостояния населения», «Качество населения» и «Качество социальной сферы» представлены в табл. 2 и на рис. 1.

Таблица 2

Результаты кластеризации субъектов РФ

Регион РФ	Уровень благосостояния населения	Качество населения	Качество социальной сферы
Алтайский край	2	3	4
Амурская обл.	2	1	2
Архангельская обл.	3	1	2
Астраханская обл.	2	3	4
Белгородская обл.	3	3	3
Брянская обл.	3	1	3
Владимирская обл.	3	1	3
Волгоградская обл.	2	3	2
Вологодская обл.	3	4	2
Воронежская обл.	3	4	3
г. Москва	1	2	3
г. Санкт-Петербург	4	3	1
Еврейская автономная область	3	1	2
Ивановская обл.	3	1	3
Иркутская обл.	3	4	1
Кабардино-Балкарская Республика	2	3	4
Калининградская обл.	2	4	1
Калужская обл.	3	3	3
Камчатская обл.	2	3	2
Карачаево-Черкесская Республика	2	3	4
Кемеровская обл.	3	1	4
Кировская обл.	3	1	2
Костромская обл.	3	1	3
Краснодарский край	2	3	4

Продолжение табл. 2

Регион РФ	Уровень благосостояния населения	Качество населения	Качество социальной сферы
Красноярский край	4	4	4
Курганская обл.	2	4	2
Курская обл.	3	3	2
Ленинградская обл.	3	4	1
Липецкая обл.	4	3	3
Магаданская обл.	4	3	3
Московская обл.	3	3	1
Мурманская обл.	4	3	4
Нижегородская обл.	3	1	3
Новгородская обл.	3	1	3
Новосибирская обл.	3	3	4
Омская обл.	3	3	2
Оренбургская обл.	2	1	1
Орловская обл.	3	1	3
Пензенская обл.	3	1	3
Пермская обл.	4	4	3
Приморский край	2	3	4
Псковская обл.	3	1	3
Республика Адыгея	3	3	4
Республика Алтай	2	1	2
Республика Башкортостан	3	3	2
Республика Бурятия	2	1	2
Республика Дагестан	2	3	4
Республика Калмыкия	3	3	2
Республика Карелия	3	1	3
Республика Коми	4	4	2
Республика Марий Эл	3	4	2
Республика Мордовия	3	3	2
Республика Саха (Якутия)	4	3	3
Республика Северная Осетия-Алания	3	3	4
Республика Татарстан	4	3	2
Республика Тыва	2	1	1
Республика Хакасия	2	1	2
Ростовская обл.	2	3	4
Рязанская обл.	3	4	3
Самарская обл.	4	3	1
Саратовская обл.	3	1	4
Сахалинская обл.	4	3	3

Окончание табл. 2

Регион РФ	Уровень благосостояния населения	Качество населения	Качество социальной сферы
Свердловская обл.	3	4	1
Смоленская обл.	3	1	3
Ставропольский край	2	3	4
Тамбовская обл.	3	4	3
Тверская обл.	3	1	3
Томская обл.	3	3	4
Тульская обл.	3	4	3
Тюменская обл.	1	2	1
Удмуртская Республика	2	4	2
Ульяновская обл.	3	1	1
Хабаровский край	3	1	2
Челябинская обл.	2	3	1
Читинская обл.	2	1	2
Чувашская Республика	3	4	2
Чукотский АО	1	2	3
Ярославская обл.	4	4	3

В категории качества жизни «Уровень благосостояния населения» образовался самостоятельный класс: город Москва, Тюменская область, Чукотский АО. Выделение указанных субъектов среди прочих регионов России в первую очередь связано с высокими значениями по показателю ВРП на душу населения. Во второй класс сгруппировались регионы Транссибирской магистрали, Южного федерального округа (за исключением республик Калмыкия, Адыгея и Северная Осетия), а также регионы, лежащие на основных направлениях российско-казахских транспортных путей. В третьем классе сосредоточились регионы средней полосы и северо-европейской части России, при этом аналогичной структурой благосостояния населения обладают центры образования Сибири (Омская, Томская и Новосибирская области). Четвертый класс регионов образуют субъекты с относительно низкой плотностью заселения территории, где большая часть населения сосредоточена в крупных населенных пунктах.

Классификация в категории «Качество населения» обладает сильной территориальной разнородностью и не позволяет сделать однозначные выводы об образованных классах. Однако в некоторых классах наблюдаются сравнительно однородные группы регионов. Так, в первом классе присутствуют регионы, расположенные вдоль Транссибирской магистрали, а также почти половина регионов средней полосы европейской части России, Республика Карелия и Архангельская область. Второй класс образован тремя субъектами: городом Москвой, Тюменской областью и Чукотским ОА, как и первый класс категории «Уровень благосостояния». В третий класс вошли все регионы Южного федерального округа, группа субъектов, состоящая из Нижегородской, Челябинской областей, республик Татарстан, Башкортостан, и группа регионов из Омской, Томской, Новосибирской областей и Республ-

Цифрами на карте обозначены:

1. Белгородская область
2. Владимирская область
3. Воронежская область
4. Ивановская область
5. Калужская область
6. Костромская область
7. Курская область
8. Липецкая область
9. Московская область
10. Орловская область
11. Рязанская область
12. Тамбовская область
13. Тульская область
14. Ярославская область
15. Вологодская область
16. Новгородская область
17. Республика Адыгея
18. Республика Ингушетия
19. Кабардино-Балкарская Республика
20. Карачаево-Черкесская Республика
21. Республика Северная Осетия-Алания
22. Чеченская Республика
23. Краснодарский край
24. Ставропольский край
25. Республика Башкортостан
26. Республика Марий-Эл
27. Республика Мордовия
28. Республика Татарстан
29. Удмуртская Республика
30. Чувашская Республика
31. Кировская область
32. Нижегородская область
33. Пензенская область
34. Пермский край
35. Самарская область
36. Саратовская область
37. Ульяновская область

Рис. 2. Результаты классификации регионов РФ

лики Алтай. Четвертый класс сформирован следующими группами регионов. В западноевропейской части России — Ленинградская, Вологодская, Ярославская, Тульская, Тамбовская, Рязанская и Воронежская области. В Приуралье — Республика Коми, Удмуртская Республика, Пермская и Свердловская области. В данный (четвертый) класс в азиатской части вошли Красноярский край и Иркутская область.

Объекты классов категории «Качество социальной сферы» также имеют сильный пространственный разброс. Однако внутри классов выделяются гомогенные структуры, территориально совпадающие с административными границами федеральных округов и экономических районов, что может быть косвенным признаком локальной непрерывности социогенеза вокруг основных центров развития территорий. В первом классе данной категории не удалось выдвинуть гипотезу о возможных причинах однородной структуры качества социальной сферы населения. Тем не менее в данную классификацию практически полностью вошел Уральский федеральный округ (исключение составила Курганская область). Второй класс включает в себя центральные субъекты Приволжского федерального округа, восточную часть Северо-Западного федерального округа и основные горнодобывающие регионы Забайкалья, включая Амурскую область. В третий класс вошла локация регионов Центрального федерального округа. Структурой, аналогичной «Качеству социальной сферы» обладает «пояс» из Республики Саха (Якутия), Магаданской области и Чукотского АО. Четвертый класс составляют субъекты Южного ФО и западные регионы Сибирского ФО.

Выводы

Рассмотренная модификация метода *k*-средних с неизвестным числом классов, общая схема которой предложена С. А. Айвазяном, позволяет устранить ряд недостатков, присущих «классическому» методу, и связанных, в первую очередь, с игнорированием зависимостей между показателями, по которым осуществляется классификация. Модифицированный метод успешно классифицирует выборки, сгенерированные методом Монте-Карло, и демонстрирует устойчивость к различному выбору начальных центров классов, а также определенную чувствительность к выбору начальной ковариационной матрицы. С помощью предложенной процедуры осуществлена классификация регионов РФ с целью нахождения кластеров, однородных по уровню благосостояния, качеству населения и социальной сферы.

Список литературы

Айвазян С. А., Мхитарян В. С. Теория вероятностей и прикладная статистика. Основы эконометрики. М.: ЮНИТИ, 2001.

Айвазян С. А. К методологии измерения синтетических категорий качества жизни населения // Экономика и математические методы. Т. 39. 2003. № 2. С. 33–53.

Айвазян С. А., Исакин М. А. Интегральные индикаторы качества жизни населения региона как критерии эффективности социально-экономической политики, проводимой органами региональной власти // *Прикладная эконометрика.* 2006. № 1.

Апраушева Н.Н. Определение числа классов в задачах классификации // Известия АН СССР, серия «Техническая кибернетика». 1981. № 3. С. 71–77. № 5. С. 153—160.

Регионы России. Социально-экономические показатели. Госкомстат, 2005.

MacQueen J. Some Methods for Classification and Analysis of Multivariate Observations // Proc. Fifth Berkeley Symp. Math. Stat. and Probab. 1967. Vol. 1. Pp. 281—297.