УДК 378:519.23:004.9

ПРИМЕНЕНИЕ МЕТОДОВ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ ДЛЯ КЛАССИФИКАЦИИ РЕЙТИНГОВ ТЕХНИЧЕСКИХ УНИВЕРСИТЕТОВ НА ОСНОВЕ ВСТУПИТЕЛЬНЫХ ИСПЫТАНИЙ

В.П. Арефьев, А.А. Михальчук

Томский политехнический университет
E-mail: vpa@ido.tpu.ru

Проведен многомерный статистический анализ качества набора абитуриентов в российские технические университеты на основе результатов вступительных испытаний 2010 г. Выявлена высоко значимая отрицательная корреляционная зависимость между долевым количеством абитуриентов, принятых по конкурсу баллов ЕГЭ и по целевому набору. Построена десятикластерная модель технических университетов в трехфакторном пространстве показателей вступительных испытаний 2010 г., позволяющая выделять группы технических университетов, однородных по совокупности факторных показателей.

Ключевые слова:
Многомерный статистический анализ, корреляционный анализ, факторный анализ, кластерный анализ, дисперсионный анализ, рейтинг, вступительные испытания.

Арефьев Владимир Петрович, канд. физ.-мат. наук, менеджер Института дистанционного образования ТПУ, доцент кафедры высшей математики и математической физики Физико-технического института ТПУ.
E-mail: vpa@ido.tpu.ru
Область научных интересов: информационные технологии в образовании и науке.

Михальчук Александр Александрович, канд. физ.-мат. наук, доцент кафедры высшей математики и математической физики Физико-технического института ТПУ.
E-mail: aamih@rambler.ru
Область научных интересов: компьютерное математическое моделирование различных процессов с использованием статистических методов.

На фоне всестороннего анализа современного состояния и проблем российского высшего образования [1] особенно активно обсуждается одна из новаций в реформировании образования в России – введение единого государственного экзамена (ЕГЭ) в систему аттестации школьных знаний и применения его как вступительного испытания для высшей школы. Результаты ЕГЭ рассматриваются как критерий оценки качества работы средней школы и качества набора абитуриентов в высшую школу [2–4] с привлечением иногда в качестве обоснования методов математической статистики [1, 5]. При реорганизации сети вузов РФ предполагается изменение механизма перераспределения средств в государственном секторе образования таким образом, что финансирование вуза будет зависеть от качества сформированного им контингента студентов, то есть от рейтинга вуза по среднему баллу ЕГЭ ($m_{\text{ЕГЭ}}$). Такой рейтинг [6] показывает, с какими знаниями абитуриенты 2010 г. поступили на бюджетные места в государственные вузы страны.

В частности, рейтинг качества приема в технические и технологические вузы, которые далее будут называться техническими университетами (ТУ), возглавляет Московский физико-технический институт (МФТИ), у которого средний балл ЕГЭ – 86,3 по 100 балльной шкале, а Московский ядерный институт (МИФИ) – на третьем месте (74,4). В рейтинге технических и технологических вузов шестым (72,3) стал Сибирский государственный университет путей сообщения (СГУПС), седьмым (71,7) – Санкт-Петербургский государственный политехнический университет (СПГПУ), девятым (68,3) – Новосибирский государственный технический университет (НГТУ), а у Томского политехнический университета (ТПУ) – 15-е место (66,7).

В работах [7–10] рассмотрено применение метода классификации вузов на мировом и региональном уровнях. В данной работе этот метод применен на федеральном уровне для классификации российских технических университетов на основе показателей вступительных испытаний 2010 г.

Визуально наблюдаемое распределение (гистограмма) $m_{ЕГЭ}$ (рис. 1) близко к теоретическому распределению по нормальному закону. Проверка нормальности распределения $m_{ЕГЭ}$ с помощью χ^2-критерия Пирсона дает незначимое (уровень значимости $p>0{,}10$) отличие от нормального закона с выборочной средней 59,153 балла по 100 балльной шкале и выборочным стандартным отклонением $\sigma = 6{,}285$. Диаграмма рассеяния с гистограммой $m_{ЕГЭ}$ по 100 балльной и стандартизированной шкалам приведена на рис. 1. Кроме вышеперечисленных ТУ на рис. 1 указаны в качестве примеров также Московский государственный технический университет им. Баумана (МГТУ), Томский государственный университет систем управления и радиоэлектроники (ТУСУР), Алтайский (АГТУ), Омский (ОГТУ) и Кузбасский (КГТУ) государственные технические университеты.

По результатам данного рейтинга только 7 ТУ смогли набрать себе отличников – т. е. тех, у кого средний результат ЕГЭ оказался выше 70 баллов, 97 ТУ смогли обеспечить себя хорошистами (55–70 баллов по ЕГЭ), а 35 ТУ были готовы принять всех, кто принес менее 55 баллов.

В данной работе на основании базы данных рейтинга качества приема в ТУ [6] проведена их кластеризация по совокупности показателей вступительных испытаний (ПВИ) 2010 г., включающих кроме $m_{ЕГЭ}$ также долевое количество абитуриентов (в % от общего количества бюджетных мест), принятых по конкурсу баллов ЕГЭ ($N_{ЕГЭ}$), по целевому набору ($N_{Ц}$), по олимпиадам ($N_{О}$) и по льготам ($N_{льг}$). Заметим, что подсистема долевых показателей является избыточной, так как $N_{ЕГЭ}+N_{Ц}+N_{О}+N_{льг} = 100$ %. В силу разнородности ПВИ они были стандартизированы.

Составляющими статистического метода исследования являются корреляционный, факторный, кластерный, дискриминантный и дисперсионный анализы. Статистический анализ проводился в системе Statistica [11].

Статистический анализ ТУ начнем с проверки ПВИ на корреляционную зависимость. Матрицы коэффициентов парных корреляций ПВИ приведены в табл. 1 (Пирсона r – в правоверхнем треугольнике над диагональю и Спирмена R – в лево-нижнем треугольнике под диагональю). Жирным шрифтом выделены высоко значимые (уровень значимости $p<0{,}0005$) корреляции.

Таблица 1. Матрица коэффициентов парных корреляций Пирсона r и ранговых корреляций Спирмена R ПВИ

ПВИ	$m_{ЕГЭ}$	$N_{ЕГЭ}$	$N_{Ц}$	$N_{О}$	$N_{льг}$
$m_{ЕГЭ}$	1,00	**–0,39**	0,23	**0,36**	0,08
$N_{ЕГЭ}$	**–0,44**	1,00	**–0,86**	**–0,49**	–0,18
$N_{Ц}$	0,29	**–0,86**	1,00	0,04	–0,00
$N_{О}$	**0,39**	**–0,31**	0,11	1,00	–0,12
$N_{льг}$	0,10	–0,29	0,07	–0,18	1,00

Диаграмма рассеяния и прямая регрессии с 95 % доверительным интервалом для наиболее сильной отрицательной корреляционной зависимости $N_{ЕГЭ}$ и $N_{Ц}$ изображена на рис. 2.

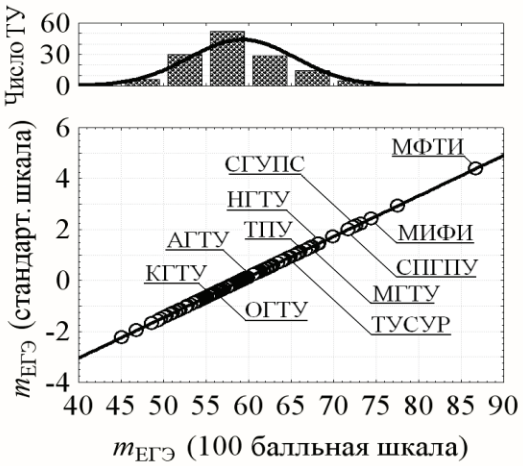

Рис. 1. Диаграмма рассеяния $m_{ЕГЭ}$ по стандартизированной и 100 балльной шкалам

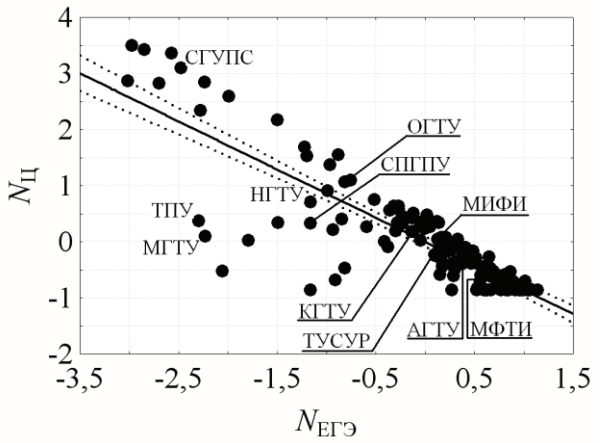

Рис. 2. Диаграмма рассеяния корреляционной зависимости $N_{ЕГЭ}$ и $N_Ц$

Наличие корреляционной связи ПВИ позволяет использовать факторный анализ для сокращения числа показателей и определения структуры взаимосвязей между показателями, т. е. классификации ПВИ.

Факторный анализ как метод классификации основан на оценках корреляций (факторных нагрузок) между исходными показателями и новыми показателями (факторами) в рамках выбранной факторной модели и позволяет узнать значимость факторов. Факторные нагрузки можно изобразить в виде диаграммы рассеяния, на которой каждый исходный показатель представлен точкой в факторном пространстве (координатах «факторные нагрузки»). С помощью типичного метода вращения – варимакс – получена простая интерпретация факторов, ясно отмеченная высокими нагрузками для некоторых показателей и низкими – для других (табл. 2), что и позволяет провести классификацию показателей.

С помощью факторного анализа построена четырехфакторная {Ф1, Ф2, Ф3, Ф4} модель ПВИ ТУ 2010 г. (табл. 2). В табл. 2 жирным шрифтом выделены наиболее значимые повернутые факторные нагрузки показателей на факторы, что позволяет по совокупности этих показателей интерпретировать соответствующие факторы, приписывая им наиболее существенные черты значимых показателей. В нижней строке приведены доли объясненной данным фактором дисперсии исходных показателей, иными словами, весовые коэффициенты факторов. Накопленная дисперсия представлена первыми тремя факторами ≈99 %.

Таблица 2. Матрица факторной структуры ПВИ ТУ 2010 г.

ПВИ	**Ф1**	**Ф2**	**Ф3**	Ф4
$m_{ЕГЭ}$	0,15	0,17	**0,97**	0,01
$N_{ЕГЭ}$	**–0,89**	–0,38	–0,19	–0,17
$N_Ц$	**0,99**	–0,07	0,10	–0,09
$N_О$	0,09	**0,98**	0,18	0,02
Доля фактора	**0,45**	**0,28**	**0,26**	0,01

Согласно табл. 2, высокие факторные нагрузки ПВИ распределились по факторам, имеющим наибольшие веса, следующим образом:

Фактор_1 (Ф1) – наиболее весомый (0,45), характеризуется $N_Ц$ и $N_{ЕГЭ}$, связанными отрицательной корреляционной связью (чем больше $N_Ц$, тем меньше $N_{ЕГЭ}$). Таким образом, положительная часть Ф1 интерпретируется как $N_Ц$ (чем правее по оси Ф1 (рис. 3), тем больше $N_Ц$), а отрицательная – как $N_{ЕГЭ}$ (чем левее по оси Ф1, тем больше $N_{ЕГЭ}$).

Фактор_2 (Ф2) – менее весомый (0,28), характеризуется $N_О$. Таким образом, фактор Ф2 интерпретируется как значение $N_О$, (чем выше по оси Ф2 (рис. 3), тем больше значение $N_О$, а чем ниже по оси Ф2, тем меньше значение $N_О$).

Фактор_3 (Ф3) – менее весомый (0,26), характеризуется $m_{ЕГЭ}$. Таким образом, фактор Ф3 интерпретируется как значение $m_{ЕГЭ}$, (чем выше по оси Ф3 (рис. 3), тем больше значение $m_{ЕГЭ}$, а чем ниже по оси Ф3, тем меньше значение $m_{ЕГЭ}$).

Заметим, что проверка нормальности распределения факторов с помощью χ^2-критерия Пирсона дает незначимое (уровень значимости $p>0{,}10$**)** для Ф3 и высоко значимые (уровень значимости $p<0{,}001$) для Ф1 и Ф2 отличия от нормального закона.

При проведении кластеризации ТУ в построенном трехфакторном пространстве {Ф1, Ф2, Ф3} в качестве меры близости выбрано евклидово расстояние, а в качестве правила объединения двух кластеров использован метод Уорда. Методом древовидной кластеризации построено иерархическое дерево (рис. 3).

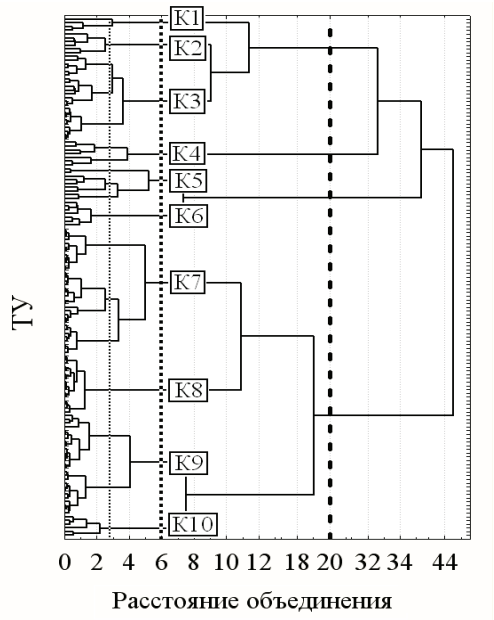

Рис. 3. Горизонтальная дендрограмма ТУ в факторном пространстве {Ф1, Ф2, Ф3}

Древовидная диаграмма начинается слева с каждого ТУ в своем собственном кластере. При движении вправо наиболее близкие в координатах Ф1, Ф2 и Ф3 ТУ объединяются и формируют кластеры. Каждый узел диаграммы представляет объединение двух или более кластеров, а положение узлов на горизонтальной оси определяет расстояние объединения соответствующих кластеров. В зависимости от выбора расстояния объединения можно получить соответствующее число кластеров. Так, например, расстоянию объединения, равному 6 (средняя пунктирная вертикальная прямая), соответствует 10 кластеров (К1–К10); равному 20 (правая крупная пунктирная вертикальная прямая) – 4 кластера (К1+К2+К3, К4, К5+К6, К7+К8+К9+К10); равному 2,8 (левая мелкая пунктирная вертикальная прямая) – 19 кластеров. Из кластера К1 вычленяется одиночный кластер МФТИ, в то время как ряд остальных кластеров дробится на более мелкие. Таким образом, выбор значения связующего расстояния позволяет проводить кластеризацию на любом уровне, т. е. строить кластерную модель с любым наперед заданным числом кластеров.

Предлагается десятикластерная модель ТУ, согласно λ-критерию Уилкса высоко значимо (на уровне значимости $p<0{,}0005$) различающая 10 кластеров ТУ по совокупности Ф1, Ф2 и Ф3. На основании *F*-критерия, а также рангового критерия Краскела–Уоллиса, можно оценить качество проведенной классификация для каждого фактора. В рассматриваемом случае *F*-критерий и ранговый критерий Краскела–Уоллиса показывают, что для каждого фактора различие между кластерами высоко значимо (на уровне $p<0{,}0005$).

После получения результатов классификации рассчитываются средние значения кластеров по каждому показателю (табл. 3).

Таблица 3. Матрица факторных средних кластеров десятикластерной модели ТУ, а также *N* – число ТУ в кластере

	Ф1	Ф2	Ф3	*N*
К1	–0,610	–0,695	**3,295**	4
К2	0,415	1,041	0,646	6
К3	–0,472	–0,240	0,893	23
К4	0,131	**3,891**	0,249	7
К5	**2,970**	–0,335	0,230	9
К6	1,418	–0,452	–0,153	7
К7	0,051	–0,341	–0,177	34
К8	–0,783	–0,302	0,004	16
К9	–0,653	–0,088	–0,888	27
К10	0,277	–0,034	**–1,731**	6

Согласно апостериорному критерию наименьших значений разности для Ф3 и ранговому критерию Краскела–Уоллиса для Ф1 и Ф2 можно выделить для каждого фактора однородные (различающиеся незначимо, то есть на уровне значимости $p>0{,}10$) группы кластеров, расположенные в порядке убывания факторных средних:

- Ф1: {К5}, {К6}, {К2, К10, К4}, {К10, К4, К7}, {К3, К1, К9}, {К1, К9, К8}. Имеются две пары пересекающихся групп так, что К2 отличается от К7 статистически значимо (на уровне $p\approx0{,}017$), а К3 отличается от К8 статистически значимо (на уровне $p\approx0{,}018$).
- Ф2: {К4}, {К2}, {К10, К9, К3, К5}, {К3, К8, К5, К7, К6}, {К5, К6, К1}. В данном случае образуются три последние последовательно пересекающиеся группы так, что К10 статистически значимо отличается от К8 и К7, а К1 слабо значимо отличается от К7 и К8. Кластер К5 настолько сильно распылен вдоль Ф2, что входит во все три группы.
- Ф3: {К1}, {К3, К2}, {К2, К4, К5}, {К4, К5, К8, К6, К7}, {К9}, {К10}. В полученном ряде образуются четыре последовательно пересекающиеся группы так, что К3 сильно значимо (на уровне $p\approx0{,}003$) отличается от К4, а К2 сильно значимо (на уровне $p\approx0{,}008$) отличается от К8 и высоко значимо (на уровне $p<0{,}0005$) отличается от К7.

Графики факторных средних кластера в рамках десятикластерной модели ТУ представлены факторной диаграммой рассеяния средних кластеров ТУ в трехфакторном пространстве {Ф1, Ф2, Ф3} (рис. 4) в виде образной формы «птицы потенциального высшего технического образования», обладающей «олимпиадным» крылом (13 ТУ кластеров К2 и К4) вдоль Ф2, «целевым» крылом (16 ТУ кластеров К5 и К6) вдоль Ф1 и возглавляемой МФТИ на вытянутой шее кластера К1 вдоль Ф3.

Проекции трехфакторной диаграммы рассеяния средних кластеров ТУ на соответствующие факторные координатные плоскости изображены в виде кластерных диаграмм рассеяния ТУ на рис. 5 (вид сверху) и рис. 6, 7 (вид сбоку). На примере отдельных кластеров прорисована их составная вложенная структура в соответствии с дендрограммой ТУ (рис. 3), разделяющей 139 ТУ на уровне расстояния объединения, равного 6, на 10 кластеров. Зримой становится процедура построения более подробной кластерной модели с большим числом кластеров. Так, например, в рамках 11-и кластерной модели от кластера К5 (рис. 6) отделяется на уровне расстояния объединения 5,22 в отдельный кластер пара ТУ, верхних по Ф3. Далее, в рамках 12-и кластерной модели кластер К7 (рис. 5, 6) на уровне 5,00 дробится по Ф1 на два кластера. Затем, на уровне 4,06 в рамках 13-и кластерной модели кластер К9 (рис. 6) дробится по Ф3 на два кластера. Следом, на уровне 3,88 в рамках 14-и кластерной модели кластер К4 (рис. 5, 7) дробится по Ф2 на два кластера. И так далее, и тому подобное. Заметим, что на уровне 3,30 из кластера К5 (рис. 6) выделяется в отдельный кластер пара ТУ, нижних по Ф3, а на уровне 2,99 из кластера К1 (рис. 6, 7) выделяется в отдельный кластер МФТИ.

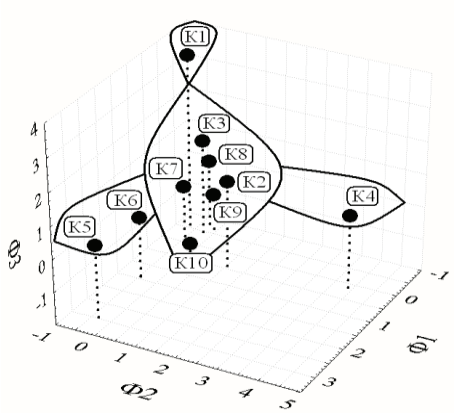

Рис. 4. Факторная диаграмма рассеяния средних кластеров ТУ

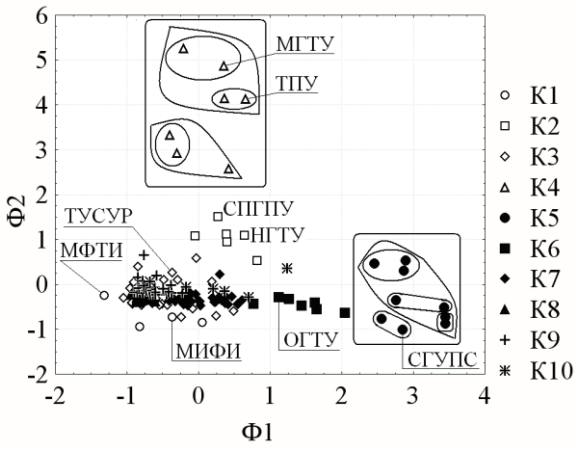

Рис. 5. Кластерная диаграмма рассеяния ТУ в факторных координатах Ф1 и Ф2

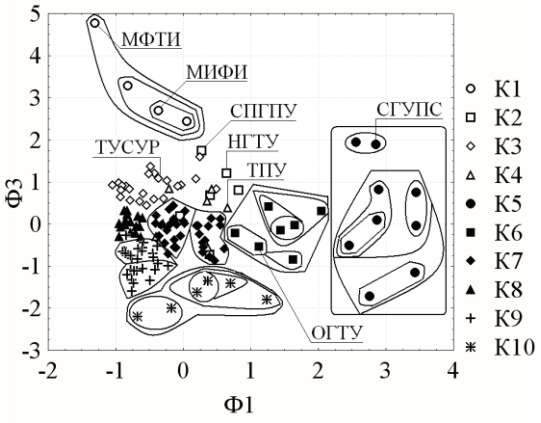

Рис. 6. Кластерная диаграмма рассеяния ТУ в факторных координатах Ф1 и Ф3

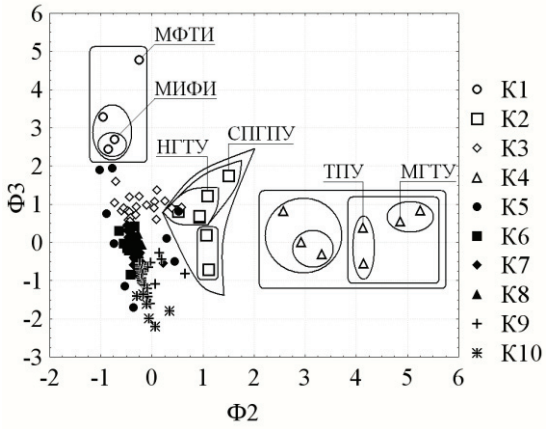

Рис. 7. Кластерная диаграмма рассеяния ТУ в факторных координатах Ф2 и Ф3

С помощью рис. 4–7 наглядным выглядят результаты выделения однородных групп кластеров для каждого фактора: Ф1 (рис. 5, 6), Ф2 (рис. 5, 7) и Ф3 (рис. 6, 7). Также наглядным и объяснимым становится вхождение в несколько групп одного кластера, например, К5 вследствие распыленности вдоль Ф2 и Ф3 (рис. 5–7).

При проведении классификации для каждого ТУ вычисляются апостериорные вероятности отнесения его к разным кластерам, что вызывает особый интерес в случае приграничных ТУ. Апостериорные вероятности ТУ определяются посредством расстояний Махаланобиса каждого ТУ от центров различных кластеров. Таким образом, каждый ТУ приписывают кластеру, к которому он ближе, т. е. когда расстояние Махаланобиса до него минимально, и для которого он имеет наивысшую апостериорную вероятность классификации. Так, например, ТУСУР с вероятностью 0,96 относится к К3, с вероятностью 0,025 – к К7 и с вероятностью 0,015 – к К8, ОГТУ с вероятностью 0,945 относится к К6 и с вероятностью 0,055 – к К7, а Волжская государственная академия водного транспорта, имеющая факторные координаты (0,36; -0,16; -1,36), с вероятностью 0,725 относится к К10, с вероятностью 0,235 – к К7 и с вероятностью 0,040 – к К9.

Результаты кластерного анализа ТУ по совокупности показателей (табл. 3) позволяют провести качественную классификацию ТУ в номинальной шкале измерений (табл. 4), полагая в качестве уровня «Средний» – стандартизированный интервал (-0,5; +0,5) для факторов. Аномально высокие значения (>+2,5) определяют уровень «**Лидер**», а аномально низкие значения (<-1,5) определяют уровень «А**утсайдер**». Промежуточные значения между средними и аномальными определяют уровень «Выше среднего» и «Ниже среднего» соответственно.

Таблица 4. Качественная классификация ТУ

Кластер	Уровень кластера на фоне среднего по фактору		
	Ф1($N_Ц$)	Ф2($N_О$)	Ф3($m_{ЕГЭ}$)
К1	Ниже среднего	Ниже среднего	**Лидер**
К2	Средний	Выше среднего	Выше среднего
К3	Средний	Средний	Выше среднего
К4	Средний	**Лидер**	Средний
К5	**Лидер**	Средний	Средний
К6	Выше среднего	Средний	Средний
К7	Средний	Средний	Средний
К8	Ниже среднего	Средний	Средний
К9	Ниже среднего	Средний	Ниже среднего
К10	Средний	Средний	**Аутсайдер**

В связи с приданием ЕГЭ обязательного статуса результаты качественной классификации ТУ в факторном пространстве (табл. 4) в силу сильной отрицательной корреляционной зависимости между $N_Ц$ и $N_{ЕГЭ}$ можно перефразировать на языке $N_{ЕГЭ}$ (среднее $N_{ЕГЭ} \approx 82$ %).

Заметим, что по показателю $N_{льг}$ лидерами являются Сибирский государственный аэрокосмический университет (36,3 %) и Восточно-Сибирский государственный технологический университет (18,5 %).

Работа выполнена при финансовой поддержке ФЦП «Научные и научно-педагогические кадры инновационной России», контракт № П691.

СПИСОК ЛИТЕРАТУРЫ

1. Сальников Н., Бурухин С. Реформирование высшей школы: актуальное состояние и проблемы // Высшее образование в России. – 2008. – № 8. – С. 3–13.
2. Грязев М.В., Хадарцев А.А., Хрупачёв А.Г., Туляков С.П. Методика интегральной оценки знаний абитуриентов // Высшее образование в России. – 2010. – № 6. – С. 28–32.
3. Гоник И.Л., Москвичев С.М., Иванов Ю.В., Гурулев Д.Н. Различные формы сдачи вступительных испытаний как элемент формирования контингента абитуриентов // Известия Волгоградского государственного технического университета. – 2009. – Т. 10. – № 6. – С. 27–28.
4. Данилов Д.А. ЕГЭ как критерий качества образования // Наука и образование. – 2008. – № 1. – С. 75–76.
5. Лапотникова И.Н. Методы математической статистики для оценки результатов ЕГЭ // Ярославский педагогический вестник. – 2008. – № 1. – С. 17–23.
6. Рейтинг вузов РФ по среднему баллу ЕГЭ 2010 года // РИА Новости. 2010. URL: http://www.hse.ru/org/hse/ex (дата обращения: 25.10.2010).
7. Кружалин В.И., Аршинова В.В., Попов Л.В., Чаплыгина А.А. Рейтинги мировых университетов как инструмент управления качеством образования // Alma mater (Вестник высшей школы). – 2010. – № 6. – С. 9–8.
8. Ильшев А.М., Шубат О.М. Многомерная классификация данных: особенности методики, анализ практики и перспектив применения // Вопросы статистики. – 2010. – № 10. – С. 34–40.
9. Сайфутдинова А.С. Проведение многомерной классификации вузов читинской области и АБАО-на основе кластерного анализа // Успехи современного естествознания. – 2008. – № 1. – С. 66–69.
10. Корсунов В.И. Классификация американских вузов и вопросы их диверсификации // Alma mater (Вестник высшей школы). – 2009. – № 2. – С. 52–60.
11. Халафян А.А. Statistica 6. Статистический анализ данных. – М.: ООО «Бином-Пресс», 2008. – 512 с.

Поступила 06.02.2012 г.