УДК 336.14:353

КЛАСТЕРНЫЙ ПРОФИЛЬ СИСТЕМЫ СТАТИСТИЧЕСКИХ ПОКАЗАТЕЛЕЙ, ОПРЕДЕЛЯЮЩИХ ЗАКОНОМЕРНОСТИ В ФОРМИРОВАНИИ НАЛОГОВОЙ СОСТАВЛЯЮЩЕЙ ДОХОДНОЙ БАЗЫ МЕСТНЫХ БЮДЖЕТОВ САМАРСКОЙ ОБЛАСТИ

© 2009 Н.А. Сбитнева*

Ключевые слова: методы анализа временных рядов, кластерный анализ, муниципальные образования, Самарская область, налог на доходы физических лиц, налоговый потенциал, земельный налог.

Исследована налоговая составляющая доходов местных бюджетов в Самарской области. В качестве методов достижения однородности данных в сформированном информационном массиве, получения логических групп показателей и построения на их основе причинно-следственных связей результативных и факторных показателей применялись процедуры кластерного анализа данных.

Согласно полученным кластерным структурам были выявлены комплексы показателей, формировавшие тенденции к скоплению переменных в группы и обуславливающие распределение муниципальных образований в кластеры различного уровня "реагирования" на происходящие структурные изменения.

Изучение общего перечня статистической информации о формировании налоговой составляющей доходов местных бюджетов за период с 2004 года по 2007 год позволило разделить полученный массив показателей по следующим блокам влияния:

- блока факторных показателей, включающего данные о демографической структуре населения в муниципальном образовании;
- совокупности факторов, характеризующих имеющиеся в наличие у муниципалитета ресурсы и эффективность их использования;
- совокупности факторов, отражающих структуру организаций в муниципальном образовании;
- блоков факторных показателей, определяющих уровень и качество жизни населения и результаты функционирования административного аппарата управления;
- совокупности факторов, отражающих структуру расходов и доходов местных бюджетов.

Однако для выбора модели описывающей исследуемую зависимость необходимо проведение анализа специфики взаимосвязи компонентов системы между собой.

Анализ дескриптивных статистик в крайних точках исследуемого периода - 2004 г. и 2007 г. - позволил сделать вывод, что ряды содержат как симметрично, так и асимметрично распределенные величины, вариация которых может характеризовать как однородность распределения величин в базовом периоде, так и ее отсутствие в отчетном периоде. Так, совокупность считается однородной, если коэффициент вариации не превышает 33% и только в данном случае целесообразно выдвижении гипотезы о нормальном распределении[1]. Проверка по критерию Колмогорова-Смирнова также подтвердила наличие нормального закона распределения только для величин с коэффициентом вариации ниже 33%. Таким образом, дальнейший анализ временных рядов предполагал приведения входящих в его состав величин к идентичному параметру распределения.

Эдельброк (1979) отметил, что переменные многомерных данных могут менять значения параметров распределения от группы к группе[2]. В данной связи нами было выдвинуто предположение, что приведение к нормальному закону распределения может и не быть равносильным преобразованием для этих переменных и, как обращает вни-

* Сбитнева Наталья Александровна, аспирант Самарского государственного экономического университета.

мание Эдельброк, может изменять соотношения между ними. Далее, как указывает Эверитт (1980), нормировка к единичной дисперсии и нулевому среднему уменьшает различия между группами по тем переменным, по которым наилучшим образом обнаруживались групповые различия [там же]. Таким образом, можно предположить, что решение в измерении меры сходства или различия анализируемых пространственно-временных структур содержится в применении методов кластерного анализа, посредством которых возможно выявить логическую взаимосвязь между переменными и дать ей математическую интерпретацию.

Учитывая основное требование в проведении кластерного анализа - измерение данных в одном регистре, и принимая во внимание факт подверженности абсолютного показателя эффекту масштаба, нами было принято решение сформировать кластерную структуру зависимых и независимых переменных в отдельности, исключив из исходного информационного массива показатели, выраженные в абсолютных величинах. Кроме этого, с целью выявления структурной составляющей, определяющей тенденции к скоплению переменных в группы, кластер-анализ проводился по полному массиву факторных показателей и усеченному - без ряда показателей блока "Доходы местных бюджетов", являющихся одновременно и результативными показателями (величина налогового поступления, приходящегося на 1 жителя/на 1 жителя, занятого в сельскохозяйственной деятельности и доля налогового сбора в общем объеме налоговых доходах местного бюджета).

Исходя из специфики решаемой прикладной задачи, кластеризация наблюдаемых пространственно-динамических единиц проводилась методом Уорда с использованием метрики Pearson-расстояния для определения кластеров переменных[3]. Для построения группировки по объектам (муниципальным образованиям) была выбрана метрика Euclidian distances, а для анализа полученного кластер-разбиение на наличие статистического выброса использовалась метрика City-block (Manhattan) distances[4]. Для оценки близости объектов была проведена итерационная процедура поиска кластер-образующих показателей методом К-средних Мак-Кина. Кластерные структуры были построены за 2004 и 2007 годы по 97 единицам и 113 единицам (за 2004 г. и 2007 г. соответственно) полного массива факторных показателей, по 82 и 98 единицам усеченного массива и по 15 единицам массива результативных показателей (количество единиц в исследуемых временных точках оставалось неизменным).

Полученная кластерная структура результативных показателей в 2004 г. и в 2007 г. представлена в виде дендрограмм на рис. 1. В обоих рассматриваемых периодах исходный информационный массив результативных показателей был распределен по трем кластерам. Процедура определения оптимального числа кластерных групп заключалась в графическом анализе взаимосвязи между величиной коэффициента слияния и числом кластеров.

На рис. 2 представлена кривая, характеризующая количество кластерных групп по результативным показателям в 2004 году. Заметное "уплощение" кривой свидетельствует о том, что дальнейшее слияние кластеров не дает новой информации, в то время как "скачок" (на данном графике соответствует величине коэффициента слияния - 1,8126 и 0,9663) означает объединение двух довольно несхожих кластеров. Таким образом, число кластеров предшествующее этому объединению, являлось наиболее вероятным решением.

При устойчивом распределении результативных показателей формирования налоговых доходов местных бюджетов Самарской области за все субпериоды 2004-2007 годов по трем кластерам, кластерная структура не была стационарна. Явной причиной подобных внутригрупповых изменений выступали вносимые изменения в Федеральное и региональное законодательства, среди которых наиболее существенным можно назвать изменение в принципе расчета земельного налога, вступившего в силу с 1 января 2006 года. В силу данного изменения, показатели, отражающие объем поступления земельного налога в местные бюджеты, по состоянию на конец 2006 года и в 2007 г. формировали отдельный кластер.

Кроме этого, анализ дендрограмм 2004 и 2007 гг. (рис. 1) свидетельствует в силу повышения однородности кластерной струк-

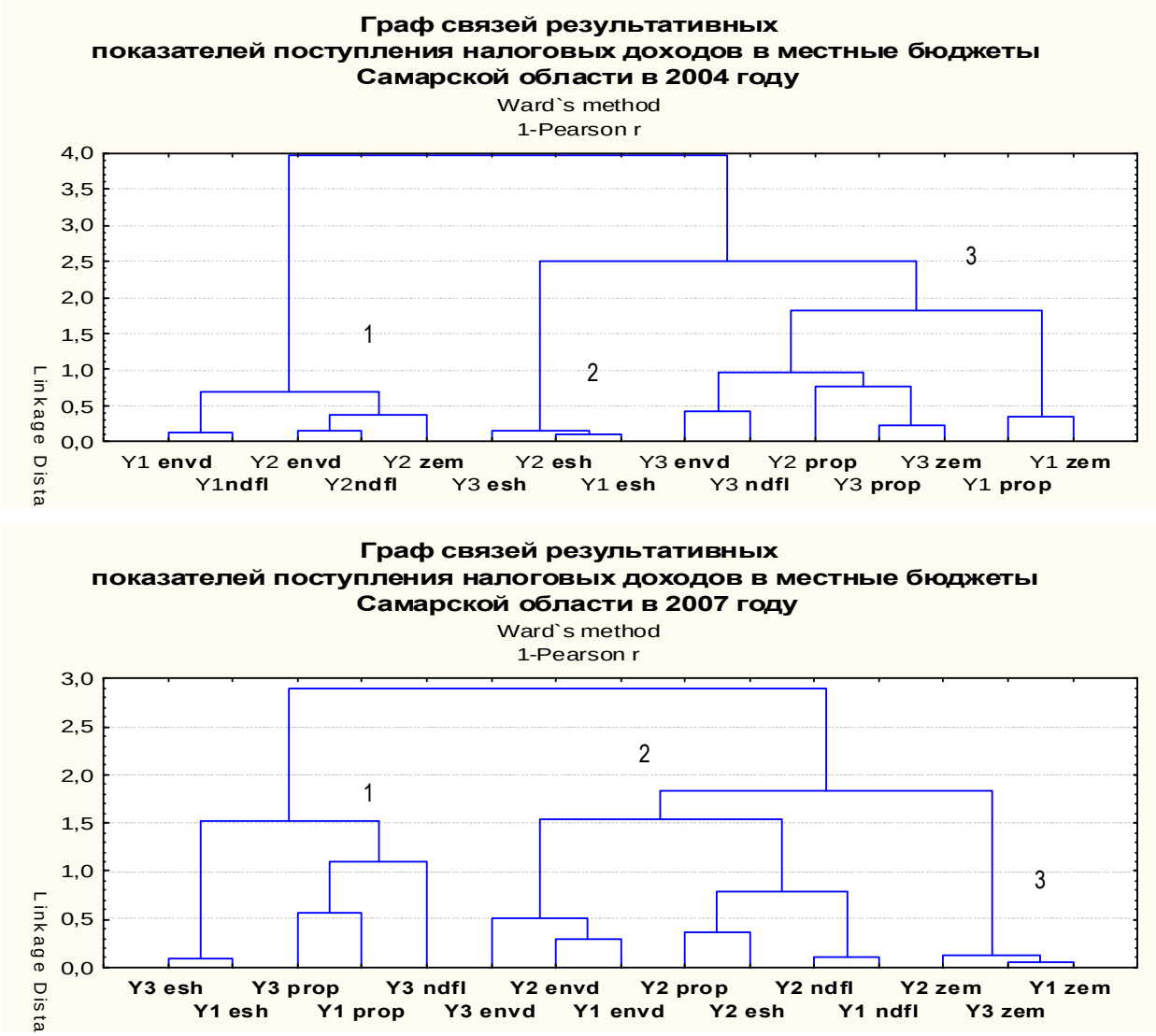

Рис. 1. **Распределение методом Уорда результативных показателей по кластерам в 2004 и 2007 гг.**

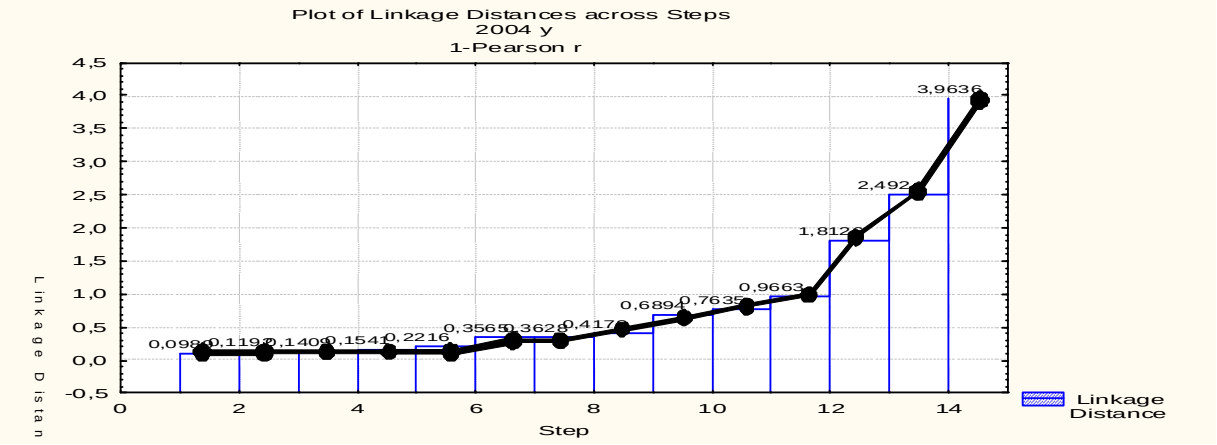

Рис. 2. **График зависимости между числом кластеров и величиной коэффициента слияния, полученный с помощью метода Уорда для результативных показателей в 2004 г.**

туры. Профили внутри каждого из скоплений имеют примерно одинаковую высоту, по мере этого, высоты больших профилей, объединяющих выделенные кластеры, в 2007 г. имеют уровень слияния ниже (максимальный уровень слияния ≈1,5), чем большие профили 2004 г. (максимальный уровень слияния ≈2,0).

Проведение итеративной процедуры кластеризации методом К-средних позволило определить показатели, формирующие кластерные группы, и кроме этого, осуществить распределение территориальных единиц по выявленным различиям. Согласно рис. 3 кластерообразующими показателями в 2004 г. являлись показатели величины земельного налога, приходящегося на 1 жителя занятого в сельскохозяйственной деятельности (Y2 zem), величины налога на доходы физических лиц, приходящегося в среднем на 1 жителя муниципального района (Y1 ndfl) и величины налога на доходы физических лиц, приходящегося на 1 жителя занятого в сельскохозяйственной деятельности (Y2 ndfl).

В 2007 г. формирование кластерных структур как и в 2004 г. было обусловлено влиянием трех показателей Y2 zem, Y2 ndfl и Y2 envd - величины единого налога на вмененный доход для отдельных видов деятельности, приходящегося на 1 жителя, занятого в сельскохозяйственной деятельности.

Таким образом, существенные различия между территориальными образованиями определялись действием указанных кластерообразующих показателей. Ранжирование средних значений данных показателей позволило распределить муниципалитеты по трем категориям в зависимости от их уровня бюджетной обеспеченности.

Средние значения кластерообразующих показателей в 2004 г. и в 2007 г. представлено в табл. 1, ротация кластерных групп по количеству входящих единиц представлено в табл. 2.

Стоит отметить, что за период 2004-2007 гг. имело место снижение доли муниципальных образований Самарской области с высокой и средней бюджетной обес-

Рис. 3. **Распределение средних значений результативных показателей (*Y*) за 2004 и 2007 гг. в кластерах, полученных методом *K*-средних при разбиении на 3 кластера**

Таблица 1

Распределение муниципальных образований Самарской области по кластерным группам в соответствии с ранжированием средних значений кластерообразующих показателей в 2004 г. и 2007 г.

2004			
Показатель	**Низкая бюджетная обеспеченность**	**Средняя бюджетная обеспеченность**	**Высокая бюджетная обеспеченность**
Y1 ndfl	643,12	1 511,29	2 892,26
Y2 zem	2 902,28	4 904,81	6 370,13
Y2 ndfl	8 763,83	28 708,20	64 473,98
2007			
Показатель	**Низкая бюджетная обеспеченность**	**Средняя бюджетная обеспеченность**	**Высокая бюджетная обеспеченность**
Y2 zem	926,81	13 019,9	10 744,8
Y2 envd	3 158,36	12 599,3	14 141,2
Y2 ndfl	25 103,08	128 816,2	466 411,4

Таблица 2

Динамика состава кластерных групп в абсолютном и процентном выражении за 2004-2007 гг.

Характеристика кластера	**Количество единиц по кластерам**				**Доля единиц, объединенных в кластер, в общем объеме совокупности**			
	2004	**2005**	**2006**	**2007**	**2004**	**2005**	**2006**	**2007**
Высокая бюджетная обеспеченность	7	4	1	1	25,9%	14,8%	3,7%	3,7%
Средняя бюджетная обеспеченность	17	6	5	5	63,0%	22,2%	18,5%	18,5%
Низкая бюджетная обеспеченность	3	17	21	21	11,1%	63,0%	77,8%	77,8%
Итого	27	27	27	27	100%	100%	100%	100%

печенностью (на 59,3 и 7,4 процентных пункта, соответственно), что, в свою очередь, определило семикратный рост (с 11,1% в 2004 г. до 77,8% к 2007 г.) доли муниципальных образований с низкой бюджетной обеспеченностью.

Аналогичная последовательность процедур кластер-анализа была осуществлена по факторным показателям полного массива (X). Усеченный перечень показателей (X_y) изучался на предмет различий с полным массивом показателей в кластерообразующих переменных.

Исследование зависимости между числом кластеров и коэффициентом слияния в обоих случаях подтвердила распределение единиц по трем кластерам

Проведение кластерного анализа по объектам (муниципальным районам Самарской области) также подтвердило значимость разбиения на три кластера как для полного, так и для усеченного массивов данных (рис. 4). Анализ на статистический выброс, согласно отмеченному ранее, был проведен посредством построения дендограммы с метрикой City-block (Manhattan), однако значительных различий в форме и структуре кластеров обнаружено не было. По мере этого статистический выброс был обозначен и при использовании Евклидовой метрики. Согласно рисункам, максимальная мера диссонанса объектов с общей совокупностью муниципалитетов была отмечена только при кластеризации полного и усеченного массивов данных за 2007 г. В обоих случаях в статистический выброс попал муниципальный район Кинельский.

Далее на основе полученного разбиения муниципалитетов на кластеры была осуществлена процедура определения значимости средних значений в кластерах. Каждое среднее значение проверялось на соответствие значимости по F-критерию. Если величина p не превышала выбранный уровень значимости в 0.05, то принималась гипотеза о том, что среднее значимо и, по всей видимости, определяет тенденцию к формированию кластера.

Следующим этапом в исследовании стала процедура - анализ методом K-средних, представляющая собой, своего рода, кластерный анализ результатов кластерного ана-

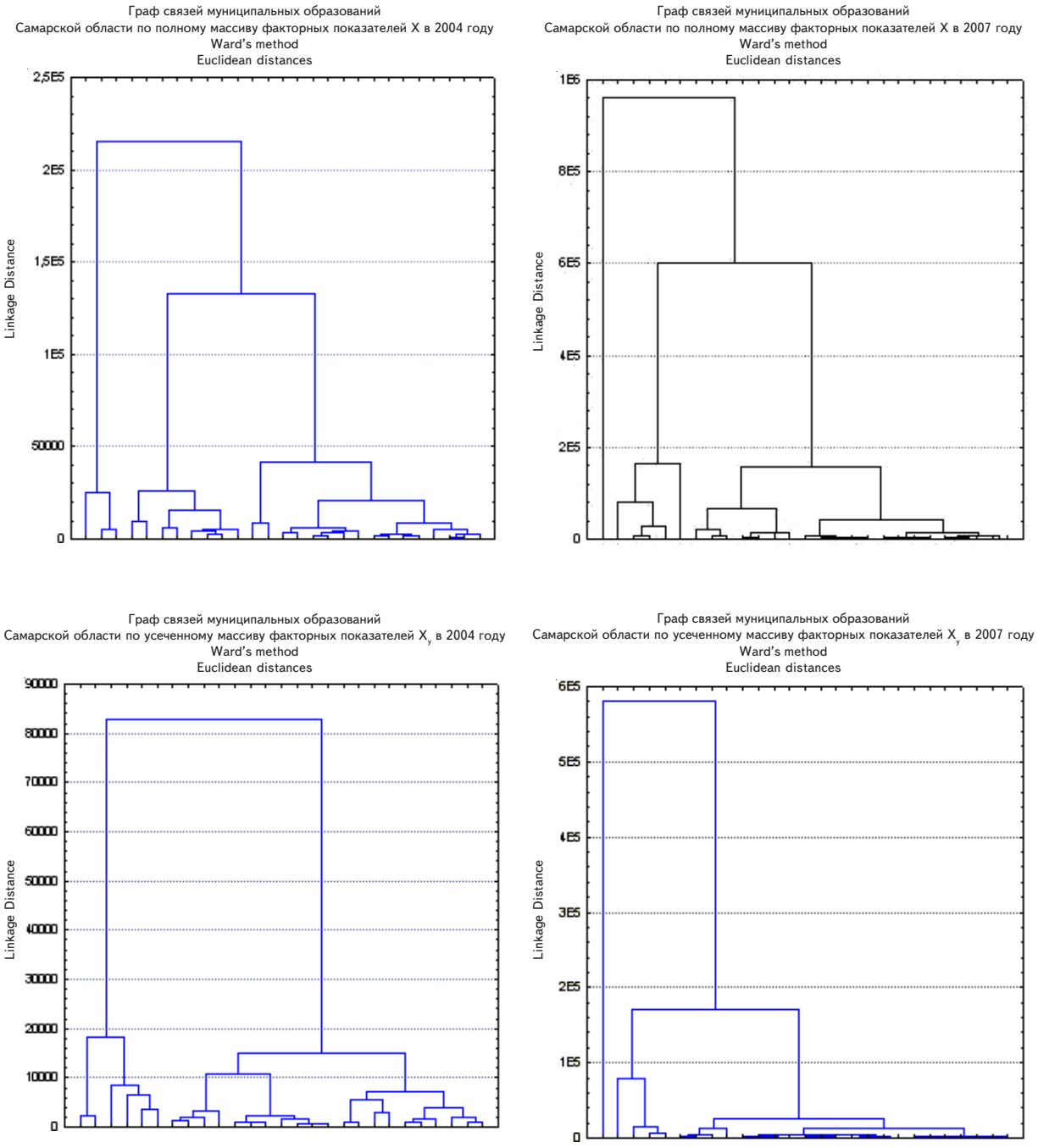

Рис. 4. **Распределение методом Уорда муниципальных образований Самарской области по полному и усеченному массивам данных в 2004 и 2007 гг.**

лиза. При этом анализ проводился только по переменным, отвечающим критерию "главного показателя в таксоне".

В результате данной операции было значительно сокращено первоначальное количество кластерообразующих переменных. Показатели, определившие скопление муниципальных образований Самарской области в классы, представлены в табл. 3.

Сравнение средних значений "лидеров" того или иного кластера с аналогичными показателями в других кластерах позволило дать объективную интерпретацию экономическому явлению, определившего тенденцию к формированию групп муниципальных образований.

Таким образом, в 2004 г. при анализе полного массива показателей *X* были вы-

Таблица 3

Кластерообразующие показатели полного и усеченного массивов данных в 2004 и 2007 гг.

Массив показателей	Кластерообразующие показатели		
	кластер 1	кластер 2	кластер 3
2004 массив X	er16	s19	i6
2004 массив X_y	s9	i24	r6
2007 массив X	-	i12	i6
2007 массив X_y	-	er10	i25

делены три группы муниципалитетов: с минимальным по области удельным весом прибыльных сельскохозяйственных организаций частной формы собственности в общей численности сельскохозяйственных организаций частной формы собственности (**er16**), муниципалитеты с максимальной по области величиной расходов социального направления, приходящейся в среднем на 1 жителя, занятого в сельском хозяйстве (**s19**) и муниципалитеты с минимальной по области величиной поступления по НДФЛ, приходящейся в среднем на 1 жителя, занятого в сельскохозяйственной деятельности (**i6**).

Картографическое распределение муниципальных образований в зависимости от принадлежности к тому или иному таксону подтвердило логику кластеризации. Муниципальные образования, вошедшие в первый кластер, территориально прилегают к крупным экономическим центрам (Самара, Тольятти, Новокуйышевск), в связи с чем, в данных муниципалитетах больше развито промышленное производство, в сравнении с юго-восточными районами, с ориентацией на сельскохозяйственное производство. Во второй кластер были сгруппированы "передовые" муниципальные районы Самарской области, с развитой собственной доходной базой, которые прилегают или в составе своей территории содержат крупные административные центры (Нефтегорск, Сызрань, Тольятти, Чапаевск и т.д.). Подобное территориальное расположение предполагает концентрацию рабочей силы, а, следовательно, значительные поступления по НДФЛ в местные бюджеты, в связи с уплатой данного вида налога по месту деятельности. Прямо противоположная ситуация наблюдается в муниципальных образованиях, образующих третий кластер, бюджеты которых в значительной части зависят от перечислений из вышестоящих бюджетов.

Согласно табл. 3 анализ методом К-средних усеченного массива данных в 2004 г. определил следующие виды кластерных структур. В первый кластер были сгруппированы районы с максимальной величиной расходов консолидированного бюджета территории на жилищно-коммунальное хозяйство, приходящихся в среднем на 1 жителя, занятого в сельскохозяйственной деятельности (**s9).** Территориально данные муниципалитеты расположены в центральной (Волжский, Красноярский) и северной части (Сергиевский, Челно-Вершинсий, Клявлинский) Самарской области, в связи с чем, высокие расходы жилищно-коммунального хозяйства в первом случае будут обоснованы завышенной, по сравнению с районами удаленными от административных центров, расценками на тарифы за электроэнергию и газоснабжение, а во втором случае - более длительной продолжительностью отопительного сезона. Во вторую группу попали муниципалитеты со средней по области величиной налоговых доходов полученных в результате подписания соглашения о передаче дополнительных нормативов отчислений от федеральных налогов (в данном случае речь идет о доле НДФЛ из бюджета субъекта), а также те муниципалитеты, кто в 2004 г. имел средний прирост налоговых поступлений (**i24**). Территориально данные муниципалитеты не имеют "точки" концентрации. Третий кластер был образован муниципальными районами со средним по области значением потребления платных услуг на душу населения (**r6**). На карте данные муниципальные образования либо прилегают и имеют хорошее транспортное сообщение с крупными административными центрами непосредственно Самарской обла-

сти, либо близко расположены к экономическим центрам соседнего субъекта (Кошкинский, Шенталинский, Камышлинский).

В 2007 г. анализ методом К-средних полного массива данных определил две категории муниципальных образований: муниципалитетов со средним по области объемом поступления по единому сельскохозяйственному налогу, приходящимся в среднем на 1 жителя, занятого в сельскохозяйственной деятельности (**i12**) и муниципалитетов с минимальным объемом по Самарской области поступлений по НДФЛ, приходящихся в среднем на 1 жителя, занятого в сельскохозяйственной деятельности (**i6**).

Картографическое распределение муниципальных образований, формирующих кластер первой категории, логически обосновано ориентацией районов юго-восточной и центральной территорий области на зерновое производство и растениеводство. Группировку муниципалитетов второго кластера определила аналогичная 2004 г. тенденция снижения поступлений по НДФЛ при увеличении расстояния межу муниципальным образованием и экономическим центром. Но при этом в 2007 г. число единиц рассматриваемого кластера значительно возросло на количество муниципальных образований, на порядок кластеризации которых повлиял статистический выброс.

Присутствие статистического диссонанса в 2007 г. было обусловлено высоким объемом поступлений в муниципальном районе Кинельский налоговых доходов в связи с заключенным соглашении о передаче дополнительных нормативов, приходящихся в среднем на 1 жителя, занятого в сельском хозяйстве (**i25**). Соотношение значительной величины аккумулированых налоговых доходов в абсолютном выражении (объем полученных доходов по соглашению в 2007 г. муниципальным районом Кинельский составил 66 025 тыс. руб., для сравнения, среднее значение по соответствующей категории районов было определено в размере 62 270 тыс. руб.) с низкой численностью населения, занятых в сельскохозяйственной деятельности в районе обусловило гипер-разрыв межу территориями в рассматриваемом показателе. В абсолютном выражении величина факторного показателя **i6** по муниципальному району Кинельский в 2007 г. составляла 314 407 руб. на 1 жителя, занятого в сельском хозяйстве. По всем остальным характеристикам муниципальный район Кинельский идентичен муниципальным районам третьего таксона.

При кластеризации усеченного массива данных в 2007 г. формирование групп муниципалитетов определили следующие факторы: доля убыточных сельскохозяйственных организаций в общей численности сельскохозяйственных организаций в районе (**er10**) и величина налоговых доходов полученных в результате подписания соглашения о передаче дополнительных нормативов отчислений от федеральных налогов (**i25**). В результате все муниципальные районы были ранжированы на кластер муниципалитетов со средним по области показателем (**er10**) и кластер муниципалитетов со средним по области показателем (**i25**).

Картографический анализ полученного кластер-распределния наглядно отразил представленную в целом по Самарской области ситуацию низкого уровня развития сельскохозяйственной отрасли. Другой кластер был сформирован из муниципалитетов, которые в 2007 г. имели наилучшие по Самарской области показатели налоговых поступлений по дополнительному нормативу. Однако в силу статистического выброса по соответствующим индикаторам Кинельского района, данным муниципалитетам было присвоено среднее место по области.

В целом анализ методом К-средних позволил перенести кластеры факторных признаков на непосредственно характеризующиеся данными признаками объекты - муниципальные образования. При этом стоит отметить, что как в 2004 г., так и в 2007 г. формирование групп муниципальных образований в полном и в усеченном массивах данных было обусловлено влиянием показателей одной смысловой нагрузки (к примеру, величина налоговых поступлений по НДФЛ и величина налоговых поступлений, полученных по дополнительным нормативам), не значительно изменяющих архитектуру кластеров.

Таким образом, в качестве объективных тенденций, определяющих распределение

муниципальных районов Самарской области по классам, можно выделить совокупность факторов, определяющих размер поступлений в местные бюджеты по НДФЛ, и систему индикаторов, отражающих уровень развития в районах сельскохозяйственного производства.

Учитывая то, что массив X представлял собой совмещение усеченного массива X_y с массивом результативных показателей Y различие в пространственно-динамических структурах X и X_y предполагает присутствие скрытой тенденции, выраженной линейной комбинацией корреляций результативных и факторных показателей.

В данной связи, для выявления существующих латентных факторов и их объективной интерпретации целесообразно применение статистического метода позволяющего осуществить логическое объединение показателей внутри каждого из кластеров по направлению и силе воздействия.

[1] *Ефимова М.Р., Петрова Е.В., Румянцев В.Н.* Общая теория статистики: Учебник. 2-е изд., испр. и доп. М., 2004. (Серия "Высшее образование").

[2] Факторный, дискриминационный и кластерный анализ: Пер. с англ. / Дж.-О. Ким, Ч.У. Мьюллер, У.Р. Клекка и др. / Под ред. И.С. Енюкова. М., 1989. С. 153.

[3] При том, что М. Кэндалл и А. Стьюарт (1973) утверждают об отсутствии как зависимости между формулой линейного коэффициента корреляции и видом распределения, так и возможности применения коэффициента корреляции Пирсона для большинства случаев закономерностей варьирования данных, использование данного коэффициента оправдано в случае распределения признаков близкому к нормальному.

[4] Евклидово расстояние зависит только от нескольких "доминирующих" разностей, поскольку они возводятся в квадрат, и практически игнорирует оставшиеся незначительные расхождения. Применение метрики Манхэттен, наоборот, в формировании расстояния преследует цель выделения "редкого" объекта, показывающего максимальный диссонанс с общей совокупностью объектов.