КЛАСТЕРНЫЙ АНАЛИЗ: СУЩНОСТЬ, ПРЕИМУЩЕСТВА И НЕДОСТАТКИ

С. А. Суслов, *к. э. н., доцент кафедры «Экономика и статистика» НГИЭИ*

Аннотация. Кластерный анализ является основой мня многих научных исследований. При его проведении автор должен выбрать метод анализа, провести сегментацию объектов исследования и проверить результаты решения на статистическую адекватность. Существует много программ для проведения кластерного анализа, одной из которых является программный комплекс статистической обработки данных - straz.

Ключевые слова. Кластерный анализ, сегментация, зерновая отрасль.

В экономической литературе есть много определений кластерного анализа, но все они понимают под собой совокупность математических методов, предназначенные для формирования относительно «отдаленных» друг от друга групп «близких» между собой объектов по информации о расстояниях или связях (мерах близости) между ними одновременно по всем наиболее существенным признакам

Кластерный анализ применяется для решения широкого спектра задач, но чаще всего речь идет именно о задаче сегментации. Все исследования, посвященные проблеме сегментации, безотносительно того, какой используется метод, имеют целью идентифицировать устойчивы группы, каждая из которых объединяет в себя объекты похожими характеристиками.

В отличие от большинства других методов многомерного анализа, кластерный анализ параллельно развивался в

нескольких дисциплинах (психология, биология, экономика...), поэтому у большинства методов существует по 2 и более названий, что существенно затрудняет взаимопонимание исследователей, в особенности, если речь идет о разных отраслях знания.

Другая проблема связана с обилием вариантов при выборе метрики и метода кластеризации, а также согласования между ними.

Выделяют две группы методов кластерного анализа: иерархические и неиерархические.

Основными методами иерархического кластерного анализа являются метод ближнего соседа, метод полной связи, метод средней связи и метод Варда. Существуют также центроидные методы и методы, использующие медиану, но их применение может привести к некоторым весьма нежелательным последствиям.

Неиерархических методов больше, хотя работают они на одних и тех же принципах. По сути, они представляют собой итеративные методы дробления исходной совокупности. В процессе деления формируются новые кластеры, и так до тех пор, пока не будет выполнено правило остановки. Между собой методы различаются выбором начальной точки, правилом формирования новых кластеров и правилом остановки. Чаще всего используется алгоритм К-средних. Он подразумевает, что аналитик заранее фиксирует количество кластеров в результирующем разбиении.

Выбирая между иерархическими и неиерархическими методами, следует обратить внимание на следующие моменты.

Неиерархические методы обнаруживают более высокую устойчивость по отношению к выбросам, неверному выбору метрики, включению незначимых переменных в базу для кластеризации и пр. Но платой за это является слово «априори». Исследователь должен заранее фиксировать ре-

зультирующее количество кластеров, правило остановки и, если на то есть основания, начальный центр кластера. Последний момент существенно отражается на эффективности работы алгоритма. Если нет оснований искусственно задать это условие, рекомендуется использовать иерархические методы. Нужно учесть также еще один момент, существенный для обеих групп алгоритмов: не всегда правильным решением является кластеризация всех наблюдений. Возможно, более аккуратным будет сначала очистить выборку от выбросов, а затем продолжить анализ. Можно также не задавать очень высокий критерий остановки (можно делать остановку, к примеру, когда кластеризовано более 90 % наблюдений).

Из информации, приведенной выше, явно прослеживается, что от аналитика в процессе применения кластерного анализа ожидается решение ряда задач. Их можно сгруппировать следующим образом:

1. Изменение исходных данных:

• выбор метрики;

• выбор метода стандартизации;

• как работать с зависимыми выборками.

2. Принятие решений:

• сколько кластеров необходимо сформировать;

• какой метод кластеризации следует использовать;

• следует ли использовать все наблюдения или необходимо исключить некоторые под выборки.

3. Анализ полученных результатов

• насколько полученное разбиение отличается от случайного;

• является ли оно надежным и стабильным на подвыборках;

• какова взаимосвязь между результатами кластеризации и переменными, не участвовавшими в процессе кластеризации;

• можно ли проинтерпретировать полученные результаты.

Для практического примера нами была проведена кластеризация муниципальных районов Нижегородской области с показателями, характеризующими производство зерна.

Для проведения данного анализа был использован программный комплекс статистической обработки данных STRAZ, разработанный научными сотрудниками Российского государственного аграрного университета - МСХА имени К. А. Тимирязева.

Вследствие ограниченности ввода в программу объектов исследования (до 100 единиц) были взяты средние показатели по 48 муниципальным районам области. При этом муниципальный район является такой единицей, в которой происходит обмен материальными ресурсами, и организации, не имеющие каких-то ресурсов, могут воспользоваться наличием их в других организациях.

В качестве оптимального деления исследуемой совокупности на кластеры была взята третья интерация, характеризующаяся наименьшими показателями коэффициента вариации (табл.).

Результаты решения показывают, что наивысшая прибыль и урожайность определяются большими показателями фондо- и энергообеспеченности - (1-й кластер). Более низкая продуктивность земельных угодий ведет к сокращению прибыли и требует меньших размеров основных и оборотных средств - (2-й и 3-й кластер). Так, урожайность в первом кластере составляет 20,88 ц с га, прибыльна 1 га посевов 89,24 руб, фондовооруженность 1 га с.-х.угодий 5122 руб., а энерговооруженность 1,46 л.с. В третьем кластере данные показатели равны: урожайность11,96 ц/га, убыток на 1 га посевов зерновых 57,58 руб., фондовооруженность и энерговооруженность га с.-х. угодий 2782 руб. и 1,04 л.с. соответственно.

Таким образом полученная информации по кластерам не противоречит экономическим законам.

Таблица 1

Результаты кластерного анализа

№ кластера	Прибыль (убыток) на 1 га посевов, руб.	Урожайность, ц/га	Приходится основных средств на 1 га с.-х. угодий, руб.	Приходится оборотных средств на 1 га с.-х. угодий, руб.	Приходится посевов зерновых на 1 сеялку, га	Приходится посевов зерновых на 1 зерноуборочный комбайн, га	Приходится с.-х. угодий на 1 автомобиль, га	Приходится л.с. на 1 га, с.-х. угодий
1	89,24	20,88	5122	3676	169	220	274	1,46
2	66,41	16,4	3351	2557	149	167	441	1,43
3	-57,58	11,96	2782	1375	160	166	418	1,04
В среднем	79,86	15,66	4496	3408	169	201	303	1,40

Однако, во многих задачах даже после того, как правило остановки рекомендовало нам прекратить дальнейшие вычисления, нет оснований считать, что полученное решение является адекватным. Результаты процедуры кластерного анализа обязательно требуют проверки с помощью формальных и неформальных тестов.

Говоря о формальных статистиках, следует рекомендовать рассчитать значение статистики «С». Однако, не следует переоценивать значение формально рассчитанных показателей: немало авторов приводит примеры, когда «хорошие»

с этой точки зрения результаты оказывались малосодержательным.

Неформальная проверка результатов кластерного анализа включает в себя такие процедуры, как анализ результатов, полученных на подвыборках, кросс-проверка на «внешних» данных, вменение порядка наблюдений, удаление небольшого количества наблюдений и повторение кластерного анализа т коротких выборках и т.д.

Список литературы

1. Орехов, Н. А.. Математические методы и модели в экономике / Н. А. Орехов, А. Г. Левин, Е. А. Горбунов. -М.: ЮНИТИ-ДАНА, 2904. - 302 с.

2. Статистическое моделирование и прогнозирование/ Под ред. А. Г. Гранберга. - М.: Финансы и статистика,1990.- 383 с.

3. www.nickartspb.ru.

THE CLUSTER ANALYSIS: THE ISSUE, ADVANTAGES AND DISADVANTAGES.

S. A. Suslov, *the candidate of economic sciences, the docent of the chair «Economics and statistics», the NGIEI*

Annotation. The cluster analysis is a basis for many scientific researches. To carry out the analysis the author should choose a method of the analysis, carry out segmentation of the objects of research and check up results of the decision on statistical adequacy. There are many programs for carrying out the cluster analysis, one of which is the program complex of statistical data processing - straz.

The key words. The cluster analysis. Segmentation. Cropsbranch.

ИТОГИ РАБОТЫ СЕЛЬСКОХОЗЯЙСТВЕННЫХ ОРГАНИЗАЦИЙ И МАЛЫХ ФОРМ ХОЗЯЙСТВОВАНИЯ В НИЖЕГОРОДСКОЙ ОБЛАСТИ

Н. П. Сидорова, *аспирант кафедры «Экономика и статистика» НГИЭИ*

Аннотация. Показана классификация групп предприятии различных форм собственности в Российской Федерации. Проанализированы основные направления работы сельского хозяйства Нижегородской области. Дана оценка работы сельскохозяйственных предприятий и малых форм хозяйствования в сельском хозяйстве Нижегородской области.

Ключевые слова. Сельскохозяйственные организации, крестьянские (фермерские) хозяйства, личные подсобные хозяйства, сельское хозяйство, Нижегородская область.

В сельском хозяйстве, в основном, создана многоукладная структура, состоящая из трех групп предприятий не равных по численности, объему производимой валовой и товарной продукции, вовлекаемых трудовых, материальных и финансовых ресурсов, эффективности ведения производства:

- сельскохозяйственные организации - коллективные и кооперативные формы собственности и хозяйствования;
- хозяйства населения (личные подсобные хозяйства), коллективные и индивидуальные сады и огороды - частные формы собственности и хозяйствования;
- крестьянские (фермерские) хозяйства и индивидуальные предприниматели - частные формы собственности и