**Методология многомерного статистического анализа**

1. *Общая оценка целесообразности применения методов многомерного статистического анализа.*

Социальные и экономические объекты, как правило, характеризуются достаточно большим числом параметров, образующих многомерные векторы, и особое значение в экономических и социальных исследованиях приобретают задачи изучения взаимосвязей между компонентами этих векторов, причем эти взаимосвязи необходимо выявлять на основании ограниченного числа многомерных наблюдений.

Многомерным статистическим анализом называется раздел математической статистики, изучающий методы сбора и обработки многомерных статистических данных, их систематизации и обработки с целью выявления характера и структуры взаимосвязей между компонентами исследуемого многомерного признака, получения практических выводов.

Методы многомерной классификации предназначены для разделения совокупностей объектов, характеризующиеся большим числом признаков, на классы, в каждый из которых должны входить объекты, в определенном смысле однородные или близкие.

Например, статистический анализ вариации показателей инновационной деятельности субъектов в большинстве случаев выявляет существенную их дифференциацию, которая может быть обусловлена различными причинами – и принадлежностью субъекта к определенной отрасли, и общим состоянием экономического потенциала территории, на которой осуществляют свою деятельность указанные субъекты, и качеством институциональной среды и т.д.

Существенная дифференциация субъектов инновационного предпринимательства по основным показателям, характеризующим инновационный потенциал, а также значительные структурные изменения в развитии инновационных процессов требуют выделения типологических групп объектов. Только в однородных группах можно качественно и адекватно оценить влияние различных факторов на инновационную деятельность, изучить присущие данным группам субъектов структурные закономерности, выделить перспективные хозяйствующие субъекты с точки зрения инновационного потенциала

Особое значение приобретают методы многомерного статистического анализа в региональных и межстрановых сопоставлениях. В этом случае, исследование пространственной структуры деятельности целесообразно проводить в типологически однородных группах регионов, имеющих общие черты и закономерности экономического развития на основе многомерных статистических и эконометрических методов.

Вспомним, что в статистических исследованиях группировка первичных данных является основным приемом решения задачи классификации, а поэтому и основой всей дальнейшей работы с собранной информацией.

Ранее было показано, что традиционно эта задача решается следующим образом. Из множества признаков, описывающих объект, отбирается один, наиболее информативный с точки зрения исследователя, и производится группировка в соответствии со значениями данного признака. Если требуется провести классификацию по нескольким признакам, ранжированным между собой по степени важности, то сначала производится классификация по первому признаку, затем каждый из полученных' классов разбивается на подклассы по второму признаку и т.д. Подобным образом строится большинство комбинационных статистических группировок.

В тех случаях, когда не представляется возможным упорядочить классификационные признаки, применяется наиболее простой метод многомерной группировки – создание интегрального показателя (индекса), функционально зависящего от исходных признаков, с последующей классификацией по этому показателю.

Развитием этого подхода является вариант классификации по нескольким обобщающим показателям (главным компонентам), полученным с помощью методов факторного или компонентного анализа.

При наличии нескольких признаков (исходных или обобщенных) задача классификации может быть решена методами кластерного анализа.

Различия между схемами решения задачи по классификации во многом определяются тем, что понимают под понятием «сходство» и «степень сходства».

После того как сформулирована цель работы, естественно попытаться определить критерии качества, целевую функцию, значения которой позволят сопоставить различные схемы классификации.

В экономических исследованиях целевая функция, как правило, должна минимизировать некоторый параметр, определенный на множестве объектов (например, целью классифицировать оборудования может явиться группировка, минимизирующая совокупность затрат времени и средств на ремонтные работы).

В случаях, когда формализовать цель задачи не удается, критерием качества классификации может служить возможность содержательной интерпретации найденных групп.

Рассмотрим следующую задачу. Пусть исследуется совокупность *n* объектов, каждый из которых характеризуется по *k* замеренным на нем признакам *Х*. Требуется разбить эту совокупность на однородные в некотором смысле группы (классы).

При этом практически отсутствует априорная информация о характере распределения измерений Х внутри классов.

Полученные в результате разбиения группы обычно называются кластерами, таксонами, образами (*cluster* (англ.) – группа элементов, характеризуемых каким-либо общим свойством; *taxon* (англ.) – систематизированная группа любой категории) методы их нахождения – кластер-анализом (соответственно численной таксономией или распознаванием образов с самообучением).

При этом необходимо с самого начала четко представить, какая из двух задач классификации подлежит решению. Если решается обычная задача типизации, то совокупность наблюдений разбивают на сравнительно небольшое число областей группирования (например, интервальный вариационный ряд в случае одномерных наблюдений) так, чтобы элементы одной такой области находились друг от друга по возможности на небольшом расстоянии.

Решение другой задачи заключается в определении естественного расслоения исходных наблюдений на четко выраженные кластеры, лежащие друг от друга на некотором расстоянии.

Если первая задача типизации всегда имеет решение, то при второй постановке может оказаться, что множество исходных наблюдений не обнаруживает естественного расслоения на кластеры, т.е. образует один кластер.

Резюмируя, можно отметить то важное обстоятельство, что при реализации эконометрических моделей, отображающих взаимосвязи между выбранными переменными (применительно к нашей области – переменными, характеризующими инновационную деятельность), часто получить адекватные зависимости не удается. Это связано во многих случаях опять-таки с тем, что выбранные для анализа совокупности субъектов инновационного предпринимательства неоднородны. Получить однородные группы объектов модно и на основе некоторых субъективных оценок, однако, более достоверным является математическое выражение близости характеристик той или иной группы к некоторому классу.

Более того, большинство методов многомерного статистического анализа данных предполагают в своей основе отсутствие мультиколлениарности между факторными признаками и, тем самым, предлагают реальные способы «борьбы» с указанным явлением.

Даже, если отвлечься от математической сущности рассматриваемых процедур, можно также констатировать, что разбиение совокупности хозяйствующих субъектов на однородные по кругу сопоставимых факторов группы имеет решающее значение для стратегического управления ими, особенно в условиях признания существенности дифференциации экономического развития на микро-, мезо- и макроуровнях и отказа от жесткого целеполагания с переходом на дифференциацию целей управления, адекватных имеющемуся экономическому и инновационному потенциалу.

2. *Методология многомерных группировок и сравнительного анализа объектов на множестве факторов, образующих общее признаковое пространство.*

Согласно общепринятой технологии проведения многомерной классификации первоначально выбранный набор факторных признаков подвергается проверке на мультиколлинеарность. Отсутствие сильной зависимости между факторными признаками является необходимым условием достоверности и надежности анализа.

Для устранения мультиколлинеарности может быть использован переход от исходных объясняющих переменных , связанных между собой достаточно тесной корреляционной зависимостью, к новым переменным, представляющим линейные комбинации исходных. При этом новые переменные должны быть слабо коррелированными либо вообще некоррелированными. Такую процедуру можно осуществить с помощью факторного анализа методом главных компонент.

Главными целями факторного анализа являются уменьшение размерности исследуемой совокупности и определение структуры взаимосвязей между переменными, т.е. классификация переменных.

Как метод сжатия факторный анализ обычно играет вспомогательную роль и служит для сокращения количества переменных до стержневого набора факторов, используемых затем кластерным анализом.

В факторном анализе предполагается, что наблюдаемые переменные являются линейной комбинацией латентных факторов. Некоторые из этих факторов допускаются общими для двух и более переменных, а другие – характерными для каждого параметра в отдельности.

Рассмотрим модель факторного анализа. Пусть задана система переменных  для каждого из регионов. Представим информацию в виде матрицы  размерности .

;

,

где  – значение -ого показателя (переменной) у -го объекта исследования;

 – -й общий фактор;

 – значение -го характерного фактора на -м объекте исследования;

 – весовой коэффициент -й переменной на -м общем факторе;

 – весовой коэффициент -й переменной на -м характерном факторе.

Факторы, связанные значимыми коэффициентами веса более чем с одной переменной, называются общими.

Множество главных компонент представляет собой удобную систему координат, а соответствующие дисперсии главных компонент характеризуют их статистические свойства. Из общего числа главных компонент для исследования, как правило, оставляют  наиболее весомых, т.е. вносящих максимальный вклад в объясняемую часть общей дисперсии.

При извлечении всех главных компонент матрица весовых коэффициентов будет квадратной порядка .

При выполнении факторного анализа существуют различные критерии значимости, которые способствуют выбору числа факторов и компонент. Некоторые из них являются альтернативными по отношению к другим, а часть этих критериев можно использовать вместе, чтобы один дополнял другой:

* *критерий Кайзера* или критерий собственных чисел. Этот критерий предложен Кайзером, и является, вероятно, наиболее широко используемым. Отбираются только факторы с собственными значениями равными или большими 1. Это означает, что если фактор не выделяет дисперсию, эквивалентную, по крайней мере, дисперсии одной переменной, то он опускается;
* *критерий каменистой осыпи* или критерий отсеивания. Он является графическим методом, впервые предложенным психологом Кэттелом. Собственные значения можно изобразить в виде простого графика. Кэттел предложил найти такое место на графике, где убывание собственных значений слева направо максимально замедляется (см. рис.). Предполагается, что справа от этой точки находится только «факториальная осыпь». Однако этот критерий отличается высокой субъективностью и, в отличие от предыдущего критерия, статистически необоснован;
* *критерий доли воспроизводимой дисперсии*. Факторы ранжируются по доле детерминируемой дисперсии, когда процент дисперсии оказывается несущественным, выделение следует остановить. Желательно, чтобы выделенные факторы объясняли более 80 % разброса. Главные факторы должны вместе объяснять не меньше 50,1 % дисперсии.

Отметим, что все указанные критерии, да и сама процедура факторного анализа полностью «автоматизирована» и реализуется в ППП «Statistica», «SPSS».

Если результаты факторного анализа удовлетворительны, то необходимо рассчитать новый набор переменных. Новые переменные вычисляются умножением исходных переменных на набор весовых коэффициентов, полученных из коэффициентов нагрузок. Далее осуществляется переход непосредственно к процедуре кластерного анализа.

Большое достоинство кластерного анализа в том, что он позволяет производить разбиение объектов не по одному параметру, а по целому набору признаков. Кроме того, кластерный анализ в отличие от большинства математико-статистических методов не накладывает никаких ограничений на вид рассматриваемых объектов, и позволяет рассматривать множество исходных данных практически произвольной природы. Это имеет большое значение, например, для прогнозирования конъюнктуры, когда показатели имеют разнообразный вид, затрудняющий применение традиционных эконометрических подходов.

Кластерный анализ позволяет рассматривать достаточно большой объем информации и резко сокращать, сжимать большие массивы социально-экономической информации, делать их компактными и наглядными.

Обычно формой представления данных в задачах кластерного анализа служит прямоугольная таблица:

,

где каждая строка представляет результат измерений  рассматриваемых признаков на одном из обследованных объектов. В конкретных случаях может представлять интерес, как группировка объектов, так и группировка признаков. В тех случаях, когда разница между двумя этими задачами несущественна, можно пользоваться понятием «объект», включая в это понятие и «признак».

Матрица  не является единственным способом представления данных в задачах кластерного анализа. Иногда исходная матрица задана в виде квадратной матрицы:

; ,

где элемент  определяет степень близости -го объекта к -му.

Если данные представлены в форме , то первым этапом решения задачи поиска кластеров будет выбор способа вычислений расстояний или близости, между объектами или признаками. Мерами близости в этом случае служат различные статистические коэффициенты связи:

– *Евклидово расстояние*.Оно является геометрическим расстоянием в многомерном пространстве и вычисляется следующим образом:

,

где  ,  – величина -ой компоненты и -ого (-ого ) объекта.

Заметим, что Евклидово расстояние (и его квадрат) вычисляется по исходным, а не по стандартизованным данным.;

– *нормализованное Евклидово расстояние*. Применяется в тех случаях, когда каждой компоненте  вектора наблюдений  удаётся приписать некоторый вес , пропорционально степени важности признака в задаче классификации. Оно более подходит для переменных, измеренных в различных единицах или сильно различающихся по величине:

;

– *расстояние городских кварталов (манхэттенское расстояние)*.Это расстояние является просто средним разностей по координатам. В большинстве случаев эта мера расстояния приводит к таким же результатам, как и для обычного расстояния Евклида.

Существует множество методов объединения кластеров[[1]](#footnote-1).

1. *Одиночная связь (метод ближайшего соседа*). В этом методе расстояние между двумя кластерами определяется расстоянием между двумя наиболее близкими объектами (ближайшими соседями) в различных кластерах. Это правило должно, в известном смысле, нанизыватьобъекты вместе для формирования кластеров, и результирующие кластеры имеют тенденцию быть представленными длинными "цепочками".
2. *Полная связь (метод наиболее удаленных соседей)*. В этом методе расстояния между кластерами определяются наибольшим расстоянием между любыми двумя объектами в различных кластерах (т.е. "наиболее удаленными соседями"). Если же кластеры имеют в некотором роде удлиненную форму или их естественный тип является "цепочечным", то этот метод непригоден.
3. *Невзвешенное попарное среднее*.В этом методе расстояние между двумя различными кластерами вычисляется как среднее расстояние между всеми парами объектов в них. Метод эффективен, когда объекты в действительности формируют различные "рощи", однако он работает одинаково хорошо и в случаях протяженных ("цепочного" типа) кластеров.
4. *Невзвешенный центроидный метод*.В этом методе расстояние между двумя кластерами определяется как расстояние между их центрами тяжести.
5. *Взвешенный центроидный метод (медиана)*.Этот метод идентичен предыдущему, за исключением того, что при вычислениях используются веса для учёта разницы между размерами кластеров (т.е. числами объектов в них). Поэтому, если имеются (или подозреваются) значительные отличия в размерах кластеров, этот метод оказывается предпочтительнее предыдущего.
6. *Метод Варда*.Этот метод отличается от всех других методов, поскольку он использует методы дисперсионного анализа для оценки расстояний между кластерами. Метод минимизирует сумму квадратов для любых двух (гипотетических) кластеров, которые могут быть сформированы на каждом шаге. В целом метод представляется очень эффективным, однако он стремится создавать кластеры малого размера.

Для сравнения качества разбиения на классы используется ряд функционалов качества. Наиболее употребляемые из них:

* + суммавнутригрупповых дисперсий



* + сумма попарных внутриклассовых расстояний между элементами



где  – заданное количество кластеров.

Число алгоритмов кластерного анализа слишком велико. Все их можно подразделить на иерархические и неиерархические.

**Иерархические** алгоритмы связаны с построением дендрограмм и делятся на:

а) *агломеративные*, характеризуемые последовательным объединением исходных элементов и соответствующим уменьшением числа кластеров;

б) *дивизимные* (делимые), в которых число кластеров возрастает, начиная с одного, в результате чего образуется последовательность расщепляющих групп.

При большом количестве наблюдений иерархические методы кластерного анализа не пригодны. В таких случаях используют неиерархические методы, основанные на разделении, которые представляют собой итеративные методы дробления исходной совокупности. В процессе деления новые кластеры формируются до тех пор, пока не будет выполнено правило остановки.

Наиболее распространен среди неиерархических методов *алгоритм k-средних*, также называемый быстрым кластерным анализом

Алгоритм k-средних строит k кластеров, расположенных на возможно больших расстояниях друг от друга. Основной тип задач, которые решает алгоритм k-средних, - наличие предположений (гипотез) относительно числа кластеров, при этом они должны быть различны настолько, насколько это возможно. Выбор числа k может базироваться на результатах предшествующих исследований, теоретических соображениях или интуиции.

После получений результатов кластерного анализа методом k-средних следует проверить правильность кластеризации (т.е. оценить, насколько кластеры отличаются друг от друга). Для этого рассчитываются средние значения для каждого кластера. При хорошей кластеризации должны быть получены сильно отличающиеся средние для всех измерений или хотя бы большей их части.

В настоящее время, несмотря на широту средств современной статистики, разнообразие методических подходов по анализу  и оценке инновационного потенциала единой научно-обоснованной методики оценки уровня инновационного потенциала не существует,  отсутствует единый методологический подход и единая система показателей. Вместе с тем, рассмотренные методы многомерного статистического анализа имеют надежную перспективу быть одним из ключевых инструментов выявления однородных по кругу сопоставимых признаков субъектов инновационной деятельности в разработке стратегических управленческих решений для развития опорных отраслевых комплексов.

1. Дубров, А.М. Многомерные статистические методы / А.М. Дубров, В.С. Мхитарян, Л.И. Трошин. – М.: Финансы и Статистика, 2000. – 362 с. [↑](#footnote-ref-1)