Bènjamin S. Duran, Patrick L. Odèll

CLUSTER ANALYSIS
A SURVEY

SPRINGER — VERLAG
BERLIN — HEIDELBERG — NEW YORK 1974

Б. Дюран и П. Оделл

КЛАСТЕРНЫЙ АНАЛИЗ

Перевод с английского *Е. З. Демиденко*
Научное редактирование и предисловие
А. Я. Боярского

МОСКВА „СТАТИСТИКА“ 1977

517.8
Д97

Дюран Б. и Оделл П.

Д97 Кластерный анализ. Пер. с англ. Е. З. Демиденко. Под ред. А. Я. Боярского. Предисловие А. Я. Боярского. М., «Статистика», 1977.

128 с. с ил.

Тема книги — обзор состояния теории и практики применения «кластерного анализа». Этот метод имеет все преимущества метода комбинационной группировки, но свободен от его главного недостатка — распыления материала, что открывает широкие перспективы применения рассматриваемого метода в статистическом анализе, в классификации объектов, в исследовании связей, типизации выборки и др. Книга отличается полнотой, доступностью и вместе с тем краткостью изложения.

Книга рассчитана на статистиков, экономистов, а также социологов, демографов, биологов и других специалистов.

517.8

Д $\frac{10803*-036}{008(01)-77}$ 40-77

* Второй индекс 10805.

О МЕТОДОЛОГИЧЕСКИХ ПРИНЦИПАХ И МНОГОМЕРНОМ АНАЛИЗЕ

(вместо предисловия)

Б. С. Ястремский, классифицируя задачи математической статистики, представил их в виде двух разрезов: статика — динамика и одномерные — многомерные задачи. Переход ко второй, более сложной ступени в обоих аспектах связан с положениями диалектики, требующей рассмотрения явлений в их развитии (динамике) и взаимосвязи, что и ведет к многомерным задачам. Не касаясь здесь первого аспекта, остановимся на втором. В область многомерных задач статистик вступает, как только он принимается за изучение совместной вариации двух признаков, т. е. их связи. Безразлично, идет ли при этом речь об аналитических группировках, комбинационной таблице по двум признакам с подсчетом числа случаев или о математических методах корреляции, дисперсионного анализа и т. п.

Развитие анализа в этом направлении приводит к рассмотрению взаимосвязи не пары признаков, а большего их числа. В области элементарных приемов это получает выражение в сложных комбинационных группировках с развитым сказуемым. В области математико-статистических методов речь идет о множественной корреляции, дисперсионном анализе зависимости от нескольких переменных и т. д.

На этом пути по мере углубления исследования рассматриваются связи все большего числа признаков. Но здесь возникают трудности технического характера. В свое время считались великолепным достижением исследования зависимостей, в которых количество одновременно учитываемых аргументов доводилось до десятка или больше. Старая литература изобилует примерами раздельного изучения зависимостей некоторого признака от каждого из большого числа других, а то и

просто перечнями влияющих признаков. Изучение их влияния в комплексе наталкивалось на два препятствия: технические трудности и ограниченность материала наблюдения. Второе чаще имеет место в естественнонаучных исследованиях и нередко оказывается снятым в социально-экономической статистике. В то время как экспериментатор должен получить значимые результаты из наблюдений над десятком кроликов, исследователь бюджета рабочей семьи располагает десятками тысяч наблюдений или социал-гигиенист — тысячами «историй болезни». Впрочем, при территориальной дифференциации или комбинации большого числа признаков не помогают и эти тысячи. В преодолении первого препятствия существенную роль играет новая вычислительная техника. Для ЭВМ получение, например, уравнения связи по сотне признаков не составляет проблемы.

Однако при старой технике роль качественного анализа (специфической логики специфического предмета) видна с первых же шагов выбора аргументов, поскольку драгоценные места для них в уравнении надо было расходовать с большой оглядкой, не допуская их траты на малосущественные связи. В условиях новой техники это соображение может показаться потерявшим значение. Для нее не так важно, если в числе сотни введенных в уравнение связи аргументов десяток окажется бесполезным, не влияющим на результативный признак: их введение не вытесняет из рассмотрения других, действительно важных.

Однако вскоре опыт показал, что голый эмпиризм остается тем, чем он был. Правда, если по его поводу Энгельс сказал, что и слепая свинья может найти свой желудь, то вооруженная современной электронной техникой она может найти целую горсть желудей. Но от этого в принципе ничего не меняется — эмпиризм остается эмпиризмом со своими возможностями и своей ограниченностью.

Сказанное мало интересует прагматически настроенных исследователей. Зато они ясно ощущают, что по мере возрастания размерности задачи все более теряется обозримость результатов. Закономерность распыляется на множество зачастую малозначащих связей.

С позиций эмпиризма сам переход от индивидуальных значений к обобщенным характеристикам есть реализация принципа «экономии мышления», а неимоверное возрастание числа этих характеристик при комплексном рассмотрении взаимосвязей многих признаков оказывается в полном противоречии с этим принципом. Исследователь снова стоит перед лицом огромной массы индивидуальных наблюдений. Так возникает задача обратного сведения множества характеристик к небольшому ряду обобщающих итогов, выражающему действительно существенное, закономерное для явления. Но пока каждый вовлеченный в анализ признак остается отдельным самостоятельным элементом, со своими характеристиками и линиями связи, число параметров, выражающих результаты обработки, не поддается уменьшению. Единственный путь к нему — либо в отсечении большинства признаков и возвращении к малоразмерным классическим задачам, либо в объединении признаков, в замене целых «гроздей» их одним, неизбежно искусственно построенным на их основе. Так появляется направление, получившее название «многомерный анализ». Его развитие и составляет новую ступень в истории математической статистики, которой отмечены последние десятилетия. Из предыдущего ясна и его связь с могущественной вычислительной техникой.

В многомерном анализе образовались разделы, которые, однако, не изолированы, а проникают и переходят один в другой. Это кластерный анализ (которому посвящена данная книга), таксономия, распознавание образов, метод главных компонент, факторный анализ. Наиболее ярко отражают черты многомерного анализа в классификации объектов кластерный анализ, а в исследовании связи — факторный анализ. Все эти разделы, закономерно обусловленные развитием математико-статистических методов и практики их применения, несомненно, несут в себе новые богатые возможности для решения многих познавательных задач, какими не располагают «классические» методы. В то же время их необходимо подвергнуть теоретическому анализу с позиций методологических принципов марксистско-ленинской теории познания. Основополагающим гносеологическим принципом статистической науки является примат качества объекта, его специфической логики. Усовершенствование формальных методов эмпирического исследования не может в какой-либо мере поколебать этот принцип. Методы многомерного анализа пока что представляются, однако, эмпиризмом, сбросившим всякие

оковы этой логики качества материального объекта (или, как теперь выражаются, «содержательного» анализа). Не случайно, например, факторный анализ возник в такой области, как психология, где для него почва особенно благоприятна, поскольку раскрытие внутренней логики в ней исключительно трудно. Задача в том, чтобы, используя приемы многомерного анализа, согласовать их теоретическую трактовку и применение с основным методологическим принципом статистического исследования. Без этого труд людей и дорогих машин легко окажется малоэффективным. Это часто видно, например, в факторном анализе, когда дело доходит до главной фазы исследования — выводов. Получив результаты вычислений, исследователю предстоит их «проинтерпретировать», иначе говоря, выяснить их смысл (экономический или иной). Выходит то, что должно быть в самом начале, выступает на сцену лишь в конце.

Сказанное можно вполне отнести и к кластерному анализу. Наиболее существенные его методологические черты сводятся к двум: образование единой меры, охватывающей ряд признаков, и чисто количественное решение вопроса о группировке объектов наблюдения.

Идея классификации по сочетанию ряда признаков не нуждается в аргументации. Как раз в одной из наиболее популярных задач кластерного анализа — группировке районов — она давно признана. Еще в 1920 г., анализируя «Связь между элементами крестьянского хозяйства в 1917 и 1919 годах» («Вестник статистики», 1920, с. 19—21), Б. С. Ястремский рассматривал 34 характеристики уездов, влиявшие на эту связь. Можно привести и другие примеры группировки территориальных единиц, по комплексу признаков, неизменно имевших место в задачах районирования. Но в кластерном анализе признаки объединяются с помощью некоторой «метрики» в один количественный показатель сходства (различия) группируемых объектов. Казалось бы, достаточно запустить в ЭВМ массив информации о них и подходящую программу классификации. На самом деле, однако, без предварительного анализа качества нельзя и приступить к делу. Уже в определении самого перечня признаков он неизбежно присутствует. Иной, быть может, скажет на это, что надо попросту ввести в ЭВМ весь материал наблюдения. Но ведь кто-то на основании чего-то составил и саму программу наблюдения.

Стоит себе отдать в этом ясный отчет и все становится на свое место, настолько, что в этой части кластерный анализ оказывается сродни идеям таких классических исследований, как ленинские группировки крестьянских хозяйств, выявившие два класса капиталистической экономики на полюсах и еще не размытую дифференциацией середину. Капиталистическая верхушка — это хозяйства эксплуататоров. Но эксплуатация могла осуществляться в разных формах: наем работников, «прокат» инвентаря, займы и т. д. Для этого хозяйство должно было чем-то располагать: землей, инвентарем, деньгами, мастерской или торговлей. Как видим, принадлежность к этой группе внешне могла получать отражение в ряде признаков. Отсюда ленинское требование изучать совокупность признаков, давать по ним группировку не параллельную, а комбинационную. При этом важность признаков зависит от особенностей района: в земледельческом—это прежде всего площадь посева, в животноводческом — численность скота и т. д. Легко видеть, что как сама задача, так и необходимость учета в ее решении ряда признаков и выбора этих признаков — все это продиктовано качественным анализом и, как всегда в статистике, через него тесными узами связано с целью исследования. Наверно, если бы цель состояла не в анализе классовой дифференциации, а в анализе уровня культуры, ведущее место заняли бы грамотность, число лет обучения, наличие в доме книг и т. п., а посевная площадь или число лошадей заняли бы место вспомогательной информации для выявления связи между уровнем культуры и социально-экономическим фактором.

Коль скоро признаки отобраны, может быть оправданным и подход кластерного анализа, но не как чисто эмпирического, а основанного на правильных методологических принципах. Так, именно из качественного анализа вытекает, что в составе капиталистической верхушки могли быть хозяйства, обрабатывающие большой земельный массив и без большого числа лошадей, и одновременно хозяйства без больших посевов, но богатые живым и мертвым инвентарем или имеющие торговлю и т. д. Следуя указаниям качественного анализа, их надо объединить в одну группу.

В кластерном анализе группировочные признаки подвергаются объединению с помощью некоторой «метри-

ки» — евклидова расстояния или иной. Но здесь возникает самое настоящее embarras de richesse, затруднение от изобилия. Метрик оказывается много и число их возрастает. Какой отдать предпочтение? Кроме того, в частности, в евклидовой результат зависит от масштаба, от выбранных единиц измерения, например, будет ли один признак измеряться в метрах, другой в килограммах или первый в сантиметрах, а второй в тоннах. Это обстоятельство вскользь отмечает и автор данной книги. Правда, есть способ выйти из затруднения путем нормирования признаков. Но нельзя доказать, что для всех признаков одно квадратическое отклонение одинаково значимо.

Вопрос о выборе метрики и масштабов имеет различное содержание в зависимости от целей. Если группировки различаются на «типологические» и «аналитические» (не настаиваем на этой терминологии), то же самое не может не относиться и к кластеризации. Между тем в литературе это игнорируется. Более того, кластеры выдаются обычно за «типы», что должно в какой-то мере подчеркивать их существенное различие, чуть ли не качественное.

Если речь идет о качестве в подлинном смысле слова, то ни метрики, ни масштабы не произвольны. Так, для выделения верхней группы крестьянских хозяйств надо было учитывать ряд признаков, но так, чтобы их сочетание давало основание для причисления хозяйства к этой группе. Критерием могла бы быть совокупная величина возможного дохода. Поставив для нее некоторую нижнюю границу, отвечающую возможности основывать хозяйство на прибавочной стоимости без участия в нем личным трудом, мы бы получили объективный критерий для масштабов, да и для метрики. В одних случаях такой подход не слишком труден. Например, мощность тракторов и число лошадей сравнительно легко бы поддавались соизмерению. В других задача гораздо труднее. Но сказанное должно служить ориентиром во всех случаях типологической кластеризации (если можно так выразиться).

Другое дело формально-количественная кластеризация. Ее цели скромнее: представить в сжатом виде массив информации с его многомерностью, но так, чтобы потеря информации не была чрезмерной. Здесь нет жестких объективных требований и решение может быть

различным. То же можно сказать по поводу любой «аналитической» группировки. Общность вопроса вытекает уже из того, что группировка по одному признаку и кластеризация по ряду признаков приводятся друг к другу. Число соединяемых при кластеризации признаков может быть равным и единице. Это приводит задачу группировки по одному признаку к кластеризации. С другой стороны, используемая при объединении признаков метрика сводит их к одному признаку и далее разбиение на кластеры равнозначно группировке по этому признаку. О путях формализации последней задачи уже немало сказано в литературе. Внедрение ЭВМ и перевод обработки информации на индустриальные рельсы не может оставить на субъективный произвол число и границы интервалов группировки. Значит, неизбежно применение в этом некоторого формального стандарта. Однако таких стандартов может быть несколько: разбиение по децилям, по квадратическим отклонениям, по максимуму «локального расстояния», по относительному расстоянию, по внутрикластерному коэффициенту вариации и т. д. Индустриальный подход, таким образом, не исключает инициативы исследователя, его выбора. Но этот выбор теперь будет состоять в выборе между несколькими стандартами, для которых имеются машинные программы. Это несколько ограничивает исследователя, но дает возможность гораздо большей сравнимости разных группировок и их быстрого получения.

Выходит, что методы кластеризации нужны при внедрении ЭВМ даже для решения задачи простой группировки. Поскольку в ней нет качественного критерия, все сводится к образованию групп по количественному сходству. А в такой постановке машина с помощью той или иной стандартной программы может с ней справиться лучше.

Из всего сказанного ясно, что по отношению к кластерному анализу, как и к другим частям многомерного анализа, необходимо, во-первых, хорошо изучить теорию и имеющуюся практику применения, во-вторых, на основе этого и все увеличивающегося нового опыта применения глубоко осмыслить его технику с позиций общих методологических принципов статистической науки.

В достижении первой цели предлагаемая книга представляет большую ценность, так как в ней читатель най-

дет богатый и в то же время сжато изложенный материал, образующий в целом прекрасный обзор теории кластерного анализа и ряда его приложений. Именно эту цель и ставили перед собой авторы и они прекрасно справились с делом. Что касается второго, то эту задачу мы здесь, конечно, могли только поставить. Она должна быть решена не столько математиками, сколько статистиками, экономистами и представителями других конкретных областей применения, не без участия философов.

В целом же книга заслуживает высокой оценки не только как монография, но и как пособие учебного характера. Хотя некоторые ее места воспринимаются не без известного труда, в целом она отличается от многих других книг по этой или примыкающим проблемам ясностью и доступностью изложения. Ее появление на русском языке несомненно принесет большую пользу советским специалистам и всем интересующимся статистической наукой.

Редактор взял на себя смелость исправить некоторые явные опечатки оригинала.

А. Я. БОЯРСКИЙ

ПРЕДИСЛОВИЕ

За последнее тридцатилетие в области кластерного анализа была проделана большая работа, причем значительная ее часть была проведена после 1960 г. Основное содержание этой книги было опубликовано в различных журналах, в том числе прикладного характера, однако до сих пор этот материал не был собран воедино.

Цель данной монографии заключается в том, чтобы объединить разрозненные статьи в виде краткого обзора по кластерному анализу.

Мы надеемся, что эта книга позволит читателю быстро ознакомиться с проблемами кластерного анализа и другими смежными вопросами.

По этой причине многие детали были опущены. Это же касается иллюстрирующих примеров. Большинство работ, на которые мы ссылаемся, содержат примеры применения кластерного анализа, поэтому читатель может воспользоваться ими для получения дополнительной информации по специальным вопросам. Мы постарались включить в библиографию все работы, которые сыграли какую-либо роль в развитии «теории» кластерного анализа. Этот список содержит также работы прикладного характера, однако наша библиография все же далеко не полная.

Эта монография была написана в значительной мере под влиянием работ многих исследователей в данной области; это в первую очередь относится к работам Хартигана, Уишарта, Брайена, Дженсена, Вайнода и Рао.

Изложение некоторых частей книги основано на исследовании, выполненном при поддержке Центра пилотируемых космических кораблей НАСА (отдел наземного наблюдения, контракт NAS 9—12775).

[397] Wishart D. An algorithm for hierarchical classifications, Biometrics, Vol. 22, No. 1, (1969), 165—170.
[398] Wishart D. FORTRAN II Programs for 8 Methods of Cluster Analysis (CLUSTAN I), Computer Contribution 38, State Geological Survey, The University of Kansas, Lawrence, (1969).
[399] Wishart D. A fortran II program for numerical classification, St. Andrew's University, Scotland, (1968).
[400] Wishart D. Numerical classification method for deriving natural classes, Nature, London, (1969), 221, 97—98.
[401] Wolf D. E. PROMENADE: Complete Listing of PROMENADE Programs, Appendix 9d to RADC-TR-68-572, Stanford Research Institute, Menlo Park, Calif., 465 pp., (1968).
[402] Wolfe J. H. A computer program for the maximum likelihood analysis of types, Tech. Bulletin, 65—15, U. S. Naval Personnel Research Activity, San Diego, Calif., (May, 1965).
[403] Wolfe John H. NORMIX-Computational methods for estimating the parameters of multivariate normal mixtures of distributions, Tech. Report, U. S. Naval Personnel Research Activity, San Diego, Calif., (Aug., 1967), 1—31.
[404] Wolfe J. H. Pattern clustering by multivariate mixture analysis, Multivariate Behavioral Research, Vol. 5, No. 3, (1970), 329—350.
[405] Young G. Factor analysis and the index of clustering, Psychometrika, Vol. 4, No. 3, (Sept., 1939).
[406] Yule G. U. On measuring associations between attributes, J. Roy. Statist. Soc., Vol. 75, (1912), 579—642.
[407] Zadeh L. A. Fuzzy sets, Information and Control, Vol. 8, (1965), 338—353.
[408] Zahn C. T. Graph — theoretical methods for detecting and describing gestalt clusters, IEEE Trans. on Computers, Vol. C-20, No. 1, (1971), 68—86.
[409] Remote multispectral sensing in agriculture, Laboratory for Applications of Remote Sensing, Purdue University, Lafayette, Ind., Annual Report, Vol. 4, Research Bulletin 873, (Dec., 1970).

Литература, добавленная при переводе

С. А. Айвазян, З. И. Бежаева, О. В. Староверов. Классификация многомерных наблюдений. М., «Статистика», 1974.
Многомерный статистический анализ в социально-экономических исследованиях. М., «Наука», 1974, гл. II.

ОГЛАВЛЕНИЕ

Дюран Б. и Оделл П.

КЛАСТЕРНЫЙ АНАЛИЗ

Редактор *З. А. Сумник*, Мл. редактор *О. В. Степанченко*
Техн. редактор *В. А. Чуракова*. Корректор *Г. А. Башарина*
Худ. редактор *Н. А. Володина*
Обложка художника *Г. Г. Васильевой*

ИБ № 345

Сдано в набор 12/X 1976 г. Подписано к печати 21/I 1977 г. Формат бумаги 84×108$^{1}/_{32}$. Бумага № 3. Объем 4,0 печ. л. Уч.-изд. л. 7,05. Усл. п. л. 6,72. Тираж 11 000 экз. (Тематич. план 1977 г. № 40). Заказ № 6796. Цена 42 коп.

Издательство «Статистика», Москва, ул. Кирова, 39.

Областная типография управления издательств, полиграфии и книжной торговли Ивановского облисполкома, г. Иваново-8; ул. Типографская, 6.