А.Г. Буховец
(*Воронеж*)

СИСТЕМНАЯ ИНТЕРПРЕТАЦИЯ РЕЗУЛЬТАТОВ КЛАССИФИКАЦИОННЫХ ЗАДАЧ

Результаты применения методов кластерного анализа интерпретируются в соответствии с положениями системного подхода. В рамках вторичного анализа проводится изучение ранговых распределений, что позволяет сделать вывод о системности объекта исследования.

Ключевые слова: методы многомерной классификации, кластерный анализ, системный анализ, ранговые распределения, типология миграционного поведения, вторичный анализ.

Методология системного подхода, окончательно сформировавшаяся в 70-х гг. XX в., в настоящее время уже считается общепризнанной [1]. Однако использование положений системного анализа в практических задачах часто носит декларативный характер, а результаты такого применения обычно бывают представлены на качественном (вербальном) уровне.

Задачу классификации нередко рассматривают как задачу системного представления объекта исследования. В этом плане отметим, что довольно часто даже термины «классификация» и «систематизация» употребляются как синонимы. На самом деле классификационные и системные представления сосуществуют и взаимно дополняют друг друга настолько часто, что происходит нередко смешение классификационных и системных методов исследования.

Основная цель нашего исследования заключается в том, чтобы показать, что в рамках системного подхода имеется возможность

Алексей Георгиевич Буховец – кандидат экономических наук, доцент Воронежского государственного агроуниверситета.

дать объяснение некоторым закономерностям, связанным с количественными оценками полученных классификационных разбиений, которые выражаются через параметры ранговых распределений.

Использование принципов системного подхода является важным элементом моделирования социально-экономических процессов. Однако корректное применение основных положений системного анализа невозможно без предварительного установления целостности системы. Иногда эту проблему формулируют как задачу выделения системы, определения ее границ. Одним из методов решения этой задачи является рассмотренный подход, основанный на исследовании ранговых разбиений, которые получаются в ходе построения типологий. В рамках предложенного подхода удается не только получить содержательно интерпретируемые классификации элементов системы, но и разрешить вопрос о целостности всей рассматриваемой совокупности.

Основные принципы системного подхода в задачах классификации

На протяжении второй половины прошлого века в различных областях науки осуществлялся переход от механистической к системной парадигме. Основное противоречие заключалось в выяснении взаимоотношений частей и целого. Акцент на выделенные части совокупности получил название *механистического*, или *редукционистского подхода*. В противоположность ему акцент на целое характерен для холистического или системного взгляда. Второй подход в науке принято называть системным, а анализ, который базируется на таких взглядах, – системным анализом. Согласно этому подходу существенными свойствами системы являются свойства целого, которыми не обладает ни одна из его частей. Такие свойства возникают из взаимодействий и связей между отдельными частями. Эти свойства нарушаются, когда система рассекается на отдельные изолированные элементы.

Механистическая парадигма сводилась к утверждению, что в любой сложной системе поведение всей системы как целого может быть полностью объяснено на основе свойств ее частей. В системном подходе приоритет имеет целостность системы, а уж затем рассматриваются ее составляющие элементы. Систему нельзя понять только посредством анализа ее частей. При системном подходе свойства частей могут быть выведены только из организации целого. Свойства частей системы не являются их внутренними свойствами, и они могут быть поняты и осмыслены лишь в контексте всего целого. Соответственно, системный анализ акцентирует внимание в первую очередь на организации множества.

Задача классификации, сформулированная на общетеоретическом уровне, выражается как задача разделения заданного множества объектов на качественно однородные группы (классы) [2]. Совокупность групп, полученных в результате применения классификационных процедур, принято называть *результирующим разбиением*. В некоторых современных научных построениях господствующим является представление о классификации как системе знаний, дающей одновременно системное представление объектов. Классификация рассматривается как такое упорядочение множества объектов, которое позволяет делать заключения относительно фактов, не содержавшихся в первичном представлении этих объектов.

Но принятие тезиса о системности рассматриваемой совокупности ведет к необходимости принятия и следствий этого тезиса. Для исследователя системность объекта проявляется в том, что возникает возможность адекватно, хотя и приближенно, описывать этот объект и его составляющие ограниченным набором переменных (признаков).

Внешняя, или классификационная целостность системы, по мнению авторов [3, с. 69], определяется как «возможность естественного объединения в классы заранее имеющихся объектов. Общность этих объектов состоит в наличии у них единой природы,

позволяющей естественным образом сопоставлять между собой эти объекты и образовывать из них естественные классы». Критерий внутренней организации для выделения системы как отличительный признак, позволяющий разграничивать систему и случайный конгломерат объектов, выдвигал также известный специалист в области систематики А.А. Любищев.

Из всех проблем, возникающих при использовании принципов системного подхода при построении типологизации, мы в нашей работе ограничимся в полной мере рассмотрением только одного аспекта, связанного с исследованием получающихся при классификации ранговых разбиений [4].

Под ранговым распределением понимают зависимость численности, соответствующей данному элементу, от его ранга при ранжировании элементов по численности. В работах социально-экономического характера, так или иначе связанных с типологизацией, обычно не уделяется должного внимания анализу результатов классификации с точки зрения исследования ранговых распределений. Вместе с тем распределение численностей классов в построенном классификационном разбиении может служить, как будет показано ниже, некоторым числовым показателем целостности системы. Особенно следует подчеркнуть, что этот показатель не является непосредственно измеримым, а получен в результате обработки первичной информации, – практически он связан со структурными особенностями многомерных данных и проявлением этих особенностей в числовой характеристике построенного классификационного разбиения.

Ранговые распределения независимо друг от друга исследовались в различных научных областях. Удивительным оказалось то, что при весьма общих ограничениях, связанных со свойствами системности рассматриваемых совокупностей, ранговые распределения подчинялись одному и тому же типу зависимостей. В наиболее простой форме эта зависимость может быть представлена гиперболой, и поэтому в математической статистике такие

законы получили название *гиперболических законов распределения*. Математически эта зависимость может быть выражена следующим образом:

$$n_i = \frac{C}{i^{1+\alpha}}, \qquad (1)$$

где $i = 1, 2, \ldots K$ – ранг класса; C – постоянная величина, обычно равная объему наибольшего (модального) класса (обычно полагают $C \approx n_1$); n_i – объем (численность, частость) класса i-го ранга; α – некоторая постоянная положительная величина, обычно не превосходящая единицы.

Примеры ранговых распределений, представленных в указанной выше форме, можно найти в экономике, географии, биологии, лингвистике, наукометрии, информатике, политологии и многих других областях. Ссылки на указанные факты приводятся в работах [3; 5].

Зависимости указанного выше вида в экономике обычно называют законом Парето, в географии – законом Зипфа, в биологии – законом Уилкса, в информатике – законом Бредфорда, в лингвистике – законом Ципфа–Мандельброта. Различия в названиях отражают лишь тот факт, что получившие статус эмпирического закона зависимости были получены разными исследователями независимо друг от друга в различных областях. При этом Г.К. Ципф был одним из первых, кто не только обнаружил выполнение на эмпирическом уровне этого закона, но и предложил объяснение его механизма формирования [5]. Поэтому в литературе ранговые распределения такого вида чаще всего называют *ципфовыми*. Будем придерживаться этого термина в дальнейшем и мы.

Закон Ципфа как общесистемная универсальная характерная закономерность был принят во многих областях. Так, в лингвистике было показано, что для законченных текстов, образующих некоторую лексическую единицу и несущих смысловую нагрузку, выполняется закон Ципфа. И наоборот, на отдельных фрагментах текста, или на совокупности различных текстов, эта закономерность

нарушается. Выполнение ципфового распределения было использовано в качестве критерия для оценки целостности некоторых старинных текстов. В наукометрии одним из критериев наличия сформировавшегося научного направления предлагается считать выполнение ципфового рангового распределения на совокупности публикаций по тематике этого научного направления. В последнее время было предложено оценивать уровень фальсификации результатов выборов по мере отклонения представленных официально результатов от ципфового распределения.

Существование закономерности, выражающейся одинаковой математической зависимостью для столь различных областей, позволило выдвинуть предположение, что эта зависимость может представлять некоторый общесистемный принцип, точнее – являться следствием выполнения такого принципа для системной совокупности объектов. Как известно, все замкнутые системы характеризуются наличием некоторых инвариант, выраженных обычно в форме законов сохранения. Закон Ципфа, переписанный в форме $n_i i^{1+\alpha} = C$, можно, очевидно, интерпретировать как некоторый закон сохранения, реализуемый в данной системе. Кстати, заметим, что в такой форме закон первоначально и был сформулирован.

Таким образом, в качестве необходимого формального признака системности (целостности) совокупности объектов нами предлагается использовать наличие ципфового распределения на этой совокупности. Очевидно, что этот признак не является единственным и достаточным, – он, безусловно, должен быть дополнен качественным анализом системообразующей совокупности.

Закон Ципфа можно, конечно, рассматривать в качестве аксиомы для некоторых целостных систем и не ставить вопрос о механизмах его возникновения. Но такой подход заведомо становится феноменологическим и не может претендовать на роль объяснения полученных результатов. Более привлекательным представляется подход, при котором имеется возможность объяснить

эмпирическую закономерность исходя из каких-то более общих системных принципов. Кроме этого, мы хотим обратить внимание на то, что нами предлагается рассматривать ранговые распределения, которые получаются в результате исследования структуры многомерных данных методами многомерной классификации, т.е. рассматривать ранговое распределение как некоторую характеристику самой многомерной структуры.

Механизмы формирования ципфового распределения

В задачах классификации, как правило, не оговариваются свойства исследуемых совокупностей, – генеральных или выборочных. Обычно не указывается, является ли рассматриваемая совокупность объектов множеством (конгломератом), отношение к которому определяется наличием у объекта какого-либо одного или нескольких свойств, или же выбранная совокупность является системой, т.е. совокупностью объектов, взаимодействие которых вызывает появление новых интегральных качеств, не свойственных отдельно взятым объектам. Как следствие, не исследуются и в дальнейшем не используются многие системные (эмерджентные) свойства, информация о которых может оказаться весьма полезной при принятии решений.

Свойство системности исследуемой совокупности объектов наглядно проявляется при построении типологий, являющихся, как известно, одним из способов описания систем. Рассмотрение общей (генеральной) совокупности классифицируемых объектов в качестве некоторой системы или ее части неявно следует из того, что все участвующие объекты должны быть описаны одним и тем же набором признаков, – особенно это касается алгоритмов классификации, использующих геометрический подход (см., например, [6, с. 148–154]). Признаки при этом имплицитно полагаются существенными, т.е. такими, что каждый из них, взятый в отдельности, необходим, а все вместе они достаточны, чтобы с их помощью

можно было отличить данный объект от всех остальных по той его стороне, познание которой выдвигается как основная задача исследования. В случае если объекты обладают различными признаками, в рассмотрение вводятся так называемые индикаторные признаки, позволяющие путем дихотомии наличия свойств привести описание объектов к единому набору признаков. В этом уже можно усматривать проявление некоторого общего взгляда на совокупность объектов как множества, объединенного этим свойством.

Проблема формирования ципфовых распределений до настоящего времени не имеет однозначного решения. Существуют различные подходы, позволяющие при тех или иных предположениях получать ранговые распределения, соответствующие (1). В литературе известно несколько вариантов объяснения (так называемых «выводов») этого соотношения.

Значительная часть выводов гиперболических распределений была получена путем предельного перехода. Поэтому кривую, соответствующую соотношению (1) иногда рассматривают как одну из кривых семейства Пирсона (X или IV типов), которые получаются в результате предельного перехода из гипергеометрического распределения [5].

Иногда распределение Парето рассматривают как вырожденный случай бета-распределения. Выводы рангового распределения часто основываются на подходах, аналогичных тем, которые встречаются в термодинамике и статистической физике при описании равновесного распределения молекул в газе. Отметим, что такого рода подходы, на наш взгляд, плохо соотносятся с понятием системности объектов, не учитывают принципиальной конечности числа элементов системы и вступают в противоречие с наличием системных связей между отдельными элементами. Проблематичным выглядит и выполнение статистических предпосылок, которые при этом используются. Однако некоторые важные особенности функционирования сложных систем такие подходы довольно хорошо отражают. В качестве примера рассмотрим формирование

распределения случайной величины, которая описывает процесс роста интенсивности некоторого источника, время существования которого является некоторой случайной величиной.

Пусть имеется некоторый источник, порождающий объекты, причем интенсивность λ этого источника пропорциональна уже достигнутому уровню значения величины X. Как известно, такое предположение математически можно представить дифференциальным уравнением $\frac{dX}{dt} = \lambda X$. Если задано значение величины $X_0 = X(t_0)$ в начальный момент времени t_0, то, интегрируя, получим соотношение $X(t) = X_0 e^{\lambda t}$.

Если предположить, что время существования и работы источника является случайной величиной, имеющей экспоненциальное распределение с параметром μ, то плотность распределения времени жизни источника будет определяться формулой $p(t) = \mu e^{-\mu t}$. Тогда, для того чтобы найти $f(x)$ – плотность распределения случайной величины X, выразим значение t из соотношения для $X(t)$, равное $t = \frac{1}{\lambda}\ln(\frac{X_0}{X})$ и подставим в выражение $f(x) = p(t(x))t'_x$.Окончательно получим

$$f(x) = \mu e^{-\mu\frac{1}{\lambda}\ln(\frac{X}{X_0})}\frac{1}{\lambda}\left(\frac{X_0}{X}\right)\frac{1}{X_0} = \frac{\mu}{\lambda}\left(\frac{X}{X_0}\right)^{-\frac{\mu}{\lambda}}\left(\frac{X_0}{X}\right)\frac{1}{X_0},$$

или, если обозначить $\alpha = \frac{\mu}{\lambda}$, то получим хорошо известное распределение Парето

$$f(x) = \left(\frac{\alpha}{X_0}\right)\left(\frac{X_0}{X}\right)^{1+\alpha},$$

которое при $0 < \alpha < 1$ можно рассматривать как непрерывный аналог ципфового распределения.

Рассмотренный подход к механизму формирования гиперболического распределения наглядно демонстрирует, что он не является ни чисто детерминистским, ни чисто стохастическим, а представляет собою объединение этих двух противоположных тенденций. Причем за целостность и замкнутость системы отвечает детерминистская составляющая закона, а стохастическая составляющая как бы накладывается на детерминистскую. Это свойство гиперболических распределений последнее время привлекает особое внимание при анализе разного рода нелинейных процессов.

В качестве другого примера рассмотрим механизм формирования ципфового распределения, полученный на основе принципа максимальной диссиметрии, или – минимума симметрии [3]. Название принципа связано с тем, что в качестве показателя упорядоченности совокупности обычно берется энтропия распределения этой совокупности, которая, как известно, достигает своего максимального значения в том случае, когда все состояния совокупности равновозможные, т.е. распределение состояний является максимально симметричным. И наоборот, отклонение от симметричности распределения связывают с наличием некоторой структуры, упорядоченности совокупности. При этом неявно предполагается, что уменьшение симметрии, или – что то же самое, – увеличение диссимметрии, свидетельствует о большей структурированности рассматриваемой совокупности. Подобные идеи широко используются в синергетике, где полагается, что хаос обладает максимальной симметрией.

Постулируя принцип минимума симметрии в качестве общесистемного принципа, можно, как будет показано ниже, получить некоторые следствия, которые представляют собою количественные характеристики системности. Таким образом, появляется возможность на практике проверить выполнение этого постулата на эмпирическом материале.

Предположим, что имеется совокупность, состоящая из N объектов, для которой построено классификационное разбиение,

состоящее их K непустых классов численностями $n_1, n_2, \ldots, n_K$, причем $\sum_{i=1}^{K} n_i = N$. Тогда в качестве меры симметрии возьмем число преобразований множества в себя, которое сохраняет данное разбиение. Для класса, содержащего n_i объектов, число таких преобразований будет $n_i!$, а для всего разбиения в целом число преобразований будет равно произведению $n_1!\, n_2! \ldots n_K!$.

Если определять классификационное разбиение, исходя из минимума только этой меры симметрии, то будет получен тривиальный результат: $n_i = 1$ для всех $i = 1, 2, \ldots, N$. Поэтому одновременно с разбиением в рассмотрение вводится сопряженное ему разбиение той же совокупности $m_1, m_2, \ldots, m_L$ ($\sum_{i=1}^{L} m_i = N$) так называемое коразбиение, которое определяется следующим образом:

1) классы n_i и m_j в пересечении имеют не более одного элемента;

2) любое укрупнение классов коразбиения этим свойством не обладает;

3) если пересечение классов n_i и m_j не пусто, то m_j пересекается с любым классом разбиения, число элементов которого не меньше, чем n_i.

Пример построения коразбиения можно представить, если в имеющемся ранговом распределении провести вертикальные полосы, которые выделят классы исходного разбиения, а горизонтальные – классы коразбиения. Как следует из определения, для каждого разбиения существует множество коразбиений, которые получаются за счет перестановки элементов из одного класса.

Очевидно, что мера симметрии для коразбиения будет иметь такой же вид: $m_1!\, m_2! \ldots m_l!$. Тогда общее число преобразований, не изменяющих исходное разбиение, будет равняться $S = (n_1!\, n_2! \ldots n_K!)^{\alpha}\, (m_1!\, m_2! \ldots m_l!)^{\beta}$, где постоянные величины α и β введены как весовые коэффициенты мер симметрии.

Наша задача будет теперь заключаться в нахождении такого разбиения, численности значений классов n_1, n_2, …, n_K которых доставляют функции S минимальное значение при условии, что $\sum_{i=1}^{K} n_i = N$.

Для минимизации величины S удобно перейти к ее логарифму, достигающему минимального значения там же, где и сама S. Кроме того, используя асимптотическую формулу Стирлинга $\ln(n!) \approx n(\ln n - 1)$, получим

$$\ln S \approx \alpha \sum_{i=1}^{K} n_i \ln n_i - \alpha \sum_{i=1}^{K} n_i + \beta \sum_{i=1}^{L} m_j \ln m_j - \beta \sum_{i=1}^{L} m_j.$$

Легко заметить, что минимизируемая функция S представляет собою с точностью до постоянных слагаемых взвешенную сумму энтропий исходного и сопряженного к нему разбиения, т.е. действительно характеризует степень упорядоченности множества, на котором задано разбиение.

Для дальнейшего решения оптимизационной задачи аппроксимируем значения численностей классов функцией $y = y(x)$, принимающей в целочисленных точках значения, совпадающие с n_i, а за m_j примем абсциссы точек с целочисленными ординатами той же функции. В результате этого получим возможность записать следующие приближенные равенства

$$\sum_{i=1}^{K} n_i \ln n_i \approx \int_1^a y \ln y dx,$$

$$\sum_{j=1}^{L} m_j \ln m_j \approx \int_b^c x \ln x dy,$$

$$\sum_{i=1}^{K} n_i = \sum_{j=1}^{L} m_j \approx \int_1^a y dx = N.$$

С учетом полученных равенств целевая функция примет вид

$$\ln S \approx \alpha \int_1^a y \ln y dx + \beta \int_c^d x \ln x dy - (\alpha + \beta) \int_1^a y dx.$$

Сделаем замену переменной во втором интеграле, учитывая что $y = y(x)$, и, следовательно, $dy = y'dx$, а $y(1) = d$, $y(a) = c$. После этих преобразований задача сводится к нахождению функции $y = y(x)$, минимизирующей следующее выражение

$$\ln S \approx \int_1^a (\alpha y \ln y - \beta x \ln x\, y' - (\alpha + \beta) y) dx \to \min$$

при заданном ограничении $-\int_1^a y dx = N$. Для дальнейшего решения полученной задачи используем метод неопределенных множителей. Функция Лагранжа после преобразований будет иметь вид:

$$\int_1^a (\alpha y \ln y - \beta x \ln x\, y' + \lambda y) dx \to \min.$$

Для решения полученной задачи используем вариационный подход и уравнение Эйлера. Полагая $F(x, y, y') = \alpha y \ln y - \beta x \ln x\, y' + \lambda y$, найдем

$$\frac{\partial F}{\partial y} = \alpha \ln y + \alpha + \lambda,$$

$$\frac{\partial F}{\partial y'} = -\beta x \ln x,$$

$$\frac{d}{dx}\left(\frac{\partial F}{\partial y'}\right) = -\beta \ln x - \beta.$$

Подставляя найденные значения в уравнение Эйлера $\frac{\partial F}{\partial y} - \frac{d}{dx}\left(\frac{\partial F}{\partial y'}\right) = 0$, окончательно получим $\alpha \ln y + \beta \ln x + \alpha + \beta + \lambda = 0$. Разрешая это уравнение относительно искомой функции, найдем $y = Cx^{-\gamma}$, где $\gamma = \frac{\beta}{\alpha}$, а $C = y(1)$.

Полученная зависимость устанавливает приближенную связь ранга класса x с y – объемом этого класса. Нетрудно заметить, что эта зависимость соответствует закону Ципфа.

Отметим, что в данном подходе связные ранги не рассматривались. Это приводит к тому, что полученное соотношение хорошо описывает распределение объектов в области малых рангов и гораздо хуже в области больших рангов, где часто фиксируются классы, имеющие одинаковую численность. Однако это в целом не снижает значения полученных в ходе рассуждений выводов. В частности, из приведенного рассуждения следует, что с точки зрения предложенного принципа максимальной симметрией в случае дискретных величин будет обладать равномерное распределение.

Еще одно соображение в пользу выполнения рангового распределения в форме (1) приводится нами в [7], где задача построения классификации сводится к нахождению решения стационарного уравнения Шредингера.

Общее решение введенного в рассмотрение уравнения может быть представлено в виде ряда $\Psi = \sum_n C_n \Psi_n$, и, принимая во внимание условие ортонормированности собственных функций задачи, можно видеть, что $(\Psi, \Psi^*) = \sum_n |C_n|^2$. Это позволяет интерпретировать коэффициенты $|C_n|^2$ как интенсивности классов, т.е. величины, пропорциональные их численности. Требование конечности функции Ψ приводит к тому, что ряд, составленный из коэффициентов, должен быть сходящимся, и, следовательно, $|C_n|^2 \to 0$. Если предположить, что коэффициенты этого ряда в простейшем случае обратно пропорциональны порядковым номерам классов, т.е. выполняется соотношение $|C_n|^2 = \frac{C}{n^P}$, где C, p – некоторые константы. Тогда для сходимости ряда, члены которого задаются таким соотношением, очевидно, должно выполняться требование $p > 1$, что

автоматически приводит к тому, что ранговое распределение численности классов должно носить гиперболический характер.

По-видимому, можно еще найти и другие соображения в пользу этой закономерности, эмпирически обнаруженной во многих областях научной деятельности. Но в любом случае этот показатель структурированности множества не должен оставаться без внимания при исследовании сложных объектов, особенно в рамках системного подхода. Там, где фиксируется выполнение соотношения (1) на эмпирическом уровне, имеет смысл искать проявление и других системных закономерностей.

Пример построения типологии увольняющихся методами многомерной классификации

В качестве примера построения типологии в социально-экономических исследованиях предлагается рассмотреть результаты, полученные автором совместно с В.М. Гаськовым в Институте социологических исследований в 1977–1978 гг. Работа была выполнена на материалах исследования, проведенного в Бурятской АССР.

Это исследование было связано с изучением структуры увольняющихся с промышленных предприятий с целью выявления классов мигрантов, обладающих различными типами трудовой мобильности. Под типом трудовой мобильности мы понимали совокупность реального и предполагаемого движения трудоспособного населения между регионами страны, которое является существенным для данной совокупности респондентов. Были выделены два основных типа поведения, соответствующих моменту исследования: территориальная стабильность (увольняющиеся переходят на другие предприятия в пределах населенных пунктов) и территориальная мобильность (увольняющиеся выезжают в другие населенные пункты). Один и тот же тип поведения может быть свойствен совокупности классов, каждый из которых объединяет группу респондентов со сходными личностными характеристиками.

В качестве признаков, образующих пространство, в котором проводилась классификация, были выбраны такие переменные как возраст, образование, длительность проживания респондента в данном населенном пункте, а также уровень жизни в районах рождения, получения образования, выбытия и предполагаемого вселения. Ряд других показателей, такие, например, как стаж работы на предприятии, использовались только на стадии интерпретации классификационных разбиений.

Входной информацией для построения классификации послужили данные анкетного опроса увольняющихся с промышленных предприятий г. Улан-Уде, проведенного в 1977 г. Из всего контингента увольняющихся была образована случайная выборка, составившая 5% объема генеральной совокупности.

Весь массив исходных данных был предварительно разбит на две группы. Первая группа характеризовалась территориальной стабильностью. В ее состав вошли анкеты тех, кто переходил на другие предприятия в пределах города. Вторую группу составили анкеты респондентов, которые собирались выехать из города Улан-Уде и указали район нового вселения. Эта группа характеризовалась территориальной мобильностью. Анкеты сформированных таким образом групп обрабатывались первоначально отдельно.

При построении типологии была использована совокупность алгоритмов многомерной классификации в соответствии с предложенной нами методикой [8, с. 39–41]. Так, на первом этапе применялся алгоритм итеративного метода классификации «Форэль». Значение управляющего параметра, радиуса гиперсферы, выбиралось исходя из содержательных оценок полученных разбиений. В нашем случае он выбирался таким, чтобы при нем выделялись качественно различные ядра классов.

На следующем этапе был применен иерархический агломеративный алгоритм, работающий по принципу «ближайшего соседа». Для удобства сравнения результатов, полученных разными

алгоритмами, была использована модификация этого метода, позволяющая получать одно разбиение на заданном уровне связности.

Третий этап исследования массива многомерных данных заключался в применении алгоритма, использующего понятие нечетких множеств. В качестве функции принадлежности бралось число точек в гиперсфере выбранного радиуса. Результаты работы этого алгоритма использовались для уточнения плотности распределения и оценки мод отдельных классов.

Для получения связных на выбранном уровне классов на четвертом этапе применялся алгоритм модального анализа, использующий градиентную процедуру. Этот алгоритм, как и предыдущий, для построения классификационного разбиения существенно использует функцию принадлежности, являющуюся в нашем случае оценкой плотности распределения. Результаты работы этого алгоритма и составили основу результирующего классификационного разбиения. Применение методов многомерной классификации позволило выделить в составе первой группы 9 классов, а в составе второй – 11 классов различных объемов.

Наиболее многочисленным из всех классов является класс № 1, который полностью состоит из местных уроженцев. Средний возраст представителей этого класса находится в интервале 23-25 лет. Представители этого класса активно меняют места работы, о чем свидетельствует отсутствие значимой корреляции между возрастом и стажем работы. В целом, среди местных уроженцев, перераспределяющихся среди предприятий города, уровень образования 8-10 классов присущ только этой группе. Отсутствие класса увольняющихся аналогичной численности в возрастах свыше 25 лет позволяет предположить, что по мере взросления эта группа рассеивается, частично превращаясь в стабильные кадры предприятий и частично переходя в другие классы, для которых характерно миграционное поведение.

Класс № 2 составили местные уроженцы, выезжающие в крупные города с более высоким уровнем жизни, в то время как поведение местных уроженцев, составляющих класс № 3, – миграция в

малые города и поселки Бурятии. По-видимому, здесь проявляется действие каких-то дополнительных, не учтенных нами факторов. Следует отметить, что среди классов, составленных из местных уроженцев, перераспределяющихся в пределах города, нет типичных групп в интервале 19-22 года. Для увольняющейся с предприятий молодежи с восьмилетним и средним образованием в этом возрастном интервале характерна категоричность в принятии решений – они выезжают из города, не попытавшись устроиться на другое предприятие. Среди мотивов выезда в этом классе наибольший удельный вес имел «выезд на учебу», поэтому и основное направление миграции – крупные города Сибири и Дальнего Востока.

Для увольняющихся местных уроженцев с относительно низким уровнем образования типичным является миграционное поведение в средних возрастах 30-33 лет (класс № 6). Для этого класса характерна весьма высокая интенсивность текучести, о чем свидетельствует наличие значимой корреляционной связи между длительностью проживания в городе и стажем работы. Для остальных классов с таким уровнем образования типичным является территориальная стабильность. Для местных уроженцев, получивших среднее специальное и высшее образование, миграционное поведение является наиболее вероятным в возрастном интервале 26-30 лет (класс № 4), для остальных характерно перераспределение между предприятиями в пределах города (классы № 5 и № 6).

Среди мигрантов, проявляющих территориальную стабильность, были получены классы № 7, 8, 9. Класс № 7 составляют мигранты из сел и городов с более низким уровнем жизни, чем их настоящее место пребывания. Для представителей этого класса характерно то, что большинство из них уже меняли место жительства до приезда в Бурятию. В класс № 8 вошли мигранты, приехавшие в Улан-Уде из других крупных городов. Причем, если представители класса № 7 к данному моменту уже успели поменять несколько раз место жительства, то в отличие от них для

представителей класса № 8 эта смена места жительства является первой. Класс № 9 представлен сельскими мигрантами, приехавшими из сел Восточной Сибири в г. Улан-Уде в возрасте 20-30 лет и проявляющие хорошую приживаемость. Относительно высокие показатели их проживания в городе последнего вселения дают основания предполагать, что представители этого класса будут проявлять территориальную стабильность и в перспективе. Очевидно, что одним из важных факторов, снижающих миграционную подвижность, является низкий уровень их образования.

Среди самостоятельных мигрантов, увольняющихся с предприятий города и выезжающих за его пределы, можно выделить пять классов. Мобильное поведение индивидов, образовавших класс № 7, принято называть «возвратной миграцией». В ней участвуют уроженцы сел Восточной Сибири, для которых Улан-Уде был первым крупным городом их вселения. Отсутствие необходимых условий жизни, важнейшим из которых является наличие жилья, побудило их вернуться в места рождения. Этот тип поведения получил широкое распространение. Возвратная миграция характерна также и для увольняющихся, вошедших в класс № 10, который составлен из специалистов, получивших среднее специальное образование и выезжающих, в основном, в места рождения – села Восточной Сибири. Специфическим поведением отличаются мигранты, выходцы из сел, образовавшие класс № 8, которые, не успев прижиться в городе, выезжают на учебу в крупные города Сибири. Это явление получило название «учебной миграции». Класс № 9 характеризует миграционное поведение уроженцев сел России, которые, однако, приехали в Улан-Уде уже из других крупных городов, имеющих более высокий уровень жизни. Снижение уровня жизни обусловило их решение выехать за пределы Улан-Уде в направлении других крупных городов. Классы № 7-10 составлены главным образом из новоселов, которые, не прижившись в городе по разным причинам, выезжают за его пределы. Малочисленным является класс № 11, в который вошли работники в возрасте 30-40 лет со

средним и специальным образованием. Для них характерно принятие решения о выезде после длительного проживания в городе.

Приведенные описания классов, полученных в результате применения методов кластерного анализа, показывают, что каждый класс обладает характерными, присущими только ему особенностями. Выявленная структура групп увольняющихся обнаруживает необходимость дифференцированного подхода при изучении причин и мотивов в проведении миграционной политики. Более подробное описание результатов можно найти в публикациях (см., например [9]).

Вторичный анализ результатов классификации: проверка выполнения ципфового распределения на совокупности объектов

Построенная типология увольняющихся не снимает вопроса о целостности всей совокупности: можно ли рассматривать эту совокупность как некоторую систему со всеми вытекающими отсюда последствиями, или рассмотренная совокупность представляет собою отдельные, независимо функционирующие группы. Если рассмотренная совокупность представляет собою часть какой-то более широкой системы, тогда при принятии управленческих решений следует учитывать влияние и других факторов, не учтенных этим исследованием. Для ответов на поставленные вопросы перейдем к исследованию ранговых распределений полученных классификационных разбиений.

Для построения ранговых распределений нами была использована система Statistica. Ранговые распределения, соответствующие результатам классификации, представлены на рис. 1 и 2. Особо отметим, при построении ранговых распределений были также учтены и не отнесенные ни в какие классы объекты, обладающие уникальными характеристиками и соответствовавшие нетипичному миграционному поведению. Кроме этого, было рассмотрено и ранговое распределение, построенное по всему массиву

исходных данных. Отметим, что классы, соответствующие различным типам мобильности и входящие в различные результирующие разбиения, не могут быть объединены в силу того, что интервалы изменения признаков для разных классов различны.

Проверка выполнения закона Ципфа на рассматриваемом множестве не является такой уж простой задачей, как может показаться на первый взгляд. Эта проблема широко обсуждалась в литературе, но до сих пор остается актуальной. Самый «простой» и, как может показаться, понятный способ сводится к построению регрессионной зависимости в дважды логарифмических координатах. Для этого логарифмируют выражение (1) и получают линейное относительно всех переменных уравнение

$$\ln(n_i) = \ln C + \gamma \ln(i), \qquad (2)$$

где $\gamma = -(1 + \alpha)$.

Отметим, что основание логарифма при этом не имеет никакого значения для определения величины параметра γ. Затем стандартным способом производится оценка параметров регрессионного уравнения.

Однако, как правило, такой подход не дает желаемого результата: визуально фиксируется значительное отклонение от гиперболического ципфового распределения, а уравнение регрессии свидетельствует о том, что соответствие вполне приемлемое, – уравнение регрессии значимо на стандартном уровне, коэффициент значимо превышает единицу, коэффициент детерминации весьма близок к единице. Сложившаяся ситуация проясняется после проверки выполнения условий теоремы Гаусса-Маркова, которая показывает, что почти все условия оказываются нарушенными. Так, и это следует отметить в первую очередь, ранжирование классов по их численности вносит существенную зависимость в сами наблюдения. Речи о независимости ошибок (отклонений) не может быть. В этом легко убедиться, исследуя остатки с помощью критерия Дарбина-Уотсона. Автокорреляция остатков, вносимая ранжировкой объектов, почти всегда значима на самом строгом уровне.

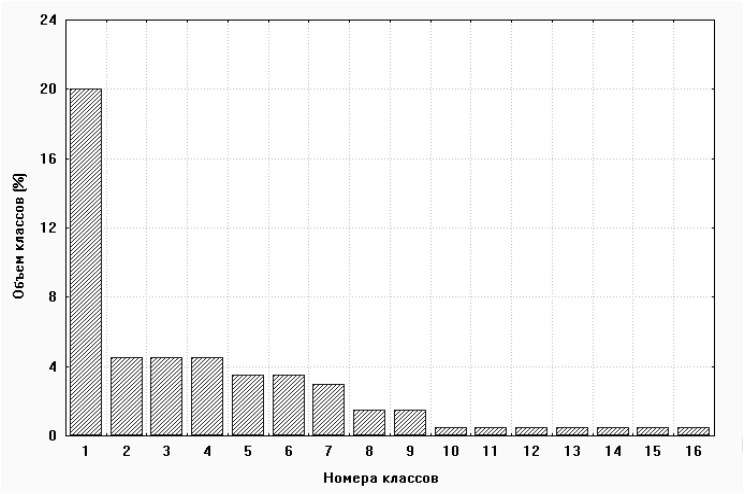

Рис. 1. Ранговое распределение классификационного разбиения № 1

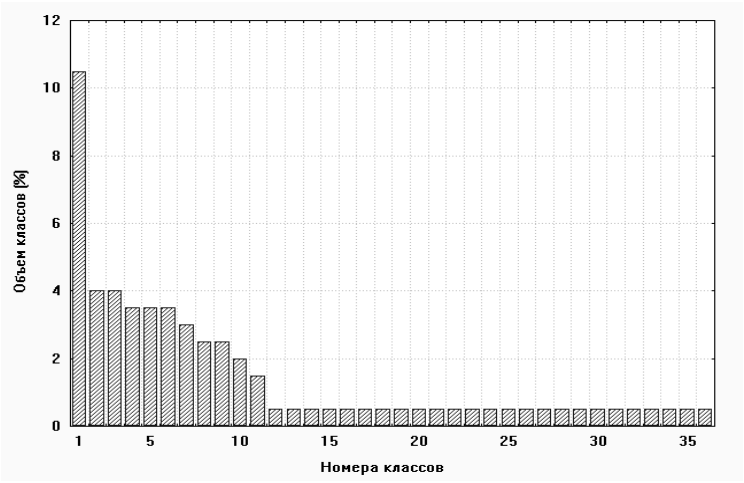

Рис. 2. Ранговое распределение классификационного разбиения № 2

Нарушения гиперболичности рангового распределения часто хорошо заметны визуально, но с большим трудом фиксируются на уровне количественных показателей. Так, если проанализировать графическое представление результатов, полученных при исследовании трудовой мобильности (см. рис. 1 и рис. 2), то легко убедиться визуально (на качественном уровне), что каждое из ранговых распределений, построенных отдельно по совокупностям увольняющихся, обладающих различными типами трудовой мобильности, заметно отличается от ципфовых. Однако анализ регрессионных уравнений, построенных в дважды логарифмических координатах, не подтверждает этого. Если обозначить через Y значения логарифмов численностей классов, а через X – значения логарифмов соответствующих рангов, то для данных первой группы, проявляющей территориальную стабильность, построенное уравнение регрессии будет иметь вид $Y = 3{,}116 - 1{,}395X$ ($R^2 = 0{,}88$), а для данных второй группы – $Y = 2{,}641 - 1{,}053X$ ($R^2 = 0{,}83$). Оба приведенных регрессионных уравнения значимы на стандартном 5%-ном уровне. Уравнение регрессии, построенное по объединенным данным, также значимо и имеет вид $Y = 3{,}348 - 1{,}107X$ ($R^2 = 0{,}87$). Как видим, сравнение результатов регрессионного анализа не позволяет ответить на вопрос о наличии распределения Ципфа и, следовательно, о целостности системы. Все это является следствием нарушения условий теоремы Гаусса-Маркова и в данном случае приводит к некорректности традиционного анализа.

Подход, предлагаемый нами, основан на проверке утверждения, что полученное распределение действительно является ципфовым, т.е. отличается от случайно сформированного равномерного рангового распределения. Сравнение с равномерным распределением, которое выбирается в данном случае в качестве распределения нулевой гипотезы, связано с тем, что, как известно, именно это распределение доставляет максимум энтропии, – функции, характеризующей степень упорядоченности рассматриваемого множества.

Для проверки этого предположения был использован метод статистических испытаний, который заключался в том, что с помощью датчика случайных чисел строилось разбиение множества на заранее заданное число классов, численности которых распределены в соответствии с равномерным распределением. Объемы классов, выраженные в процентах, ранжировались, а затем по полученному таким образом ранговому распределению производилась оценка параметра γ. Эта процедура повторялась определенное, достаточно большое число раз, например 1000. Значения параметров в дальнейшем рассматривались как случайные величины. На основании полученной в эксперименте гистограммы можно подобрать функцию распределения или плотность. С помощью уровня значимости оценивался интервал, в котором находился интересующий нас параметр рангового распределения с заданной вероятностью. Или, другими словами, определяется критическая область для проверяемой гипотезы.

Дальнейшая проверка гипотезы о том, что построенное ранговое распределение действительно отлично от случайного, производится стандартным способом. Если рассчитанные на основании эмпирических данных параметры ранговых распределений соответствуют значениям, полученным имитационным способом, т.е. попадают в область принятия нулевой гипотезы, то это означает, что нет достаточных причин для того, чтобы отказаться от нулевой гипотезы. Другими словами, отличие эмпирического рангового распределения от равномерного не является значимым, а носит случайный характер.

И наоборот, если эмпирические значения параметров ранговых распределений попадают в критическую область, то это свидетельствует о том, что рассматриваемое ранговое распределение вряд ли можно признать случайным. Иначе говоря, полученное ранговое распределение следует признать ципфовым с заданным уровнем надежности. Последнее влечет за собой ряд общеизвестных следствий, указанных выше, основным из которых является утверждение о целостности системы.

Для реализации этого подхода нами была составлена программа в среде MathCAD, позволяющая для заданного числа классов строить ранговые распределения, в которых численности классов подчинялись равномерному распределению. Для полученных в результате эксперимента данных в дважды логарифмическом масштабе строилось уравнение парной регрессии в соответствии с (2). Рассчитанные по этим данным значения коэффициентов регрессии рассматривались как значения случайной величины. На основании полученных таким образом данных в дальнейшем строились доверительные интервалы, позволяющие с заданным уровнем надежности оценить значение параметра рангового распределения.

В ходе эксперимента было установлено, что при сравнительно небольших объемах повторения (в пределах от 100 до 200) полученное распределение хорошо аппроксимируется нормальным распределением. Характерный случай представлен на рис. 3.

В этом случае границы области принятия нулевой гипотезы определялись на основе плотности нормального распределения с соответствующими параметрами, оценки которых были получены в ходе эксперимента. Результаты экспериментов представлены в табл. 1.

В том случае, когда число испытаний превышало 1000, наблюдалось отличие полученного распределения от нормального. Типичный случай можно видеть на рис. 4, где представлено распределение параметра, полученное в ходе 100000 статистических испытаний.

В этом случае построение критической области проводилось посредством определения соответствующих процентилей эмпирического распределения. Один из результатов такого эксперимента представлен в табл. 1.

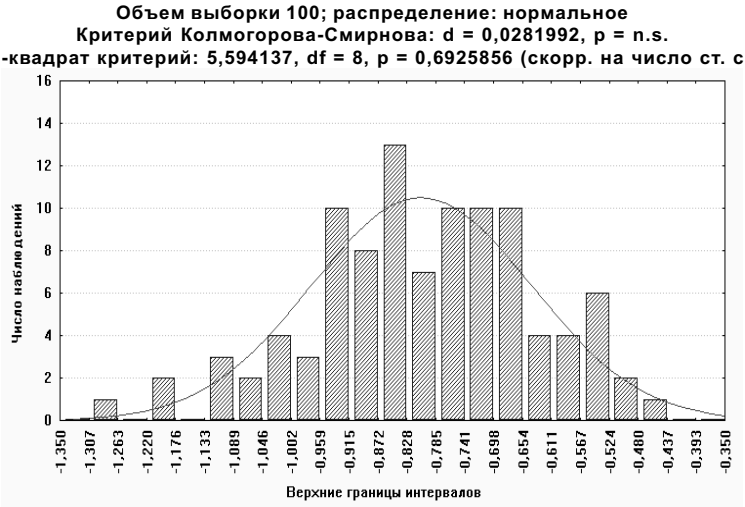

Рис. 3. Гистограмма коэффициента рангового распределения 30 классов в выборке объемом 100 единиц

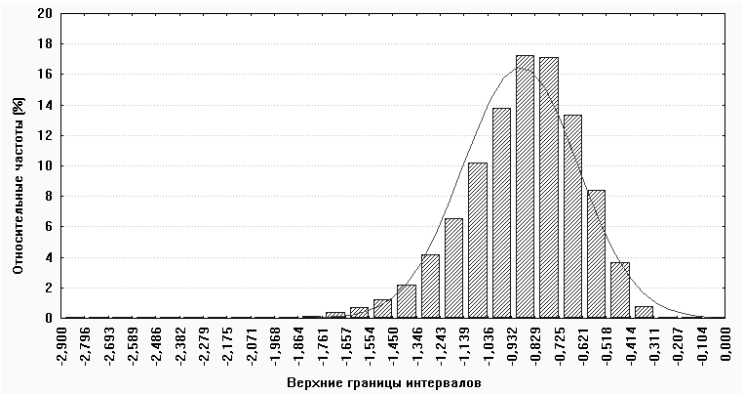

Рис. 4. Гистограмма коэффициента рангового распределения 16 классов в выборке объемом 100000 единиц

Таблица 1

РЕЗУЛЬТАТЫ ОЦЕНИВАНИЯ ЗНАЧЕНИЯ ПАРАМЕТРА РАНГОВОГО РАСПРЕДЕЛЕНИЯ (НОРМАЛЬНАЯ АППРОКСИМАЦИЯ)

№	Численности классов	Расчетное значение параметра	Нормальная аппроксимация (n = 100)			
			Среднее значение	Стандартное отклонение	Левая граница	Правая граница
1	16	–1,395	–0,889	0,267	–1,412	–0,367
2	30	–1,053	–0,845	0,167	–1,172	–0,518
3	46	–1,107	–0,748	0,112	–0,968	–0,529

Таблица 2

РЕЗУЛЬТАТЫ ОЦЕНИВАНИЯ ЗНАЧЕНИЯ ПАРАМЕТРА РАНГОВОГО РАСПРЕДЕЛЕНИЯ (МЕТОД СТАТИСТИЧЕСКИХ ИСПЫТАНИЙ)

№	Численности классов	Расчетное значение параметра	Машинная (числовая) аппроксимация (n = 100000)			
			Среднее значение	Стандартное отклонение	Левая граница	Правая граница
1	16	–1,395	–0,893	0,247	–1,441	–0,481
2	30	–1,053	–0,844	0,167	–1,241	–0,569
3	46	–1,107	–0,813	0,124	–1,036	–0,551

Анализ представленных результатов показывает, что в действительности ранговое распределение, построенное в отдельности по данным каждой группы, нельзя признать ципфовым, так как они незначимо отличаются от случайных распределений, численности классов которых имеют равномерное распределение. Этот вывод следует из того, что значения соответствующих параметров распределения не выходят за границы области принятия нулевой гипотезы. В табл. 1 можно видеть, что значение –1,395 принадлежит интервалу (–1,412; –0,367), а значение –1,053 – интервалу (–1,172; –0,518). В табл. 2 построенные интервалы (–1,441; –0,481) и (–1,241; –0,569) также включают в себя эти значения. Практически это означает, что эти совокупности мигрантов вряд ли стоит рассматривать как целостные замкнутые системы. Другими словами, проведение миграционной политики только в отношении одной из этих групп без учета интересов другой группы не является целесообразным, а рассмотрение в отдельности представителей различных групп миграции не позволяет составить представление о всей совокупности как едином целом.

С другой стороны, объединяя данные в одну совокупность, можно получить системный объект, для которого характерным является целостность (системность). Вывод о выполнении закона Ципфа на совместном ранговом распределении можно сделать из того, что значение расчетного коэффициента, равное –1,107, не принадлежит ни интервалу (–0,968; –0,529) табл. 1, ни интервалу (–1,036; –0,551) табл. 2. Значит, следует признать, что ранговое распределение, построенное по всей исследуемой совокупности, является ципфовым. Это означает, что только вся исследуемая совокупность может быть представлена как система, а отдельные ее части этим свойством не обладают.

Таким образом, как было установлено при исследовании ранговых распределений, при совместном рассмотрении всей совокупности к ней следует подходить как к некоторой сложившейся целостной системе, функционирование которой в значительной степени

определяется местными условиями и взаимодействием отдельных ее составляющих. Полученные выводы были использованы в практической деятельности по проведению демографической и миграционной политики в районах Сибири и Дальнего Востока.

***

Принятие решений по управлению системным объектом должно обязательно учитывать особенности его функционирования. Так, например, одним из следствий выполнения ципфового распределения является наличие большого числа малочисленных классов. Это положение следует иметь в виду, если придерживаться принципов системного подхода. Игнорирование этого следствия приводит, как правило, к нарушению целостности системы и мешает нормальному ее функционированию. Достаточно, например, вспомнить окончательные результаты опыта по укрупнению сельскохозяйственных предприятий в советское время. Или, с этой точки зрения, попытаться оценить недавно принятые Государственной думой поправки в федеральный закон «О политических партиях», согласно которым минимальная численность политических партий в России ограничивается снизу 50 тыс. членов.

ЛИТЕРАТУРА

1. *Буховец А.Г.* Критерий системности социально-экономических объектов // Математические методы в социологических исследованиях / А.Г. Буховец, А.С. Соловьев. М.: ИСИ АН СССР, 1984. С. 28–36.

2. *Буховец А.Г.* Кластерный анализ как метод решения классификационной задачи // Вестник факультета прикладной математики и механики. Воронеж: ВГУ, 2000. С. 248–253. Вып. 2.

3. *Шрейдер Ю.А.* Системы и модели / Ю.А. Шрейдер, А.А. Шаров. М.: Радио и связь, 1982.

4. *Буховец А.Г.* Использование ранговых распределений при интерпретации результатов кластерного анализа // Методы социологических исследований. 3-я Всесоюзная конференция / А.Г. Буховец, А.С. Соловьев. М., 1989.

5. *Яблонский А.И.* Математические модели в исследовании науки. М.: Наука, 1986.

6. Типология и классификация в социологических исследованиях. М.: Наука, 1982.

7. *Буховец А.Г.* Об одном подходе к задаче классификации // Социология: методология, методы, математические модели. 2004. № 18. С. 82–105.

8. *Буховец А.Г.* Последовательное применение алгоритмов многомерной классификации // Многомерный анализ социологических данных (методические указания, алгоритмы и описания программ). М.: ИСИ АН СССР, 1981. С. 24–73.

9. *Буховец А.Г.* Изучение трудовой мобильности методами многомерной классификации // Проблемы воспроизводства и миграции населения / А.Г. Буховец, В.М. Гаськов. М.: ИСИ АН СССР, 1981. С. 215–228.