УДК 519.86

СИСТЕМНЫЙ ПОДХОД И РАНГОВЫЕ РАСПРЕДЕЛЕНИЯ В ЗАДАЧАХ КЛАССИФИКАЦИИ

А. Г. Буховец

Воронежский государственный аграрный университет им. К. Д. Глинки

Сейчас уже не вызывает споров методология системного подхода, окончательно сформировавшаяся в семидесятых годах прошлого века. Несложно указать список литературы, включающий множество статей и даже учебников, в которых сформулированы разработанные основные положения системного анализа в самом общем виде. Однако, в практических задачах использование системного анализа часто носит декларативный характер, а результаты оцениваются на качественном уровне.

Задачу построения классификации нередко рассматривают как задачу построения 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.

Алгоритмы классификации как таковые не связаны с системными представлениями совокупности объектов, но критерии оценки результатов применения алгоритмов уже должны учитывать системные представления обо всей совокупности. Задача нашей работы заключается в том, чтобы показать, что в рамках системного подхода имеется возможность дать объяснение некоторым закономерностям, связанным с количественными оценками полученных классификационных разбиений, которые выражаются через параметры ранговых распределений.

1. ПРИНЦИПЫ СИСТЕМНОГО ПОДХОДА В КЛАССИФИКАЦИОННЫХ ЗАДАЧАХ

Основные принципы системного подхода удалось сформулировать почти одновременно в нескольких дисциплинах. Одними из первых ввели в рассмотрение понятие системы и системного подхода биологи, которые придерживались взгляда на живой организм как интегрированное целое. Идеи, выдвинутые биологами, способствовали появлению нового способа анализа, — системного анализа, опирающегося на связность и взаимоотношения отдельных элементов системы. Согласно этому подходу существенными свойствами организма или живой системы, являются свойства целого, которыми не обладает ни одна из его частей. Такие свойства возникают из взаимодействий и связей между частями. Эти свойства нарушаются, когда система рассекается на отдельные изолированные элементы.

Убеждения, что в любой сложной системе поведение целого (всей системы) может быть полностью понято на основе свойств его частей, было краеугольным камнем в механистической парадигме. В системном подходе приоритет имеет целостность системы, а уж затем рассматриваются ее составляющие элементы. Шагом вперед было сформулированное утверждение, что 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). 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. Аналогичным является и такое утверждение: «Назначение всякой классификации ... заключается, прежде всего, в том, чтобы быть средством лучшего познания изучаемых объектов, о которых еще не имелось сформировавшихся понятий» (Формальная логика. Учебник для философских факультетов университетов. — М.: 1977, С. 140).

Из всех проблем, возникающих при использовании принципов системного подхода в ходе построения типологизации, мы в своей работе ограничимся в полной мере рассмотрением только одного аспекта, связанного с исследованием получающихся при классификации ранговых разбиений.

Под ранговым распределением понимают зависимость численности, соответствующей данному элементу, от его порядкового номера (ранга) при расположении элементов по убыванию этой численности. В работах социально-экономического характера, так или иначе связанных с типологизацией, 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. В математической форме эта зависимость наиболее просто может быть выражена в следующем виде:

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

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

Приведем некоторые наиболее известные примеры ранговых распределений, представленных в указанной выше форме:

1. В географии — распределение городов по численности населения в некоторых замкнутых регионах (государствах);

2. В биологии — распределение биологических родов по числу видов в них, а также распределение видов по занимаемому ими ареалу;

3. В экономике — распределение населения по уровню доходов;

4. В лингвистике — распределение отдельных слов по частоте их появления в лексически правильном тексте;

5. В наукометрии — распределение научной продуктивности ученых; распределение индекса цитируемости; распределение числа публикаций по некоторой тематике по журналам;

6. В информатике — распределение информационных потерь; популярность веб-сайтов;

7. В политологии — голосование избирателей в случае свободного волеизъявления;

8. В сельском хозяйстве — распределение урожайности некоторых сельскохозяйственных культур; распределение цепного индекса урожайности.

Ссылки на указанные факты приводятся в работах [8—11].

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 Ципф [G. K. Zipf] был одним из первых, кто не только обнаружил выполнение на эмпирическом уровне этого закона, но и предложил объяснение его механизма формирования [10]. 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 простых, 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, такую, например, как фрактальная размерность.

2. О МЕХАНИЗМАХ ФОРМИРОВАНИЯ ЦИПФОВОГО РАСПРЕДЕЛЕНИЯ

В задачах классификации, как правило, не оговариваются свойства исследуемых совокупностей, — будь то генеральных или выборочных. Обычно не указывается, явля-

ется 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 объекты, участвующие в рассмотрении должны быть описаны одним и тем же набором признаков, — особенно это касается алгоритмов классификации, использующих геометрический подход. 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 т.н. индикаторные признаки, позволяющие путем дихотомии наличия свойств привести описание объектов к единому для всех объектов 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 (т.н. «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 [11].

Иногда распределение Парето рассматривают как вырожденный случай бета-распределения. В [9] для вывода рангового распределения были использованы подходы, аналогичные законам термодинамики и статистической физики, применяемые для описания 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 [11].

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, определяется 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 $t = \frac{1}{\lambda}\ln\left(\frac{X_0}{X}\right)$ 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 собой 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 синергетике.

В [9] показано, что известные закономерности ранговых распределений являются следствием некоего общесистемного принципа максимума диссимметрии. Этот чисто математический вывод дает некоторые основания полагать, что закономерности ранговых распределений действительно имеют системный характер. Там, где они проявляются на эмпирическом уровне, имеет смысл искать проявление и других системных закономерностей. Другими словами, такие закономерности являются необходимым условием внешней системы: невыполнение этого условия означает, что рассматривается некоторый конгломерат стихийно отобранных объектов, а выполнение такого условия еще не гарантирует, что рассматриваемая совокупность объектов обязательно является системой со всеми вытекающими из этого определения следствиями.

Еще одно соображение в пользу выполнения рангового распределения в форме (1) приводится нами в [3], где задачи 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 эмпирически обнаруженной закономерности. Но в любом случае этот показатель структурированности множества не должен оставаться без внимания при исследовании сложных объектов, особенно в рамках системного подхода.

3. СОДЕРЖАТЕЛЬНЫЙ ПРИМЕР ПОСТРОЕНИЯ ТИПОЛОГИИ УВОЛЬНЯЮЩИХСЯ МЕТОДАМИ МНОГОМЕРНОЙ КЛАССИФИКАЦИИ

В качестве примера построения типологии в социально-экономических исследова-

ниях предлагается рассмотреть результаты, полученные автором совместно с В. М. Гаськовым в Институте социологических исследований в 1977—1978 гг. Работа была выполнена на материалах Бурятской АССР, для хозяйства которой привлечение и закрепление кадров имело большое практическое значение. Более подробное описание результатов можно найти в публикациях (см., например [4]).

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 свойственен 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 алгоритмом многомерной классификации в соответствии с преложенной нами методикой [1, 2]. 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 использован 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. Окончательные результаты классификации представлены в таблицах 1 и 2 соответственно.

Более подробные описания классов, полученных в результате применения методов кластерного анализа, приведены в [4]. Содержательный анализ результатов классификации показывает, 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 и в дальнейшем послужила базой для решения ряда задач прогнозирования.

4. ИССЛЕДОВАНИЕ РАНГОВЫХ РАСПРЕДЕЛЕНИЙ КЛАССИФИКАЦИОННЫХ РАЗБИЕНИЙ

Однако наряду с этим возникает вопрос о достаточности полученной информации о структуре всей совокупности мигрантов. В частности, возникает вопрос о целостности всей совокупности: можно ли рассматривать

Таблица 1

Характеристики классов работников, перераспределяющихся между предприятиями города

№ класса	Удельный вес класса в составе увольняющихся, %	Возраст, лет	Образование	Стаж работы на последнем предприятии, лет	Район рождения*	Район выбытия*	Время проживания в городе, лет
Местные уроженцы							
1	20,0	23—25	8—10	1—4	1,0	—	—
2	3,5	34—38	7—9	10—12	1,0	—	—
3	1,5	27—29	10	8—10	1,0	—	—
4	1,5	43—48	7—10	2—6	1,0	—	—
5	3,5	24—26	Сред. спец.	1—3	1,0	—	—
6	3,0	35—40	Сред. спец. и высшее	2—3	1,0	—	—
Мигранты							
7	4,5	22—28	8—10	2—6	0,5—1,0	0,5—1,0	2—10
8	4,5	20—25	8—10	3—6	1,5—1,7	1,5—1,7	3—8
9	4,5	35—45	6—9	2—10	0,4—1,0	0,4—1,0	10—15

Таблица 2

Характеристики классов в составе увольняющихся, выезжающих за пределы города

№ класса	Удельный вес класса в составе увольняющихся, %	Возраст, лет	Образование	Стаж работы на последнем предприятии, лет	Район рождения*	Район выбытия*	Время проживания в городе, лет
Местные уроженцы							
1	10,5	19—22	8—10	1—3	1,0	—	—
2	2,0	22—24	10	5—6	1,0	—	—
3	3,0	24—26	10	2—4	1,0	—	—
4	3,5	26—30	Сред. и сред. спец.	4—10	1,0	—	—
5	2,5	35—40	10	8—10	1,0	—	—
6	1,5	30—33	7-8	1—3	1,0	—	—
Мигранты							
7	4,0	22—28	10	1—3	0,5—0,7	0,5—0,7	1—3
8	3,5	20—25	10	2—5	0,5	0,5	2—5
9	4,0	35—45	10	1—2	0,5—0,9	1,2—2,2	1—2
10	3,5		Сред. и сред. спец.	2—4	0,2—1,0	0,5—1,3	2—4
11	2,5		Сред. и сред. спец.	2—3	0,8—1,5	0,8—1,5	Свыше 10

эту совокупность как некоторую систему со всеми вытекающими отсюда следствиями, или — рассмотренная совокупность представляет собой отдельные, независимо функционирующие группы. Или рассмотренная совокупность представляет собой часть какой-то более широкой системы и тогда при принятии решений следует учитывать влияние других факторов, не учтенных этим исследованием. Для ответов на поставленные вопросы следует перейти к построению и исследованию ранговых распределений полученных классификационных разбиений.

Для построения ранговых распределений нами была использована система STATISTICA. В соответствии с результатами таблиц 1 и 2 построенные ранговые распределения представлены на рисунках 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 различных классов различны. Это хорошо видно при анализе таблиц 1 и 2.

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

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

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 ситуа-

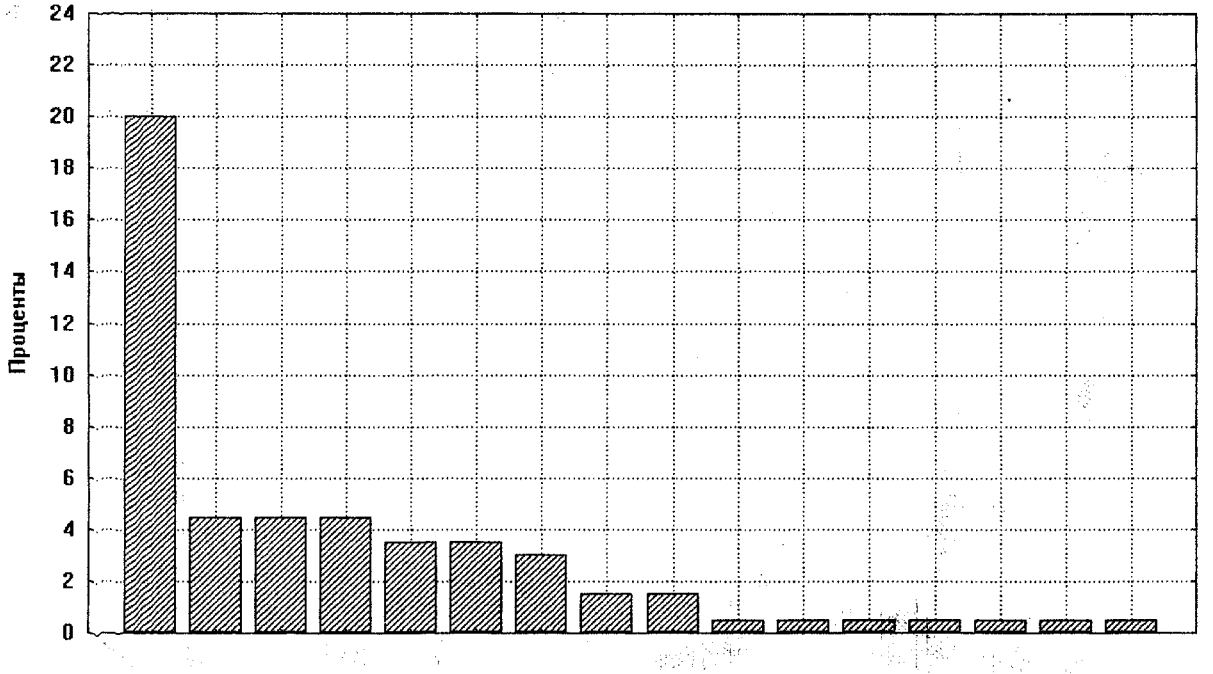

Рис. 1. Ранговое распределение классификационного разбиения таблицы 1

ция 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.

Нарушение 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 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 таблицы 1 построенное уравнение регрессии будет иметь вид $Y = 3,116 - 1,395X$, $(R^2 = 0,88)$, а для данных таблицы 2 — $Y = 2,641 - 1,053X$, $(R^2 = 0,83)$. Оба приведенных регрессионных уравнения значимы на стандартном 5 % 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 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

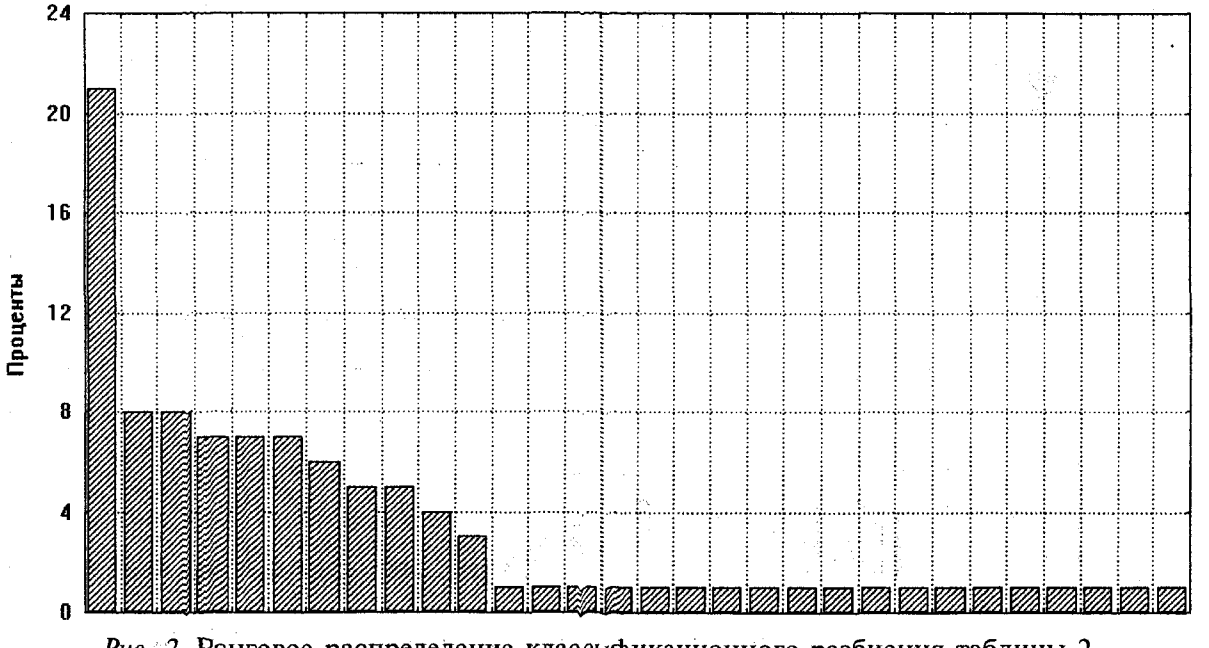

Рис. 2. Ранговое распределение классификационного разбиения таблицы 2

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 параметров. Эта процедура повторялась определенное, достаточно большое число раз, например 1000. Полученные значения параметров в дальнейшем рассматривались как случайные величины. На основании полученной в эксперименте гистограммы можно подобрать функцию распределения или плотность. Задаваясь уровнем 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 оснований для того, чтоб 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 таблице 3.

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

Таблица 3

Результаты оценивания значения параметра рангового распределения (нормальная аппроксимация)

№	Численности классов	Расчетное значение параметра	Нормальная аппроксимация (n = 100)			
			Среднее значение	Стандарт. 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

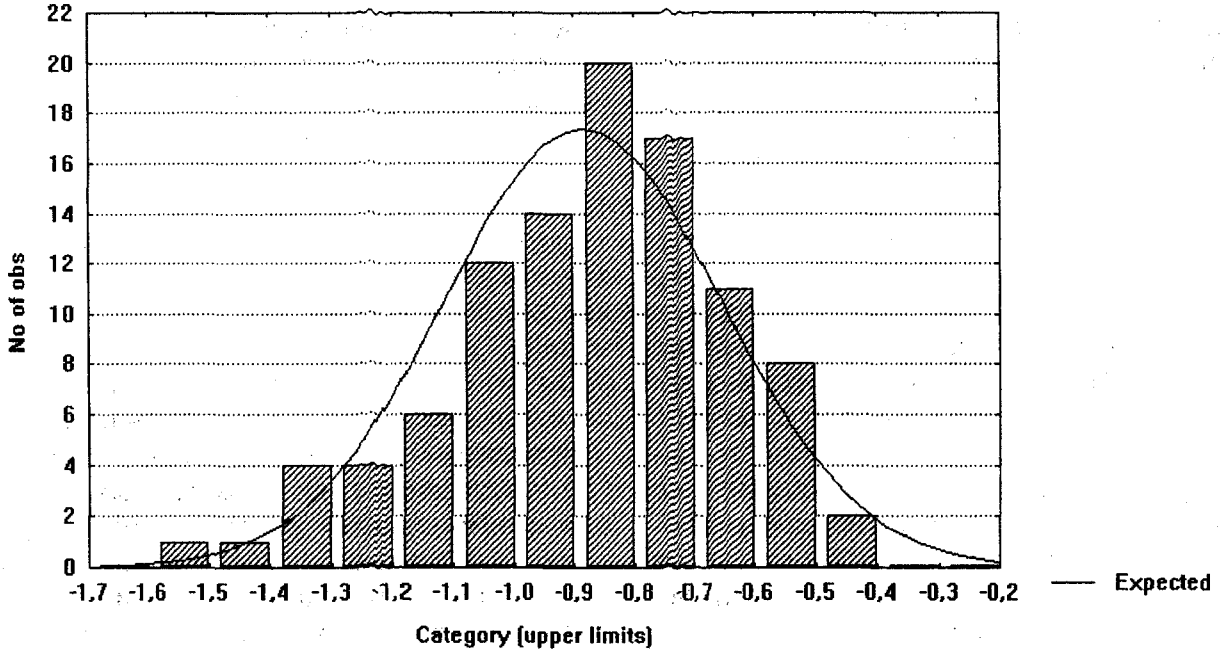

Рис. 3. Гистограмма коэффициента рангового распределения 16 классов в выборке объемом 100 единиц

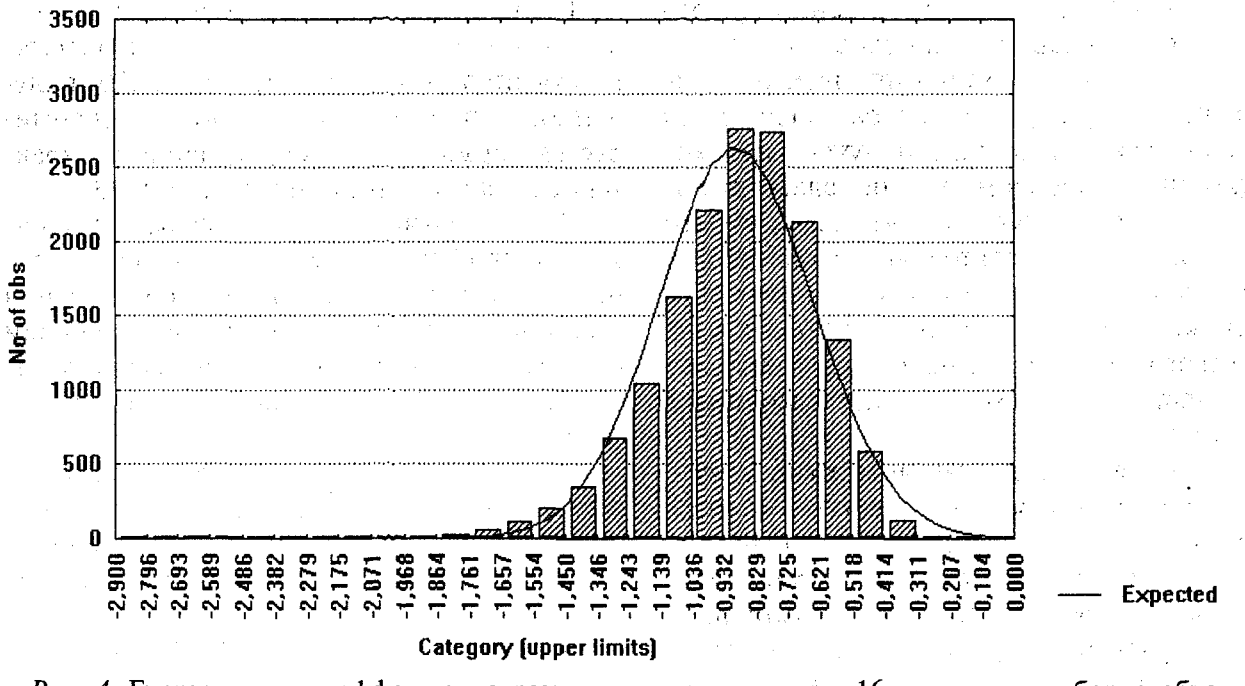

Рис. 4. Гистограмма коэффициента рангового распределения 16 классов в выборке объемом 100 000 единиц

В этом случае построение критической области проводилось посредством определения соответствующих процентилей эмпирического распределения. Один из результатов такого эксперимента представлен в таблице 4.

Анализ представленных результатов показывает, что в действительности ранговое распределение, построенное в отдельности по данным каждой группы нельзя признать ципфовским, т.к. 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 таблице 3 можно видеть, что значение −1,395

Таблица 4

Результаты оценивания значения параметра рангового распределения (метод статистических испытаний)

№	Численности классов	Расчетное значение параметра	Машинная (числовая) аппроксимация (n = 100 000)			
			Среднее значение	Стандарт. 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

принадлежит интервалу (−1,412; −0,367), а значение −1,053 — интервалу (−1,172; −0,518). В таблице 4 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 представляется 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). Графические результаты такого объединения можно видеть на рис. 5, где для сравнения приведены и расчетные значения численностей отдельных классов.

Вывод о выполнение 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) таблицы 3, ни интервалу (−1,036; −0,551). Значит, следует признать, что ранговое распределение, построенное по всей исследуемой совокупности, является ципфовым. Это означает, что только вся исследуемая может быть представлена как система, а отдельные ее части этим свойством не обладают.

Таким образом, как было установлено при исследовании ранговых распределений,

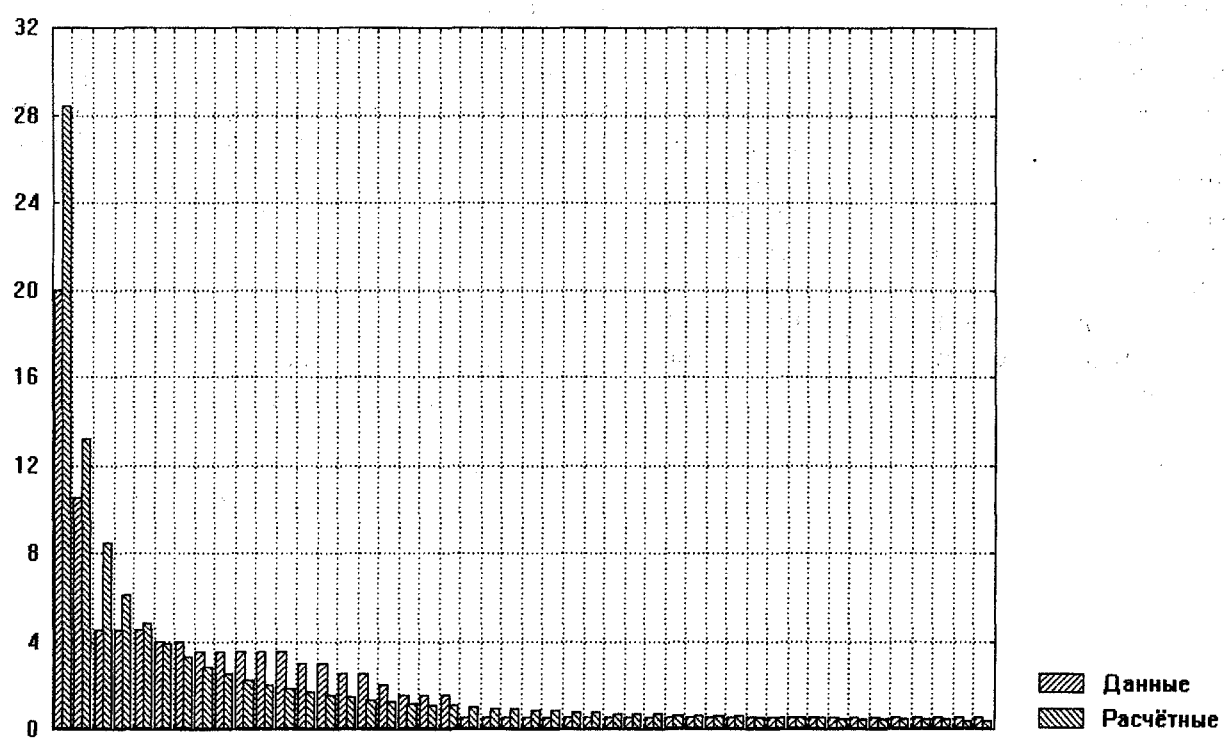

Рис. 5. Ранговое распределение объединенной совокупности

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.

ЗАКЛЮЧЕНИЕ

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 ответить на вопрос о целостности рассматриваемой совокупности.

Принятие решений по управлению системным объектом также должно 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 вспомнить окончательные результаты опыта по укрупнению сельскохозяйственных предприятий в советское время.

ЛИТЕРАТУРА

1. Типология и классификация в социологических исследованиях. — М. : Наука, 1982. — 296 с.

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

3. *Буховец, А.Г.* Модель классификационной задачи / А. Г. Буховец // Вестник. Научно-технический журнал Воронежского государственного технического университета. — Воронеж : 2002, — С. 40—45.

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

5. *Буховец, А.Г.* Использование принципов системного подхода при построении типологий. / А. Г. Буховец // II Всесоюзная конференция «Системное моделирование социально-экономических процессов». — Таллин : 1983. — Ч. I, — С. 25—26.

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

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

8. *Буховец, А.Г.* О механизме формирования ципфовских распределений при моделировании урожайности зерновых культур / А. Г. Буховец, С. Н. Дементьев, Л. П. Яновский, Т. Е. Хоршева // Труды III Международной конференции «Математика. Компьютер. Образование» — М., 1996, — С. 71—76.

9. *Шрейдер, Ю.А.* Системы и модели / Ю. А. Шрейдер, А. А. Шаров. — М. : Радио и связь, 1982. — 152 с., ил.

10. *Хайтун, С.Д.* Наукометрия. Состояние и перспективы / С. Д. Хайтун. — М. : Наука, 1983, — 344 с.

11. *Яблонский, А.И.* Математические модели в исследовании науки / А. И. Яблонский. — М.: Наука, 1986.