УДК 311.34

НЕКОТОРЫЕ ПОДХОДЫ К АНАЛИЗУ ДАННЫХ В АРХЕОЛОГИИ

Ю.П. Холюшкин
Государственная публичная научно-техническая
библиотека СО РАН
E-mail: Kholush45@gmail.com

В.С. Костин
Институт экономики и организации
промышленного производства СО РАН
E-mail: kostin@ieie.nsc.ru

Для реализации современных задач археологии необходимо использовать одно из новых направлений искусственного интеллекта – «интеллектуальный анализ данных», который является кратким и неточным переводом с английского языка терминов Data Mining и Knowledge Discovery in Data bases (DM&KDD).

Авторами на протяжении ряда лет проводились исследования по статистическому анализу данных в археологии. В ходе этих исследований была разработана последовательность применения методов в археологии, при анализе цитирования, выявления научных школ.

В ходе этих исследований применялись: методы дисперсионного анализа; методы факторного анализа и многомерного шкалирования; кластеризация. С помощью кластеризации средства Data Mining самостоятельно выделяют различные однородные группы данных; процедура выявления структуры таблицы. В статье приводятся специально разработанные средства для упорядочения неоднородной археологической информации и выявления ее структуры. Метод повторной выборки с возвращением. Сравнение классификаций и построение обобщенной классификации. Кроме того, в работе реализован метод построения сводной обобщенной классификации, основанный на анализе совпадения разных классификаций одних и тех же объектов.

Ключевые слова: дисперсионный анализ, кластерный анализ, метод повторной выборки с возвращением, сравнение классификаций, обобщенная классификация.

SOME APPROACHES TO THE ANALYSIS OF DATA IN ARCHEOLOGY

Yu.P. Kholyushkin
State Public Scientific Technological Library of the SB RAS
E-mail: Kholush45@gmail.com

V.S. Kostin
Institute of Economics and Industrial Engineering of the SB RAS
E-mail: kostin@ieie.nsc.ru

In order to accomplish up-to-date tasks, archeology must use one of the recent areas of Artificial Intelligence – «intellectual data analysis» which is a brief and imprecise translation from English of the terms «Data Mining» and «Knowledge Discovery in Databases» (DM & KDD).

For several years, the authors have been conducting research on the statistical analysis of data in archeology. In the course of these studies we have developed a coherent applica-

tion of methods in archeology with the analysis of citation and identification of scientific schools.

In the course of the studies we used: the methods of analysis of variance; methods of factor analysis and multidimensional scaling; clustering. With the help of clustering, Data Mining allocates various homogeneous groups of data; a procedure to detect the structure of the table. The article provides specially designed tools for ordering an inhomogeneous archaeological information and identifying its structure. The method of re-sampling with replacement. Comparison of the classifications and forming of a generalized classification. In addition, the article works out the method of forming a generalized consolidated classification, based on an analysis of coincidence of different classifications of the same object.

Keywords: analysis of variance, cluster analysis, a method of re-sampling with replacement, comparing of classifications, generalized classification.

Постановка задачи. Первым шагом процесса анализа данных в археологии является четкое определение проблемы и рассмотрение способов использования данных для их решения. «Действительно от характера задач зависит выбор исследуемых археологических материалов, способ их видения, классификация, комментирование и, наконец, даже мера ценности результатов, о чем можно судить, только соотнеся их с целью и задачами построения» [2].

Этот шаг включает анализ требований, определение области проблемы, метрик, по которым будет выполняться оценка модели, а также определение задач для проекта анализа данных.

Все шаги располагаются один за другим при следующих условиях.

1. Постановка задач последовательно оказывает влияние на все остальные звенья исследования, включая оценку достоверности полученных результатов.

2. Если создается компиляция, процесс останавливается на шагах «подготовка данных» и «просмотр данных».

3. Блок построения моделей играет двоякую роль: как динамическую связь между звеньями в пределах одного построения, либо между разными, альтернативными или дополняющими дуг друга построениями. В приблизительном виде это напоминает аналогичные построения Ж-К. Гардена.

Таким образом, логистическое требование, предъявляемое к первому базовому шагу схемы, состоит в перечислении возможных альтернативных задач и в обосновании выбора той или иной проблемы.

Именно из этого главного научного метода вытекает вся совокупность приемов (частных методов) и этапов исследования, которую мы зовем методикой исследования. Ведь нужно осуществить такую группировку фактов, из которой можно получать обобщения и предположения о закономерностях, причинах и зависимостях, чтобы затем извлекать следствия из этих гипотез (ожидания) и проводить систематическую проверку этих следствий на все новых наблюдаемых фактах [8].

Эти задачи можно сформулировать в виде следующих вопросов:

– Что необходимо найти? Какие типы связей необходимо выявить?

– Отражает ли поставленная задача логические правила или процессы в прошлом?

– Надо ли делать прогнозы на основании модели анализа данных или просто найти содержательные закономерности и взаимосвязи?

– Какой результат или атрибут необходимо спрогнозировать?

– Какие виды данных нужно иметь и какого рода информация находится в каждом столбце таблицы?

– Если существует несколько таблиц данных, то как они связаны между собой?

– Нужно ли выполнять очистку данных от «шума», статистическую обработку, чтобы данные стали применимыми?

– Каким образом распределяются данные? Дают ли данные точное представление об исторических процессах далекого прошлого?

Однако на пути реализации этой творческой работы в археологической практике имеются существенные трудности, обусловленные информационными проблемами археологии. Все они в той или иной мере связаны со сбором и отбором наблюдений, их анализом и интерпретацией. Среди этих проблем следует отметить особо значимые проблемы:

Неполнота и фрагментарность археологической информации, объясняемая как дискретностью самих археологических данных, так и ограниченностью их использования. В первом случае неполнота зависит от степени сохранности и исследованности археологического памятника, а во втором определяется недостаточностью списка признаков, используемых в исследовательских процедурах. Очевидно, такая информация не дает адекватного представления о действительном состоянии исследуемой проблемы и особенно в тех случаях, когда остается неизвестной та ее часть, которая учтена и использована в каждой конкретной процедуре. Наиболее остро эта проблема возникает, когда археолог производит слабо обоснованную селекцию информации. Зачастую подобная селекция производится при отсутствии концепции у исследователя, а это в свою очередь не позволяет сделать выбор из большого ряда потенциальных характеристик, присутствующих в массиве изучаемого материала. Возникает задача восполнения данных, решение которой – самостоятельная проблема.

Несопоставимость данных. Одна из форм несопоставимости связана с использованием различных классификационных построений. Решение проблемы требует, как правило, проведения комплексных исследований, в которых каждый аспект изучается на основе наполнения некоторой однородной формы данных.

Неадекватность применяемых процедур статистического анализа данных поставленной задаче. Часто археологи применяют статистические методы не потому, что он необходим, а потому, что его знают. Отсюда возникают заблуждения относительно того, что применяемый метод дает сразу ответ положительный или отрицательный на поставленную задачу.

Устойчивость исходных и выделенных структур. В предлагаемом нами контексте эта проблема в археологии ранее не рассматривалась. Что касается традиционной археологической парадигмы, то в ней предполагается постепенность культурных изменений, а в случае, когда они не наблюдаются, разрывы в структурных культурных образованиях объясняются сменой населения. При таком подходе обычно высказываются следующие предположения:

а) коллекции, относящиеся к одному и тому же времени, должны быть примерно одинаковы;

б) различия между коллекциями фиксируют направленные изменения [13], отражающие «развитие» форм артефактов из предшествующих форм [11]. Указанные взгляды покоятся на модели культуры, учитывающей лишь три показателя изменений в устойчивости структур: миграция, изобретение и диффузия, как процессы культурной истории [18]. При этом остаются в стороне другие факторы, связанные с контекстом памятника, сырьем, репрезентативностью используемых выборок, субъективностью, вносимой исследователем в источник исследования (специальная подборка материала и др.). Все эти факторы могут оказывать немаловажное влияние на устойчивость выделяемых археологических структур;

в) внесение ошибок в археологические данные (погрешности арифметических подсчетов, недочеты измерений, ошибки в определении типов артефактов (неправильные дефиниции).

В свое время Д. Кларк указывал на возникновение опасности, что альтернативное или противоречащее определение типов артефактов коренным образом изменяет подробно рассмотренные процентные соотношения и соответственно их смысловую интерпретацию [12]. На Ближнем Востоке пытаются решить эту проблему преодоления ошибок, внутренне присущих типологическому анализу, путем использования в археологических штудиях только тех типов ретушированных орудий, которые всегда могут быть идентифицированы и отделены от других орудий любым исследователем. Для иллюстрации можно привести цитату из статьи А. Маркса: «.. каждый, работавший в Леванте после Д. Гаррод, установил для себя разницу между продольными скреблами и концевыми скребками, между ножами с обушком и плоскоретушированными скреблами. Таким образом, эти группы орудий пригодны для нашего исследования. С другой стороны, ракле и псевдолеваллуазские острия, например, не всегда определимы и до сих пор еще не всеми единообразно включены в типологические списки комплексов. То же самое можно сказать о выемчатых орудиях и мустьерских транше» [16];

г) методические просчеты;

д) неверно поставленные задачи и т.д.

Здесь следует не только выделить такие этапы экспертизы подлинности археологического источника, как проверка сохранности памятника, степень его изученности, информативности и субъективизм выборки, произведенной археологом, но и учитывать проблему дальнейшей судьбы той или иной коллекции после раскопок.

Просмотр данных. Следующим шагом процесса интеллектуального анализа данных является просмотр подготовленных данных. Для принятия правильных решений при создании моделей интеллектуального анализа данных необходимо понимать данные. Методы исследования данных включают в себя расчет минимальных и максимальных значений, вычисление средневероятного и стандартного отклонения и изучение распределения данных. Например, по максимальному, минимальному и среднему значениям можно заключить, что выборка данных не является репрезентативной для имеющихся процессов, и поэтому необходимо получить более сбалансированные данные или изменить предположения, лежащие в основе ожидаемых результатов. Стандартное отклонение и другие характеристики распределения могут сообщить полезные сведения о стабильности и точ-

ности результатов. Большая величина стандартного отклонения может свидетельствовать о том, что добавление новых данных поможет усовершенствовать модель. Данные, которые сильно отклоняются от стандартного распределения, могут оказаться искаженными или представлять точную картину реальной проблемы, которая делает сложным подбор соответствующей модели для данных.

Изучение данных в свете собственных представлений о проблеме может привести к выводу о наличии ошибок в наборе данных, и затем можно выработать стратегию для устранения проблем или получить более глубокое представление о моделях археологических данных.

Предварительный анализ данных. Прежде чем приступать к глубокому статистическому анализу данных, полезно ознакомиться с каждым элементом данных по отдельности. Будем предполагать, что данные представлены таблицей «объект–свойство», т.е. каждый объект, будь то отдельная находка или целый слой археологического памятника, характеризуется некоторым набором свойств (признаков), в совокупности составляющих его описание.

Например, свойствами слоя могут быть различные типы артефактов или фаунистические остатки, обнаруженные в этом слое. Для находок, например, каменных артефактов, это может быть материал и количественное и качественное описание артефактов. Каждое свойство обладает своей описательной силой, зависящей от двух моментов: во-первых, насколько разнообразны его значения и, во-вторых, насколько существенные стороны предмета исследования оно отражает. Описательная статистика помогает оценить разброс значений признаков путем построения гистограмм частот, распределений и различных статистик.

В статистике наиболее сильные (т.е. наиболее общие и важные) выводы можно сделать относительно «хороших» переменных. Такими переменными являются количественные, нормально распределенные случайные величины. С помощью описательной статистики исследователь может выяснить, насколько его данные близки к идеалу. Но даже если эти данные далеки от идеала, то в статистике всегда найдутся средства, чтобы сделать обоснованные выводы из их анализа.

Первая задача исследователя, приступающего к статистическому анализу, состоит в определении для каждого признака типа шкалы, в которой он измерен. Для этого достаточно различать три шкалы: номинальную, порядковую и количественную.

В задачу предварительного анализа входит проверка корректности данных. Ошибку в данных легче увидеть на графике, чем в таблице. Например, для количественной переменной ошибки (опечатки) часто проявляются в виде выпадающих значений, отстоящих на значительном расстоянии от основной массы значений.

Другой, не менее важной задачей предварительного анализа данных является поиск ответа на вопрос, обладает ли какой-либо (явной или скрытой) структурой анализируемая таблица данных. Достаточно простым и эффективным средством является «серый» (или «спектральный») анализ (рис. 1). Его суть состоит в том, что анализируемая таблица дополняется графической схемой, которая представляет собой образ таблицы в виде прямоугольника, разделенного на ячейки, подобно клеткам исходной та-

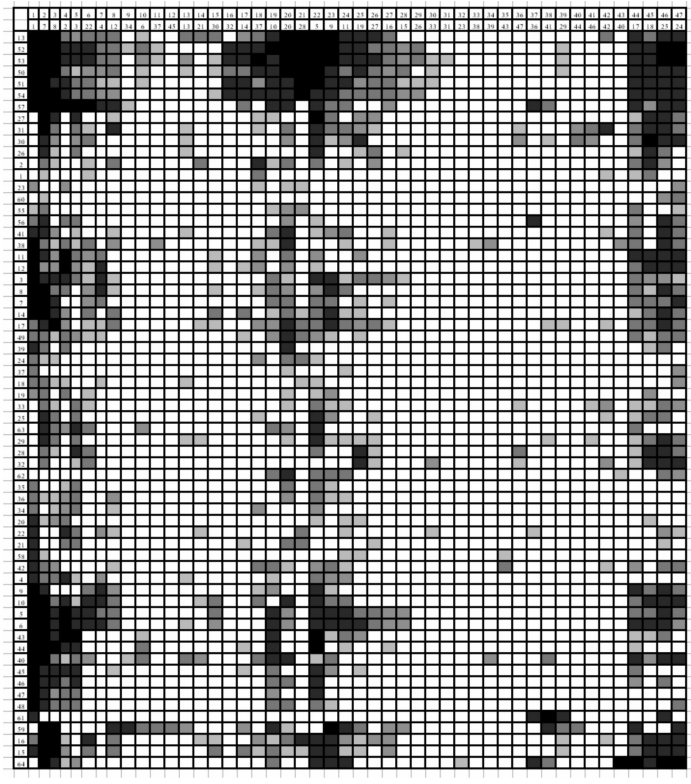

Рис. 1. Выявленная с помощью «серого» анализа структура данных

блицы. При «сером» анализе каждая клетка схемы заполняется (заливается) оттенком серого цвета в зависимости от того, какие значения принимает соответствующий признак для данного объекта. Предварительно промежуток, в который попадают числовые значения всех признаков, разбивается на конечное число равных интервалов. Каждому интервалу сопоставляется определенный оттенок серого цвета по правилу – чем больше значения признаков, которые попадают в данный интервал, тем темнее окрашиваются в серый цвет соответствующие клетки таблицы. Результатом серого анализа является наглядный образ данных, где их структура представлена наиболее отчетливо.

Таким образом, графическая схема выглядит как своего рода плоская географическая карта, выполненная оттенками серого цвета, чем и объясняется название соответствующего метода анализа данных. На аналогичном принципе построен метод анализа с помощью оттенков разного цвета («спектральный» анализ), при котором данные таблицы представляются некоторой палитрой разных цветов. Этот метод дает еще более наглядную картину. Явная структура обычно обнаруживается при взгляде на графическую схему, если на ней контрастно выделяются зоны (области) сгущений и разреженностей. В зонах сгущений (кластерах) концентрируются клетки с заметными (существенными) значениями признаков. В зонах разреженностей значения признаков представлены малыми (или нулевыми) значениями признаков.

Для выявления скрытой структуры требуется соответствующее преобразование исходной таблицы данных, достигаемое с помощью перестановки строк и (или) столбцов.

Рассмотрим пример данных, взятый из [4]. На рис. 1 приведена таблица, в которой затененная верхняя строка фиксирует номера орудийных комплексов, а левый затененный столбец – номера археологических памятников, на которых эти орудийные комплексы были найдены. Соответственно на пересечении строк и столбцов указано количество находок. Нулевые ячейки (означающие, что данные пропущены или соответствующие орудия на памятниках не найдены) не заполнены с той целью, чтобы значимые данные были более заметны. При беглом взгляде на рис. 1 видно, что если эта таблица и имеет какую-либо структуру, то она (эта структура) для наблюдателя представляется скрытой. Поэтому, упорядочивая таблицу перестановкой строк и столбцов, добиваемся того, чтобы эта структура обнаружилась. В наглядной форме эта структура представлена с помощью «серого» анализа.

Для этой цели перед предварительным упорядочиванием сначала были пронормированы значения признаков таблицы так, чтобы каждому исходному значению было поставлено в соответствие значение его ранга в таблице (ранговая статистика). Затем построенная таким образом таблица ранговых статистик была упорядочена перестановкой строк и столбцов. Таким образом, было построено наглядное представление структуры данных, выявленное в результате их «серого анализа».

Перестановка строк и столбцов при упорядочении таблицы ранговых статистик осуществлялась на использовании следующей идеи. Как правило, хорошо структурированной таблицей является та, в которой не очень часто происходят скачки по величине значений соседних элементов. Поэтому при перестановке строк и столбцов таблицы ранговых статистик эти данные были упорядочены по строкам и соответственно по столбцам таким образом, чтобы суммы расстояний между соседними элементами стали минимальными. Благодаря такой перестановке, строки, соответствующие памятникам, оказались упорядоченными по близости их распределений по артефактам.

Полученная картина показала, что данные таблицы на самом деле обладают некоторой структурой, и, таким образом, имеет смысл с помощью статистических методов анализа исследовать ее более тщательно.

Для этой цели количественные данные таблицы на рис. 1 были вновь упорядочены по строкам и соответственно по столбцам. Причем упорядочение данных исходной таблицы выполнялось по такой же схеме, что и упорядочение таблицы ранговых статистик.

На основе этого упорядочения было произведено разбиение матрицы на существенные с точки зрения информативности области.

Было выделено 20 связных областей.

Для визуализации результатов выделения каждая из смежных областей была окрашена единым оттенком серого цвета. В итоге структура данных исходной таблицы предстала еще более отчетливой (рис. 2).

Наконец, завершается предварительный анализ данных визуализацией наполненности связных областей числовыми данными (рис. 3).

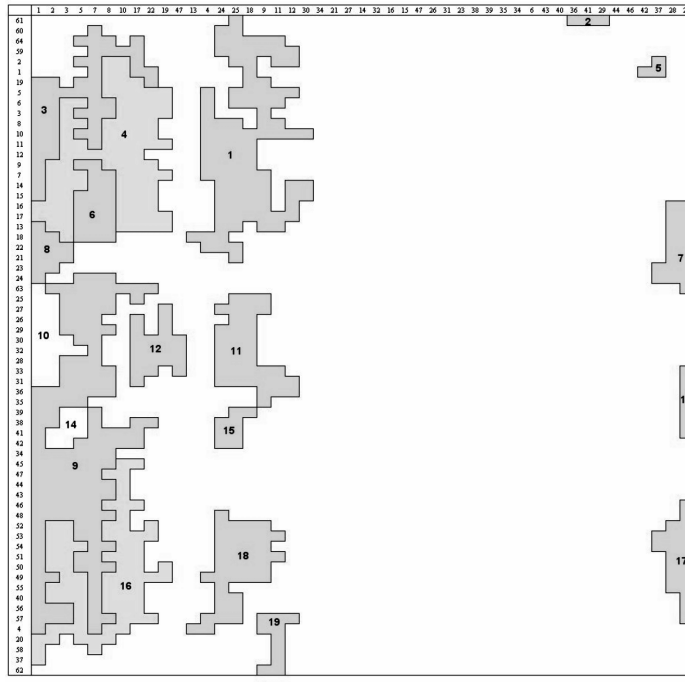

Рис. 2. Выделение смежных областей

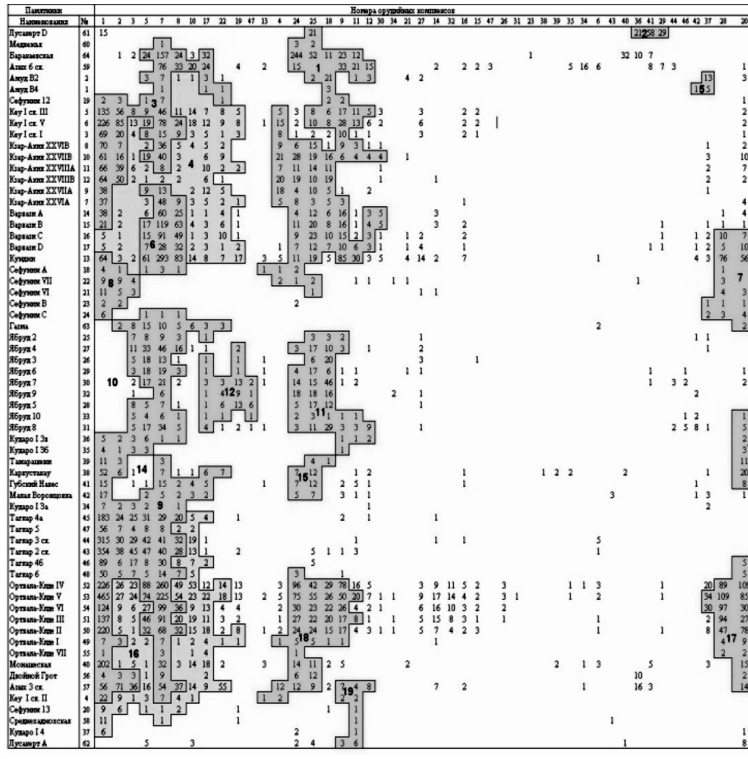

Рис. 3. Наполнение смежных областей

Исследование связей и зависимостей. Следующим шагом анализа является обнаружение взаимосвязей и зависимостей между признаками. В статистике разработано множество критериев для проверки наличия связей в данных. Но все они построены по одному принципу: в каждом критерии формулируется своя нулевая гипотеза, которая утверждает, что исследуемые признаки являются независимыми случайными величинами, связь между которыми если и проявляется, то исключительно в силу случайного совпадения. Проверка любого критерия начинается с вычисления своей статистики – величины, характеризующей степень отклонения от независимости. Вычисляемая статистика является количественной переменной и подчиняется в условиях выполнения нулевой гипотезы определенному распределению, которое может быть аналитически рассчитано или аппроксимировано программой. Таким образом, значение статистики переводится в так называемую значимость, которая является не чем иным, как вероятностью наблюдения полученного значения этой статистики при выполнении нулевой гипотезы. Если эта вероятность ниже некоторого заранее выбранного порога, например, 5 %, то исследователь имеет основания утверждать, что нулевая гипотеза не подтверждается на его данных, из чего с большой вероятностью следует вывод, что между признаками обнаруживается определенная связь.

Поскольку признаки могут быть измерены в любой из трех шкал (номинальной, порядковой и количественной), то для каждого сочетания шкал надо применять свои критерии. Например, если обе переменные измерены в шкале наименований, то можно применять критерий Хи-квадрат, если одна из них – номинальная, а другая – количественная, то можно пользоваться дисперсионным анализом, а если обе количественные, то подойдет корреляция по Пирсону. При сочетании порядковой и количественной переменных приходится огрублять количественную переменную до порядковой и применять методы ранговой корреляции.

Для обычного исследователя, не владеющего в совершенстве методами анализа данных, одна из основных трудностей при работе с пакетами статистических программ заключается в сложности ориентации среди большого числа предлагаемых методов и критериев, которые в основном носят имена их авторов. Чтобы разобраться в том, а для решения каких именно задач применяется тот или иной метод, необходимо самостоятельно изучать специальную литературу.

В данной разработке для преодоления этого препятствия предполагается упорядочить методы по решаемым задачам и условиям (например, сочетаниям шкал признаков). Кроме того, вместо теоретических соображений в пользу того или иного критерия, мы можем привести практические доводы. А точнее – непосредственно на данных пользователя проверять применимость каждого критерия и рекомендовать лишь те, которые покажут свою работоспособность прямо «на глазах изумленной публики». Для проверки работоспособности критерия связи необходимо смоделировать условия нулевой гипотезы об отсутствии связи между переменными. Это можно сделать, разрушив связь между переменными, для чего достаточно перемешать любую из них, т.е. переставить ее значения в случайном порядке, не изменяя самих значений. Многократное применение критерия к пе-

ремешанным данным должно давать случайные результаты – значимость при этом должна быть равномерно распределена в интервале от нуля до единицы. Если распределение значимости не будет равномерным, значит, критерий на предоставленных данных не работает. Если все критерии откажутся работать, то такие данные подлежат отбраковке как не пригодные к статистическому анализу связей.

Дополнительным сервисом системы может быть проверка чувствительности критериев к связям разного рода – линейным и нелинейным. Для такой проверки необходимо найти способ моделировать связи между переменными путем неслучайного перемешивания.

Методы снижения размерности. К методам снижения размерностей относятся факторный анализ (метод главных компонент) и многомерное шкалирование.

Эти методы позволяют из многих десятков малоинформативных признаков построить несколько высокоинформативных факторов, содержащих «отжатую» информацию, неравномерно разбросанную по исходным признакам. На примерах можно убедиться, что оба этих метода дают близкие результаты, но метод главных компонент алгоритмически более простой и эффективный, поэтому мы в дальнейшем сосредоточимся на нем.

Факторный анализ, кроме снижения размерности, дает косвенную возможность исследовать связи многих признаков, ибо факторы можно представить в виде линейной комбинации исходных признаков. Те признаки, которые входят с наибольшими коэффициентами в разложение факторов, образуют группы высококоррелированных признаков (табл. 1 – пример взят из монографии) [4].

Здесь из 47 признаков (типов орудий) предварительно выделены 7 главных компонент, по факторным нагрузкам которых из множества типов орудий были отобраны 19 типов как наиболее значимых. Методом вращения Varimax, при котором максимизируется дисперсия (разброс) факторных нагрузок каждого компонента на исходные переменные, получена матрица нагрузок этих переменных на отобранные главные компоненты (табл. 1).

Таким образом, факторный анализ за один проход выделяет все группы связанных переменных. Этот побочный результат данного метода часто оказывается полезен при изучении связей. Поскольку факторы представляют линейную комбинацию исходных переменных, то полный набор факторов содержит в точности то же количество информации, что и набор исходных признаков. Сущность метода можно понять, представив себе данные как точки в пространстве с размерностью, равной числу исходных переменных. Точки в этом многомерном пространстве сосредотачиваются в некотором компактном облаке рассеяния. Поиск главных компонент сводится к такому вращению системы координат, при котором вдоль первого фактора наблюдается наибольший разброс точек, вдоль второго – меньше и далее по убыванию.

Но дело в том, что случайные величины всегда содержат шум, вызванный ошибками сбора данных и случайными отклонениями в параметрах объектов. Природа этих ошибок может быть самой разной, но они всегда есть. Значит, и полный набор факторов также содержит шум. Главные компоненты, начиная с фактора номер один, характеризуются наиболь-

Таблица 1

Факторные нагрузки на типы орудий

№ типов орудий	Типы орудий	Факторные нагрузки на типы орудий						
		1	2	3	4	5	6	7
7	Продольные скребла	0,880	0,196	0,041	–0,006	0,154	–0,140	0,070
8	Двойные скребла	0,815	–0,209	–0,092	0,055	–0,133	–0,021	0,056
9	Конвергентные скребла	0,671	–0,102	–0,093	–0,045	0,026	0,476	–0,057
25	Зубчатые	–0,050	0,891	0,074	0,027	–0,079	0,003	0,068
24	Выемчатые	–0,096	0,738	–0,146	–0,337	–0,001	–0,083	–0,014
1	Леваллуазские сколы	–0,404	–0,507	–0,459	–0,176	–0,218	0,023	0,145
2	Леваллуазские острия	–0,351	–0,433	–0,229	–0,132	–0,247	–0,132	0,305
18	Резцы	–0,016	0,182	0,838	0,085	–0,116	–0,099	0,036
37	Разные	–0,133	–0,293	0,680	–0,183	0,060	–0,175	–0,289
22	Усеченные отщепы	0,044	–0,040	0,608	–0,024	0,198	–0,008	0,369
19	Проколки	–0,119	0,367	0,432	0,015	–0,287	0,152	0,206
5	Мустьерские острия	0,186	0,004	0,049	0,871	–0,012	0,053	–0,052
3	Леваллуазские ретушированные острия	–0,173	–0,139	–0,088	0,855	–0,067	–0,156	0,049
17	Скребки	–0,027	0,062	0,130	–0,145	0,869	–0,137	–0,023
10	Угловатые скребла	0,112	–0,142	–0,188	0,107	0,673	0,471	–0,037
11	Поперечные скребла	–0,025	0,020	–0,096	–0,092	–0,011	0,885	–0,117
20	Ножи	–0,252	–0,037	–0,141	–0,036	0,185	0,272	–0,654
4	Псевдолеваллуазские острия	–0,158	–0,119	–0,136	–0,259	0,221	–0,178	0,623
28	Сколы, ретушированные со спинки	0,005	–0,142	–0,167	–0,273	0,162	–0,340	–0,613

шим отношением сигнала к шуму. Чем больше номер фактора, тем меньше полезной информации он содержит. При некотором критическом номере фактора уровень шума становится выше уровня сигнала. Этот и все последующие факторы должны быть отброшены.

Определить количество факторов, которые надо отбросить, достаточно просто. Для этого воспользуемся тем же подходом, что при проверке применимости критериев связи. Действительно, мы можем сформулировать нулевую гипотезу: все исходные переменные независимы друг от друга. Промоделировать выполнение нулевой гипотезы не представляет труда – достаточно перемешать каждую исходную переменную (может быть, кроме одной) и повторить расчет факторов. Все полученные таким образом факторы, начиная с первого, не содержат никакой информации о связях, а только статистический шум, поскольку это гарантировано выполнением условий нулевой гипотезы. Такой статистический эксперимент можно провести многократно. Если теперь сравнить факторы, полученные по исходным данным с теми, что получились в результате экспериментов с перемешиванием, мы сможем определить количество факторов, действительно содержащих больше полезной информации о связях признаков, чем статистического шума.

Сжатие информации с помощью главных компонент является часто подготовительным этапом для структурного анализа, к рассмотрению которого мы и переходим.

Анализ структур. Если анализ связей выявляет признаки, значения которых согласованно изменяются от объекта к объекту, то анализ структур выявляет объекты, на которых согласованы значения определенного набора признаков.

Промежуточное место между анализом связей и анализом структур занимает метод прямого кластерного анализа, разработанный П.С. Ростовцевым. Этот метод позволяет непосредственно на таблице объект–признак выделить после переупорядочения строк и столбцов области неправильной формы, выделяющиеся близкими значениями признаков. Каждая такая область объединяет несколько объектов и несколько признаков в «пятно». Попытки проинтерпретировать наблюдаемую картину областей могут натолкнуть исследователя на новые, интересные гипотезы. Пример применения прямого кластерного анализа взят из [5] (табл. 2, рис. 4).

В частности, рассмотрение структуры областей на рис. 5 приводит к выводу о том, что хорошими признаками для выделения структурной информации являются IF и ILam. Как для всех вышерассмотренных методов, здесь также имеется возможность сформулировать и проверить нулевую гипотезу об отсутствии выделенных областей. В качестве статистики для измерения отклонений от независимости можно выбрать долю объясненной дисперсии (в нашем примере 88,5 %), также называемой коэффициентом детерминации R^2. Такое дополнение превращает метод из эвристического в статистический метод. А это позволяет использовать полученные результаты не только для научного поиска, но и делать вполне обоснованные утверждения.

К методам анализа структур относятся в первую очередь методы автоматической классификации, среди которых наиболее распространен кластерный анализ. Для проведения кластерного анализа необходимо выбрать несколько признаков – построить так называемое признаковое пространство. Задача состоит в том, чтобы выделить в этом пространстве отдель-

Таблица 2

Дисперсионный анализ областей

Область	Среднее	Ср. кв. откл.	Объем	Объясняет долю дисперсии (в %)	
				область	элемент
0	30,920	9,469	10	4,4	0,440
1	51,000	4,546	6	21,7	3,618
2	7,356	4,844	52	32,5	0,624
3	30,657	7,647	7	2,9	0,419
4	26,860	5,459	5	0,9	0,171
5	25,800	7,354	2	0,2	0,121
6	62,000	0,000	1	6,7	6,661
7	43,889	6,889	9	19,3	2,141

Примечание. Среднеквадратичное отклонение по таблице составляет 16,99 %. Объясненная дисперсия составляет 88,5 %.

№	Комплексы	IF	IFst	IL	ILam
1	Семиганч	37,10	14,20	20,20	30,70
24	Оби-Рахмат 15–18	–	–	–	35,50
25	Оби-Рахмат 10–14	–	–	–	34,40
9	Чингиз	54,00	29,00	41,10	0 23,00
11	Актогай	48,60	8,60	44,00	6,60
13	Семизбугу B	57,60	15,20	9,80	2,80
12	Семизбугу A	1 50,80	6,00	9,80	2,80
5	Георгиевский Бугор	44,30	8,80	9,50	10,80
18	Кутурбулак	50,70		2,40-	20,00
19	Зирабулак	36,90	27,30	2,00	3,10
10	Кош-Курган	42,30	13,70	3,90	6,10
16	Хантау	3 35,70	3,60	3,80	2,10
17	Бурма	26,80	9,80	0,80	2,10
6	Кара-Бура	23,60	11,50	4,20	6,30
22	Кызыл-Тау пл.2 сдф	22,00	15,00	2 3,30	0,60
20	Кызыл-Тау пл.1 сдф	15,80	11,80	7,10	0,10
21	Кызыл-Тау пл.1 слдф	9,80	7,50	0,00	0,40
15	Семизбугу D	14,30	4,80	6,80	4,10
14	Семизбугу C	21,40	11,90	11,40	7,00
3	Тоссор	25,60	4,40	12,20	5 20,60
1	Хонако 3	4 24,40	9,00	14,00	31,00
2	Худжи	27,00	6,00	14,50	6 62,00
8	Огзи-Кичик	35,90	4,60	32,70	44,00
26	Оби-Рахмат 6–9	–	–	–	43,40
27	Оби-Рахмат 2–5	–	–	–	7 43,90
23	Оби-Рахмат 19–21				44,20
4	Джар-Кутан	55,00	41,00	37,80	53,00

Рис. 4. Результат работы метода прямого кластерного анализа

ные сгущения точек – кластеры. Разнообразные типы кластерного анализа активно применялись в археологических исследованиях. В ходе таких исследований было обнаружено, что кластеры, замечательным образом найденные в первый раз и разумно описанные исследователем, после повторного сбора информации (новых раскопок и нового применения кластерного анализа) могут «рассыпаться» из-за случайности выявленной кластерной структуры (при малых выборках, ненормальных распределениях, плохо обусловленных моделях и т.д.).

С этой целью используется метод повторной выборки с возвращением, известный как метод boot-strap [14]. К сожалению, этот метод мало затронул археологию. Так, Кинти [15] использовал выборку методом Монте-Карло, чтобы генерировать псевдодоверительные интервалы для результатов анализа многообразий *k*-значной кластеризации пространственных данных. Рингроуз [17] использовал boot-strap для оценки подобным способом результатов анализа соответствия.

Суть метода состоит в имитации повторного сбора данных, в ходе которой генерируется выборка, совпадающая с исходными данными [6].

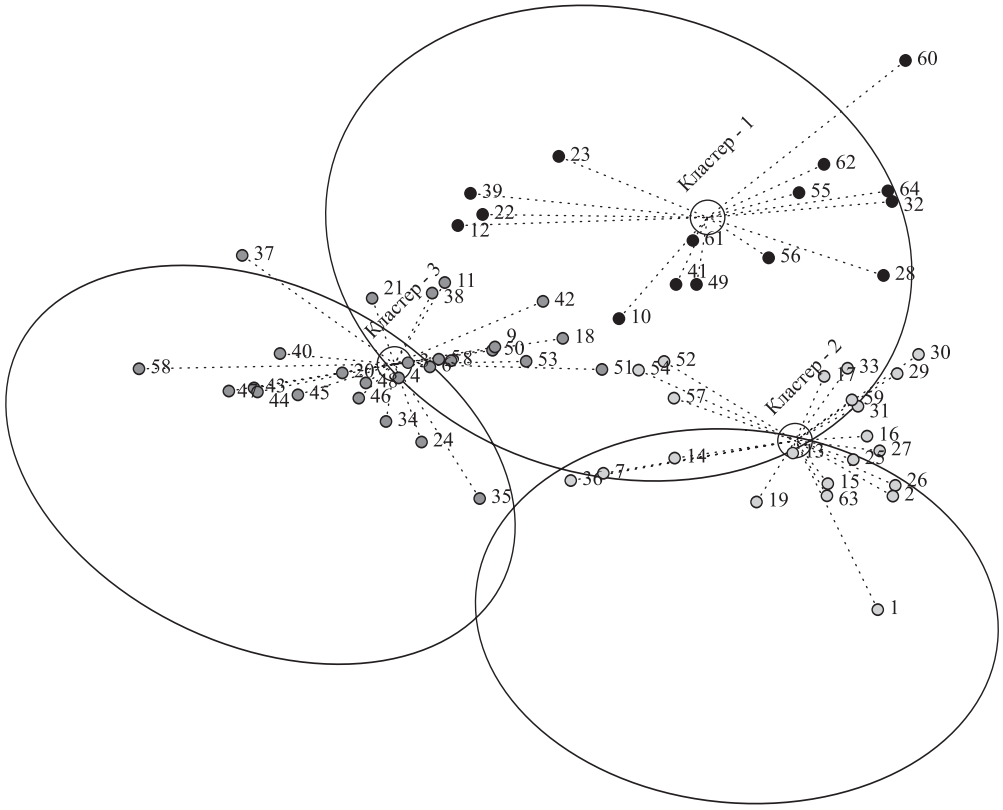

Рис. 5. Результат работы кластерного анализа методом *k*-средних

Существует множество алгоритмов для выделения сгущений, основанных на различных подходах. Мы выбрали один из простейших и вычислительно наиболее эффективных методов кластерного анализа – метод *k*-средних. Согласно этому методу, принадлежность объекта к кластеру определяется эвклидовым расстоянием между объектом и центром кластера. Объект приписывается к ближайшему кластеру. Процедура начинается с некоторого начального приближения, а затем запускается итерационный процесс, на каждом шаге которого объекты перемещаются между кластерами, что приводит к изменению координат центров кластеров (см. рис. 5).

Итерации продолжаются до тех пор, пока объекты не перестанут перебегать из одного кластера в другие. При этом достигает своего минимального значения оптимизируемый функционал – остаточная дисперсия, которая вычисляется как сумма квадратов отклонений координат объектов от центров своих кластеров.

И в этом случае имеется возможность сформулировать и проверить нулевую гипотезу, которая звучит так: в признаковом пространстве точки рассеяны так, что образуют единственный кластер [7]. Правда, в этом случае дело обстоит сложнее, чем во всех предыдущих, поскольку разрушить кластерную структуру, оставив в то же время в нетронутом виде связи между переменными, намного сложнее, чем просто разрушить связи. Простое перемешивание признаков здесь не подходит. Приходится вводить дополни-

тельное предположение о том, что точки в пространстве признаков имеют многомерное нормальное распределение, которое не всегда выполняется. Но если мы принимаем такое допущение, то далее обычным путем генерируем данные с разрушенной кластерной структурой, получая экспериментальным путем распределение значений оптимизируемого функционала в условиях нулевой гипотезы об отсутствии многокластерной структуры. По значимости нулевой гипотезы можно обоснованно сказать, наблюдается ли кластерная структура на нашей выборке. Более того, по наименьшей значимости можно даже определить наиболее вероятное количество кластеров.

Но даже возможность получить оптимальную кластерную структуру не позволяет говорить о том, что мы действительно можем извлечь из данных содержащуюся там структурную информацию. Трудность состоит в том, что методы кластерного анализа хорошо работают при небольшой размерности признакового пространства (2-3), а выбор наиболее информативного подпространства признаков превращается в неподъемную переборную задачу. К тому же появляется проблема сравнения результатов классификаций и выбора наилучшей из них.

Задача сравнения классификаций была поставлена при анализе совпадения классификаций (рис. 6), построенных на основе данных по типологии орудий среднепалеолитических индустрий Ближнего и Среднего Востока и Кавказа [3].

Для каждой классификации были выбраны наиболее значимые переменные, позволившие построить наиболее отчетливые группировки:

1) классификация, построенная методами *k*-средних в пространстве факторов 1, 2 (выделено 5 кластеров);

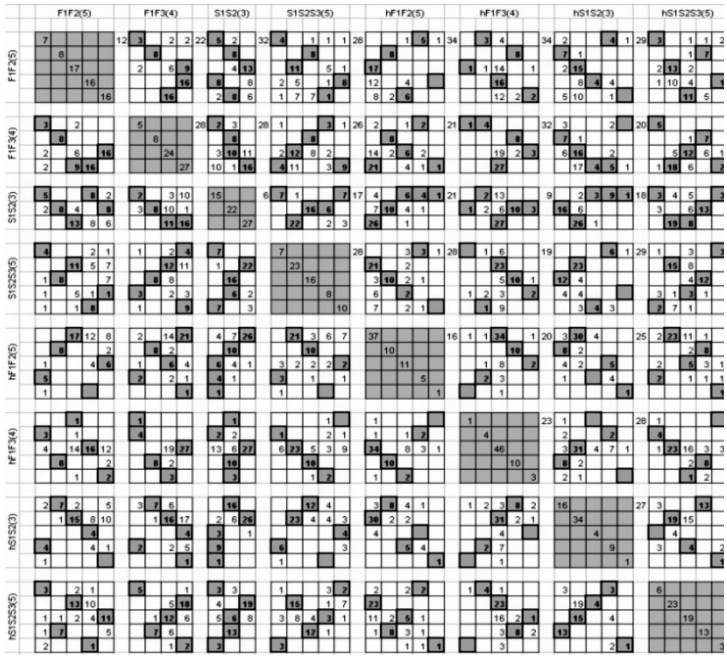

Рис. 6. Покластерное совпадение кластеров, построенных по разным классификациям

2) классификация, построенная методами *k*-средних в пространстве факторов 1, 3 (выделено 4 кластера);

3) классификация, построенная методами *k*-средних в пространстве шкал 1, 2 (выделено 3 кластера);

4) классификация, построенная методами *k*-средних в пространстве шкал 1–3 (выделено 5 кластеров);

5) классификация, построенная методами иерархического кластерного анализа в пространстве факторов 1, 2 (выделено 5 кластеров);

6) классификация, построенная методами иерархического кластерного анализа в пространстве факторов 1, 3 (выделено 4 кластера);

7) классификация, построенная методами иерархического кластерного анализа в пространстве шкал 1, 2 (выделено 3 кластера);

8) классификация, построенная методами иерархического кластерного анализа в пространстве шкал 1–3 (выделено 5 кластеров).

По этим классификациям построены попарные классификации для перечисленных разбиений. Результаты анализа совпадений кластеров при попарном сравнении состава их элементов приведены в табл. 3.

В.С. Костиным был предложен вариант решения, где в качестве статистики, измеряющей степень отклонения от независимости классификаций, выбрана максимальная доля совпадающих объектов при оптимальном соответствии кластеров [9]. А задача выбора наилучшей классификации была трансформирована в задачу объединения результатов большого количества независимо построенных классификаций [10] и построения на основе этого объединения классификации обобщенной. Наглядное представление обобщенной классификации предлагается на рис. 7. В клетках представленной таблицы указана степень согласованности включения объектов в одни и те же кластеры. Более темным оттенкам серого соответствует высокая степень согласованности, более светлым – менее высокая.

Оценивая весь оригинальный (не полностью описанный из-за ограничений на объемы) инструментарий, следует указать наиболее важную особенность методологии и методики, на которой он выстроен: все процедуры и методы завершаются обязательной проверкой статистической значимости полученных результатов.

Таблица 3

Совпадение кластеров, построенных по разным классификациям (в %)

Исходные классификации	Сопряженные классификации							
	1	2	3	4	5	6	7	8
1	100.0	81.3	65.6	50.0	56.3	46.9	46.9	54.7
2	81.3	100.0	56.3	56.3	59.4	67.2	50.0	68.8
3	65.6	56.3	100.0	90.6	73.4	67.2	85.9	71.9
4	50.0	56.3	90.6	100.0	56.3	56.3	70.3	54.7
5	56.3	59.4	73.4	56.3	100.0	75.0	68.8	60.9
6	46.9	67.2	67.2	56.3	75.0	100.0	64.1	56.3
7	46.9	50.0	85.9	70.3	68.8	64.1	100.0	57.8
8	54.7	68.8	71.9	54.7	60.9	56.3	57.8	100.0

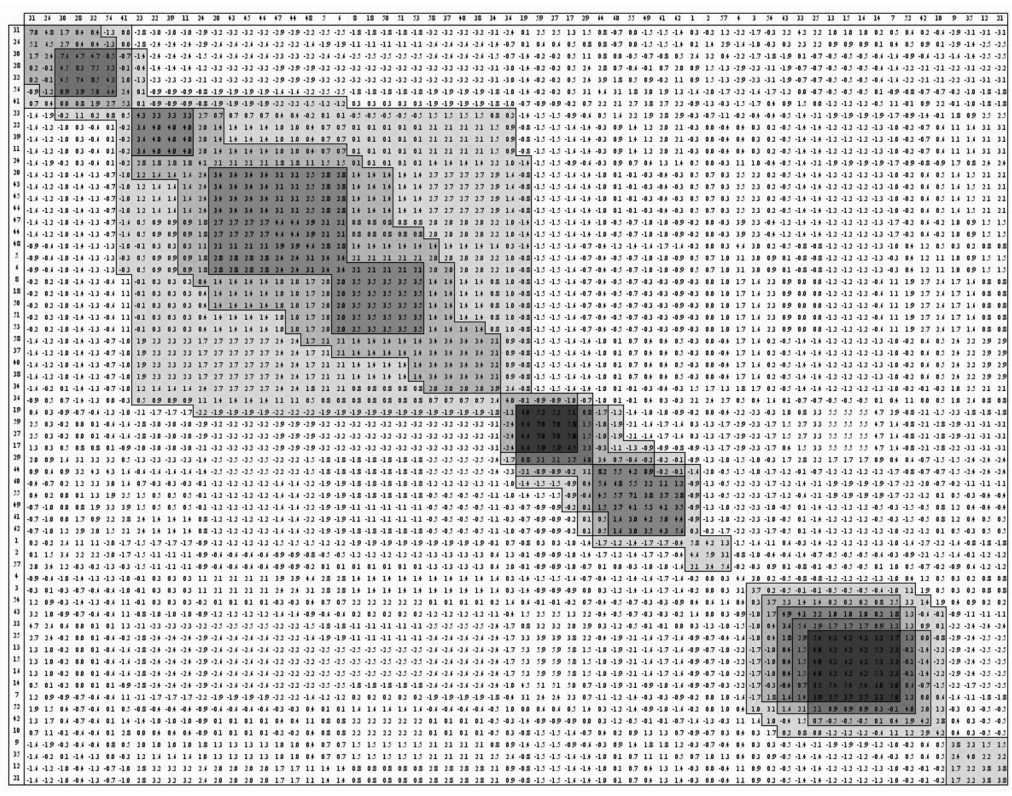

Рис. 7. Обобщенная классификация

Следует заметить, что непрерывное пополнение описанного нами выше инструментального арсенала для статистического анализа археологических данных продолжается. Дальнейшим шагом в этом направлении будет получение не одной, а нескольких обобщенных кластерных структур. Проблема состоит в том, что весь набор признаков, описывающий исследуемые объекты, как правило, отражает их с самых разных сторон, выделяя разные предметы исследования одного и того же объекта, разные уровни и формы организации и развития мира [1].

Решение следует искать в разделении всего набора предварительно проведенных аспектно-предметных классификаций на группы. Внутри каждой из подобных групп попарная близость классификаций друг к другу должна быть в среднем больше, чем близость классификаций из разных групп. Тогда на базе каждой из этих групп классификаций уже можно строить «чистую» обобщенную классификацию.

Литература

1. *Витяев Е.Е., Костин В.С.* Естественная классификация, систематика, онтология // Информационные технологии в гуманитарных исследованиях. Новосибирск: ИАЭТ СО РАН, 2009. Вып. 13. С. 65–75.
2. *Гарден Ж.-К.* Теоретическая археология. М.: Прогресс, 1983. С. 211.
3. *Деревянко А.П., Холюшкин Ю.П., Воронин В.Т., Ростовцев П.С.* Статистическое изучение мустьерских индустрий Кавказа и Ближнего Востока. Проблемы со-

поставимости // Информационные технологии в гуманитарных исследованиях. Новосибирск: РИЦ НГУ, 2003. Вып. 5.
4. *Деревянко А.П., Холюшкин Ю.П., Костин В.С., Воронин В.Т.* Структурный анализ орудийных комплексов Ближнего и Среднего Востока и Кавказа // Информационные технологии в гуманитарных исследованиях. Новосибирск: Редакционно-издательский центр НГУ, 2004. Вып. 7. С. 78–90.
5. *Деревянко А.П., Холюшкин Ю.П., Воронин В.Т., Костин В.С.* Статистическое исследование среднепалеолитических индустрий Средней Азии и Казахстана. Новосибирск, 2005. С. 45–46.
6. *Деревянко А.П., Холюшкин Ю.П., Воронин В.Т., Костин В.С.* Корреляция среднепалеолитических индустрий Ближнего Востока и Кавказа. Ч. 2. Типология. Новосибирск, 2005. 94 с.
7. *Жданов А.С., Костин В.С.* Значимость и устойчивость автоматической классификации в задаче поиска оптимального разбиения // Информационные технологии в гуманитарных исследованиях. Новосибирск: РИЦ НГУ, 2002. Вып. 3. С. 36–42.
8. *Клейн Л.С.* История археологической мысли. СПб., 2005.
9. *Костин В.С.* Статистика для сравнения классификаций // Информационные технологии в гуманитарных исследованиях. Новосибирск, 2003. Вып. 6. С. 57–65.
10. *Костин В.С., Корнюхин Ю.Г.* Построение обобщенной классификации // Информационные технологии в гуманитарных исследованиях. Новосибирск, 2003. Вып. 6. С. 65–72.
11. *Clarke D.L.* Analytical archaeology. L.: Methuen, 1968. 684 p.
12. *Clarke D.L.* Analytical archaeology. L.: Methuen, 1968. P. 188.
13. *Deetz J.* Invitation to Archaeology. N.Y., 1967. P. 26–37.
14. *Efron B.* Better bootstrap condidence intervals // J. American Statist. Association, 1986, 81.
15. *Kintigh K.* Measuring archaeological diversity by comparison with Simulated assemblages // American Antiquety. 1984. Vol. 49. P. 44–54.
16. *Marks A.E.* Typological Variability in the Levantine Middle Paleolithic // The Middle Paleolithic: Adaptation, Behavier and Variability. University Museum series. 1992. Vol. 2. P. 127–142.
17. *Ringrose T.* Bootstrapping and correspondenting analysis in archaeology // Journal of Archaeological Sciences. 1992. Vol. 19. P. 615–629.
18. *Trigger B.C.* Settlement archaeology – its goals and promise // American antiquity. 1967. Vol. 32. № 2. P. 26–31.

Bibliography

1. *Vitjaev E.E., Kostin V.S.* Estestvennaja klassifikacija, sistematika, ontologija // Informacionnye tehnologii v gumanitarnyh issledovanijah. Novosibirsk: IAJeT SO RAN, 2009. Vyp. 13. P. 65–75.
2. *Garden Zh.-K.* Teoreticheskaja arheologija. M.: Progress, 1983. P. 211.
3. *Derevjanko A.P., Holjushkin Ju.P., Voronin V.T., Rostovcev P.S.* Statisticheskoe izuchenie must'erskih industrij Kavkaza i Blizhnego Vostoka. Problemy sopostavimosti // Informacionnye tehnologii v gumanitarnyh issledovanijah. Novosibirsk: RIC NGU, 2003. Vyp. 5.
4. *Derevjanko A.P., Holjushkin Ju.P., Kostin V.S., Voronin V.T.* Strukturnyj analiz orudijnyh kompleksov Blizhnego i Srednego Vostoka i Kavkaza // Informacionnye tehnologii v gumanitarnyh issledovanijah. Novosibirsk: Redakcionno-izdatel'skij centr NGU, 2004. Vyp. 7. P. 78–90.
5. *Derevjanko A.P., Holjushkin Ju.P., Voronin V.T., Kostin V.S.* Statisticheskoe issledovanie srednepaleoliticheskih industrij Srednej Azii i Kazahstana. Novosibirsk, 2005. P. 45–46.

6. *Derevjanko A.P., Holjushkin Ju.P., Voronin V.T., Kostin V.S*. Korreljacija srednepaleoliticheskih industrij Blizhnego Vostoka i Kavkaza. Ch. 2. Tipologija. Novosibirsk, 2005. 94 p.
7. *Zhdanov A.S., Kostin V.S*. Znachimost' i ustojchivost' avtomaticheskoj klassifikacii v zadache poiska optimal'nogo razbienija // Informacionnye tehnologii v gumanitarnyh issledovanijah. Novosibirsk: RIC NGU, 2002. Vyp. 3. P. 36–42.
8. *Klejn L.S*. Istorija arheologicheskoj mysli. SPb., 2005.
9. *Kostin V.S*. Statistika dlja sravnenija klassifikacij // Informacionnye tehnologii v gumanitarnyh issledovanijah. Novosibirsk, 2003. Vyp. 6. P. 57–65.
10. *Kostin V.S., Kornjuhin Ju.G*. Postroenie obobshhennoj klassifikacii // Informacionnye tehnologii v gumanitarnyh issledovanijah. Novosibirsk, 2003. Vyp. 6. P. 65–72.
11. *Clarke D.L*. Analytical archaeology. L.: Methuen, 1968. 684 p.
12. *Clarke D.L*. Analytical archaeology. L.: Methuen, 1968. P. 188.
13. *Deetz J*. Invitation to Archaeology. N.Y., 1967. P. 26–37.
14. *Efron B*. Better bootstrap condidence intervals // J. American Statist. Association, 1986, 81.
15. *Kintigh K*. Measuring archaeological diversity by comparison with Simulated assemblages // American Antiquety. 1984. Vol. 49. P. 44–54.
16. *Marks A.E*. Typological Variability in the Levantine Middle Paleolithic // The Middle Paleolithic: Adaptation, Behavier and Variability. University Museum series. 1992. Vol. 2. P. 127–142.
17. *Ringrose T*. Bootstrapping and correspondenting analysis in archaeology // Journal of Archaeological Sciences. 1992. Vol. 19. P. 615–629.
18. *Trigger B.C*. Settlement archaeology – its goals and promise // American antiquity. 1967. Vol. 32. № 2. P. 26–31.