УДК 902

В. Т. Воронин [1, 2], **В. С. Костин** [2], **Ю. П. Холюшкин** [2]

[1] Новосибирский государственный университет
ул. Пирогова, 2, Новосибирск, 630090, Россия;
[2] Институт археологии и этнографии СО РАН
пр. Акад. Лаврентьева, 17, 630090, Россия
E-mail: hol@archaeology.nsc.ru

ОН-ЛАЙН СТАТИСТИЧЕСКИЙ АНАЛИЗ ДАННЫХ В ОБУЧЕНИИ СТУДЕНТОВ-АРХЕОЛОГОВ

Повышение качества фундаментального археологического образования требует не только установки на приоритетное освоение интеллектуальной составляющей социальных и культурных явлений прошлого, но и овладение в совершенстве современными информационными технологиями. Такая постановка вопроса вызывает необходимость разработки и внедрения современных общеобразовательных обучающих программ и создание более широких возможностей для индивидуализации обучения студентов, в том числе за счет использования дистанционных форм обучения. В статье обсуждается адекватный вариант решения этой задачи – создание специального Web-сервера, включающего подсистемы ввода/хранения данных, методов статистического анализа данных и представления результатов.

Ключевые слова: статистический пакет, статистический анализ данных, Интернет, иерархический кластерный анализ, метод k-средних, дисперсионный анализ, критерий хи-квадрат.

Повышение качества фундаментального археологического образования настоятельно требует не только установки на приоритетное освоение интеллектуальной составляющей социальных и культурных явлений прошлого, но и овладения в совершенстве современными информационными технологиями. Такая постановка вопроса вызывает необходимость разработки и внедрения современных общеобразовательных обучающих программ и создание более широких возможностей для индивидуализации обучения студентов, в том числе за счет использования дистанционных форм обучения.

За более чем десятилетнюю историю в секторе археологической теории и информатики ИАЭТ СО РАН было разработано несколько оригинальных методов анализа данных, учитывающих специфические особенности археологической информации [Холюшкин и др., 2007. С. 254–259]. Применение этих методов отражено в многочисленных статьях и монографиях, написанных сотрудниками сектора. К сожалению, программные реализации этих методов сделаны в разные годы и представляют собой разношерстный «зоопарк». Первые программы написаны на языке C++ и в качестве входных данных принимают устаревший формат файлов SPSS / PC+, вышедший из употребления с переходом пользователей с MS DOS на Windows. Последние разработки сделаны в виде макросов Excel, что позволило обойтись без программирования интерфейса пользователя, поскольку сама электронная таблица является интерфейсом для ввода табличных данных и вывода результатов анализа в виде диаграмм деловой графики. Как известно, программирование пользовательского и графического интерфейса, а также интерфейса к данным при разработке программного продукта, занимает 90 % времени. Когда мы создаем программы «для внутреннего пользования», разумно ограничиться реализацией только расчетной части, на порядок сократив затраты времени на программирование. Единственное неудобство сложившегося «зоопарка» в том, что применение каждого метода анализа связано с поиском необходимой программы и восстановлением в памяти всей последовательности действий.

В последнее время появилась мысль объединить все разрозненные программные разработки в едином пакете и сделать их доступным более широкому кругу пользователей.

ISSN 1818-7919
Вестник НГУ. Серия: История, филология. 2008. Том 7, выпуск 3: Археология и этнография

Наиболее адекватным решением этой задачи, на наш взгляд, является создание Web-сервера, включающего подсистемы ввода/хранения данных, методов анализа данных и представления результатов. Для расширения возможностей сервера предполагается использовать также отдельные модули пакета PSPP, аналогичного по функциональности SPSS, но распространяемого бесплатно по соглашению «Open source», которое предусматривает возможность использовать и модифицировать исходный код программы для включения в некоммерческие программные продукты.

Первый год работы потребовал значительных затрат на выявление аналогов предлагаемой разработки. Выяснилось, что все существующие статистические системы, а их более 1000, не являются Online-проектами. Все они – локальные приложения, к которым нет доступа из сети Интернет. Кроме того, статистические программы относятся к наукоемкому программному обеспечению, цена которых часто недоступна индивидуальному пользователю. Профессиональные пакеты имеют большое количество методов анализа, а популярные пакеты лишь какое-то количество функций, достаточных для универсального применения. Специализированные же пакеты ориентированы на какую-либо узкую область анализа данных.

Создатели программных статистических пакетов часто заявляют, что их продукт превосходит аналоги. Отсутствие у большинства исследователей времени и должного образовательного уровня для освоения нескольких программ делает непростым их выбор. Приведем несколько самых распространенных систем статистического анализа.

MS Excel является самой упоминаемой (и используемой) в археологической практике и отечественных статьях пакетов офисной программой компании Microsoft. Причины этого кроются в широком распространении данного программного обеспечения, наличии русскоязычной версии, тесной интеграцией с MS Word и PowerPoint. MS Excel – это электронная таблица с достаточно мощными математическими возможностями, где некоторые статистические функции являются просто дополнительными встроенными формулами. Однако расчеты, сделанные при ее помощи, не признаются авторитетными биомедицинскими журналами. Также в MS Excel невозможно построить качественные научные графики. Безусловно, MS Excel хорошо подходит для накопления данных, промежуточного преобразования, предварительных статистических прикидок, для построения некоторых видов диаграмм. Однако окончательный статистический анализ необходимо делать в программах, которые специально созданы для этих целей.

SPSS (Statistical Package for Social Science) – наиболее часто используемый пакет статистической обработки данных с более чем 30-летней историей. Отличается гибкостью, мощностью и применим для всех видов статистических расчетов в социологии и биомедицине. Недавно вышла 13-я англоязычная версия. Существует русскоязычное представительство компании, предлагающее полностью русифицированную версию SPSS 12.0.2 для Windows. Однако стоимость ее лицензии на два года еще недавно составляла 15 000 долларов.

Производителем программы STATISTICA является фирма StatSoft Inc. (США), которая выпускает статистические приложения, начиная с 1985 г. STATISTICA включает большое количество методов статистического анализа (более 250 встроенных функций), объединенных следующими специализированными статистическими модулями: основные статистики и таблицы; непараметрическая статистика; дисперсионный анализ; множественная регрессия; нелинейное оценивание; анализ временных рядов и прогнозирование; кластерный анализ; факторный анализ; дискриминантный функциональный анализ; анализ длительностей жизни; каноническая корреляция; многомерное шкалирование; моделирование структурными уравнениями и др.

Несложный в освоении, данный статистический пакет может быть рекомендован для археологических, социологических и биомедицинских исследований любой сложности. Но в связи с широкими возможностями таких систем у них имеются существенные для археологов недостатки, сложности для быстрого освоения и использования. Это малое количество экранных подсказок и поэтому требование наличия у пользователей профессиональных навыков и высокой квалификации, широкого первоначального статистического образования,

доступной литературы и консультационных служб, внимательного изучения документации на английском языке; отсутствие подробной документации, доступной для начинающих и информативной для специалистов-статистиков (исключение SPSS); необходимость больших финансовых затрат (профессиональные статистические пакеты обычно стоят от 1 до 10 тыс. долларов и более).

Поэтому на начальном этапе были сформулированы основные принципы построения нашей системы и вытекающие из них требования:

1. Поскольку основными пользователями системы являются археологи, имеющие собственные данные или гипотезы, которые могут быть проверены на уже опубликованных данных, сервер должен предоставлять пользователю:

а) возможность вводить, сохранять и редактировать собственные данные, которые должны сохраняться на сервере как в течение сеанса работы, так и (для зарегистрированного пользователя) между сеансами (кроме того, должна существовать возможность сохранения данных на стороне клиента, то есть экспорт / импорт);

б) доступ к архиву обобществленных данных (со ссылками на источник); из общего архива данные могут быть скопированы в личный архив, после чего их можно редактировать и сохранять; необходима также функция пополнения общего архива, хотя она может быть реализована не полностью автоматически, а с участием человека, администратора данных.

2. Пользователи могут хорошо ориентироваться в собственных данных, но не иметь специальной подготовки в методах статистического анализа; в связи с этим:

а) пользователю нужна не только возможность применения методов анализа, но и, прежде всего, объяснение, в каких случаях их применять, какие выводы можно сделать из полученных результатов, т. е. необходима большая обучающая подсистема, включая тексты с описанием методов и, обязательно, примеры анализа реальных данных;

б) пользователям, которых шаблонные интерпретации методов не устраивают, следует предложить самим разобраться в их сущности, предоставив пошаговую демонстрацию производимых преобразований на живых данных; при этом каждый шаг должен быть снабжен описанием, достаточным для понимания и самостоятельного воспроизведения.

3. В своей работе научные сотрудники применяют методы анализа данных не только в процессе исследования, но и для подготовки научных публикаций; в связи с этим:

а) вместе с выводом результатов анализа, система должна предлагать и формулировки выводов, достаточно строгие, точные, хорошо интерпретируемые и, самое главное, понятные исследователю;

б) все графические иллюстрации должны быть информативными, выполненными на хорошем дизайнерском уровне, и отвечать всем требованиям к рисункам в научных публикациях.

Данная статья призвана осветить вопросы, связанные с реализацией первой версии статистического пакета для археологов. В процессе выполнения указанной задачи были созданы следующие возможности пакета: метод анализа связей (дисперсионный анализ, коэффициент корреляции Пирсона, Хи-квадрат Пирсона); два метода анализа структуры (метод «k-средних» и агломеративный метод иерархической классификации); различные виды отображения результатов анализа (HTML-таблица и VRML-диаграмма).

Целью анализа связей является обнаружение взаимосвязей зависимостей между признаками. В статистике разобрано множество критериев для проверки связей в базах 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 становится не чем иным, как вероятностью наблюдения получен-

ного значения этой статистики при выполнении нулевой гипотезы. Если эта вероятность ниже заранее выбранного порога, например 5 %, то исследователь имеет основания утверждать, что нулевая гипотеза не подтверждается на его данных, из чего с большей вероятностью следует вывод об обнаружении между признаками определенной связи. 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.

Целью дисперсионного анализа является проверка значимости различия между средними в разных группах с помощью сравнения дисперсий этих групп. Разделение общей дисперсии на несколько источников позволяет сравнить дисперсию, вызванную различием между группами, с дисперсией, вызванной внутригрупповой изменчивостью. Проверяемая гипотеза состоит в том, что различия между группами нет. При истинности нулевой гипотезы, оценка дисперсии, связанной с внутригрупповой изменчивостью, должна быть близкой к оценке межгрупповой дисперсии, при ложности – значимо отклоняться.

Корреляция представляет собой меру зависимости переменных. Наиболее известна корреляция Пирсона. При ее выполнении предполагается, что переменные измерены, как минимум, в интервальной шкале (эта шкала измерений позволяет не только упорядочить наблюдения, но и количественно выразить расстояния между ними, причем на шкале не обязательно присутствует *абсолютная* нулевая отметка). Некоторые иные коэффициенты корреляции могут быть вычислены для менее информативных шкал. Коэффициенты корреляции изменяются в пределах от –1.00 до +1.00. Следует обратить внимание на крайние значения коэффициента корреляции. Значение –1.00 означает, что переменные имеют строгую отрицательную корреляцию. Значение +1.00 означает, что переменные имеют строгую положительную корреляцию. Значение 0.00 означает отсутствие корреляции.

Отрицательная корреляция означает две переменные, могущие быть связаны таким образом, что при возрастании значений одной из них значения другой убывают. Это и показывает отрицательный коэффициент корреляции. Про такие переменные говорят, что они отрицательно коррелированны.

Положительная корреляция – это ситуация, когда связь между двумя переменными может быть следующей: значения одной переменной возрастают, значения другой переменной также возрастают. Это и показывает положительный коэффициент корреляции. Про такие переменные говорят, что они положительно коррелированны.

Хи-квадрат Пирсона – это наиболее простой критерий проверки значимости связи между двумя категоризованными переменными. Критерий Пирсона основывается на том, что в двухвходовой таблице ожидаемые частоты при гипотезе «между переменными нет зависимости» можно вычислить непосредственно. Значение статистики «хи-квадрат» и ее уровень значимости зависят от общего числа наблюдений и количества ячеек в таблице, относительно малые отклонения наблюдаемых частот от ожидаемых будут доказывать значимость, если число наблюдений велико. Имеется только одно существенное ограничение использования критерия хи-квадрат (кроме очевидного предположения о случайном выборе наблюдений), которое состоит в том, что ожидаемые частоты не должны быть очень малы.

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

В создаваемом пакете реализованы два метода анализа структур: иерархический кластерный анализ и метод k-средних.

Дополнением к статпакету является VRML-графика, которая обеспечивает возможность создания некой комплексной модели, состоящей из трехмерных и двумерных объектов, а также звука и прочей мультимедиа информации Пример вывода результата анализа методом «k-средних» в 3-d диаграмме: см. рисунок. При этом двумерные объекты могут быть как

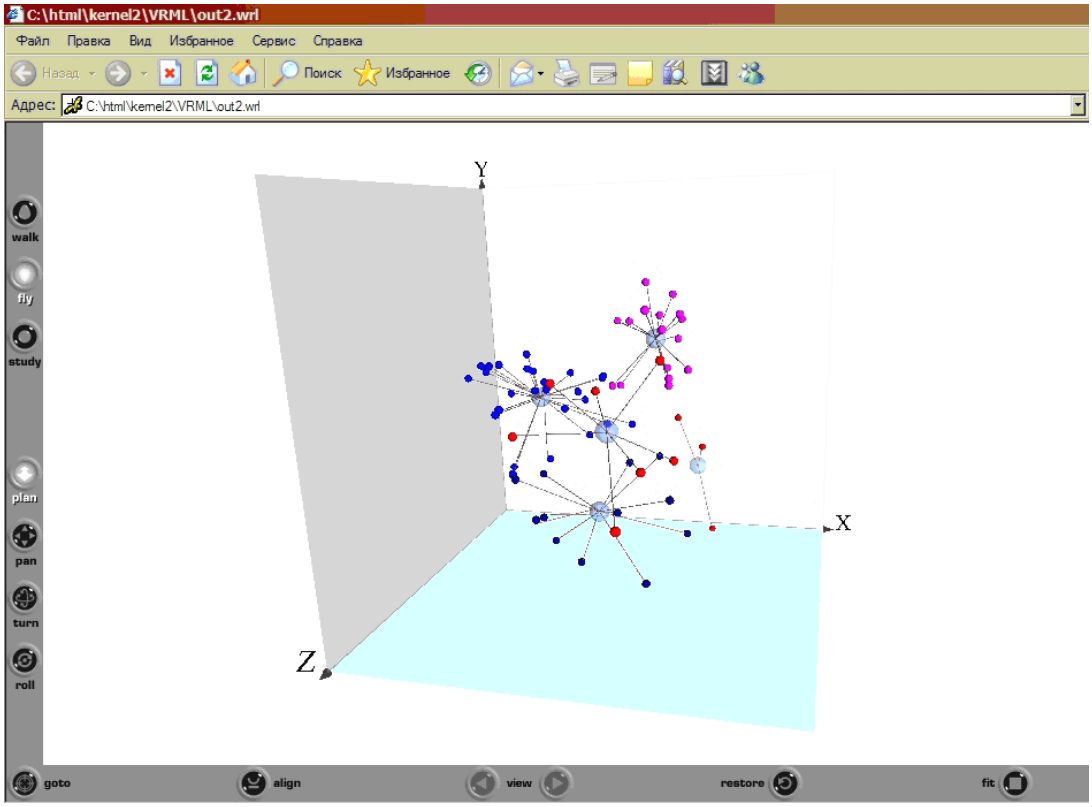

Рис. 1. Пример вывода результата анализа методом «k-средних» в 3-d диаграмме

текстовыми, так и иллюстрациями и фотографиями. При всем этом VRML не предъявляет никаких особых требований к компьютерной платформе или программному обеспечению. Как стандарт, разрабатываемый специально для Интернета, VRML открывает широкие возможности в сфере представления и передачи информации благодаря малым размерам исходных VRML-документов и высокой информативности трехмерных образов. Возможное внедрение интерактивности делает эти представления более реалистичными и удобными в понимании и освоении демонстрируемого объекта. Именно этот аспект выдвигает VRML в категорию средств предоставления доступа к информации, более полно описывающей реальные объекты. Бесспорным преимуществом технологии VRML является возможность ее сетевого использования в силу сравнительно небольшого объема передаваемых данных. Недостатком метода VRML в настоящее время является необходимость передачи на компьютер пользователя всех необходимых для визуализации файлов перед ее началом.

Таким образом, он-лайн статистический анализ данных, включающий подсистемы ввода / хранения данных, методы анализа и представления результатов, вполне может быть внедрен в образовательный процесс (включая дистанционные формы обучения) по подготовке высококвалифицированных специалистов-археологов, способных эффективно применять на практике современные информационные технологии.

Список литературы

Холюшкин Ю. П., Воронин В. Т., Костин В. С. О комплексной обработке археологических данных методами математической статистики // Северная Евразия в антропогене: человек, палеотехнологии, палеоэкология, этнология и антропология: Материалы международной конференции. Иркутск, 2007. Т. 2. С. 254–259.

Материал ***поступил в редколлегию 12.10.2007***

V. T. Voronin, V. S. Kostin, Yu. P. Kholjushkin

ON-LINE STATISTICAL ANALYSIS OF DATA IN THE TEACHING OF STUDENTS-ARCHAEOLOGISTS

Improvement of quality of fundamental archaeological education demands not only aiming at priority development of an intellectual component of the social and cultural phenomena of the past, but also mastering in perfection modern information technologies. Such statement of a question demands development and introduction of modern general educational training programs and creation more ample opportunities for an individualization of training of students, including due to use of remote forms of training. In our paper is discussed adequate decision of this task: creation of the special Web-server including subsystems of an input/data storage, methods of the statistical analysis of the data and representation of results.

Keywords: statistical package, statistical analysis of data, Internet, hierarchical cluster analysis, k-means method, dispersing analysis, chi-square