С.Н. МАРТЫШЕНКО,
Н.С. МАРТЫШЕНКО,
Д.А. КУСТОВ

Совершенствование математического и программного обеспечения обработки первичных данных в экономических и социологических исследованиях

Излагаются принципы и структура разработанного комплекса программных средств анализа анкетных данных, встраиваемого в EXCEL. Программный комплекс включает средства анализа, не входящие в распространенные пакеты статистической обработки информации.

Сбор первичной информации и ее анализ – два этапа маркетинговых и социологических исследований, в которых опрос является основным методом. Он используется более чем в 90% исследований [7]. В нашей работе мы ориентировались в большей степени на маркетинговые исследования, но поскольку с методологической точки зрения эти два вида исследований имеют много общего и очень часто пересекаются, как возможную область применения мы рассматриваем и социологические. Кроме того, в социологии в настоящее время данная проблема проработана гораздо лучше. Об этом свидетельствуют многочисленные научные публикации отечественных [3, 5, 8–10] и зарубежных [7, 11] ученых.

В маркетинговых исследованиях применяются несколько видов опросов: опрос экспертных групп, опрос методом фокус-групп, глубинное интервью, анкетирование [7]. Первые три могут рассматриваться как предварительные этапы, предшествующие анкетному опросу. На этих этапах проводятся качественные исследования, а результатом анкетного опроса являются количественные данные (табл. 1). Качественные исследования позволяют понять суть ситуации, сложившейся вокруг изучаемой проблемы, и сформулировать гипотезы, которые могут быть проверены в ходе анкетного опроса.

Данная работа посвящена совершенствованию методов и технологий обработки количественной информации, поэтому далее мы будем рассматривать только системы сбора первичной информации, основанные на проведении анкетных опросов.

Таблица 1

Характеристики качественного и количественного исследования

Элементы исследования	Качественные исследования	Количественные исследования
Цель	Определение качественного понимания скрытых мотивов и побуждений	Представление данных в количественной форме и обобщение результатов анализа данных на всю генеральную совокупность
Выборка	Небольшое количество объектов	Большое количество объектов
Сбор информации	Неструктурированный	Структурированный
Анализ информации	Нестатистический	Статистический
Результат	Начальное представление и формулировка гипотез	Рекомендации для принятия управленческого решения

В России анкетные опросы пока не приобрели столь массового характера, как в странах с развитой рыночной экономикой. Большая часть анкетных опросов до последнего времени проводится в научных целях. Можно привести несколько причин, по которым в нашей стране использование этого метода в маркетинговых исследованиях не получило большого распространения:

- нежелание большинства респондентов идти на контакт и предоставлять информацию. Если в развитых странах население в большинстве своем считает, что предоставленная ими информация может положительно отразиться на их жизни, то у нас недоверие к любым контактам с незнакомыми людьми преодолеть очень трудно. Сохраняющаяся в нашей стране высокая криминальная напряженность также не способствует установлению контактов. По этим причинам сложно организовать опросы по почте или телефону. В ближайшие годы основным методом останется личный опрос с заполнением анкетных форм на бумажном носителе, будут развиваться и электронные опросы [7];
- маркетинговая деятельность в нашей стране пока недостаточно развита. Чтобы получить действительно полезную для бизнеса информацию в ходе опросов, необходим большой опыт работы. Сегодня многим российским предприятиям не по средствам организовывать собственные маркетинговые исследования, а тем более содержать подразделение, работающее на перспективу. В странах с развитой рыночной экономикой предприятие, которое не в состоянии организовать собственные маркетинговые исследования, всегда может обратиться за помощью в многочисленные специализированные фирмы. В России таких фирм пока очень мало;
- большинство маркетологов не владеет в достаточной степени необходимыми математическими знаниями и определенными навыками для обработки статистических данных. Отечественных программных средств, пригодных для этих целей, недостаточно. Нелицензированные зарубежные пакеты по статистике весьма доступны, но большинство из

них не имеет документации на русском языке, и даже профессионал далеко не всегда может в них разобраться;

- отсутствуют предназначенные не для научных, а для практических работников, доступные и понятные широкому кругу маркетологов методики и технологии компьютерной обработки анкетных данных;
- данные анкетных опросов имеют свою специфику, которая не всегда укладывается в рамки классической статистики [2, 3]: они в большинстве своем носят нечисловой характер, либо являются разнотипными. Наиболее распространенные пакеты по статистике в основном рассчитаны на данные, измеренные в одной относительной шкале;
- известный специалист в области анализа анкетных данных Ю.Н. Толстова отмечает, что отечественных публикаций по этому вопросу очень мало, в то время как за рубежом этой проблеме уделяется пристальное внимание. Но проблема не может быть решена только с помощью перевода на русский язык иностранной литературы [10].

Теория обработки нечисловых данных находится в стадии развития и в последние годы привлекает внимание ученых-статистиков, но проводимые исследования носят скорее научный, чем практический характер [3, 9, 10].

Настоящая работа имеет целью обеспечить широкий круг практиков дополнительными средствами анализа анкетных данных, учитывающими специфику решаемых задач. Предлагаемый комплекс компьютерных программ – это не просто набор независимых функций обработки данных, а инструмент для проведения системных исследований, т.е. разрабатываемые средства предназначаются не столько для тех, кто использует анкетный опрос для решения частной задачи, сколько для тех, кто проводит или собирается проводить анкетные опросы на профессиональной основе.

Профессиональное использование анкетных опросов предполагает проведение не одного опроса, а множества. Один опрос является только базой для дальнейших исследований. В специальной литературе можно найти рекомендации по проведению опросов [7, 1], но без накопления собственного опыта невозможно получить качественные результаты и избежать погрешностей, т.к. перечень их возможных источников слишком широк. На рис. 1 представлены компоненты общей ошибки анкетного опроса. Подробный анализ ошибок можно найти в работе Н.К. Малхотры [7].

Предлагаемая нами система анализа данных разрабатывалась для обслуживания крупномасштабных исследований, когда в опросах участвуют тысячи респондентов (анкеты иногда содержат сто и более вопросов). Для обработки таких анкет даже простые операции могут превратиться в проблему. В этой ситуации особенно важна технологичность выполнения любой операции. Особые требования предъявляются к работе программного обеспечения: ставится задача получить не просто результат, а результат в рамках реального времени.

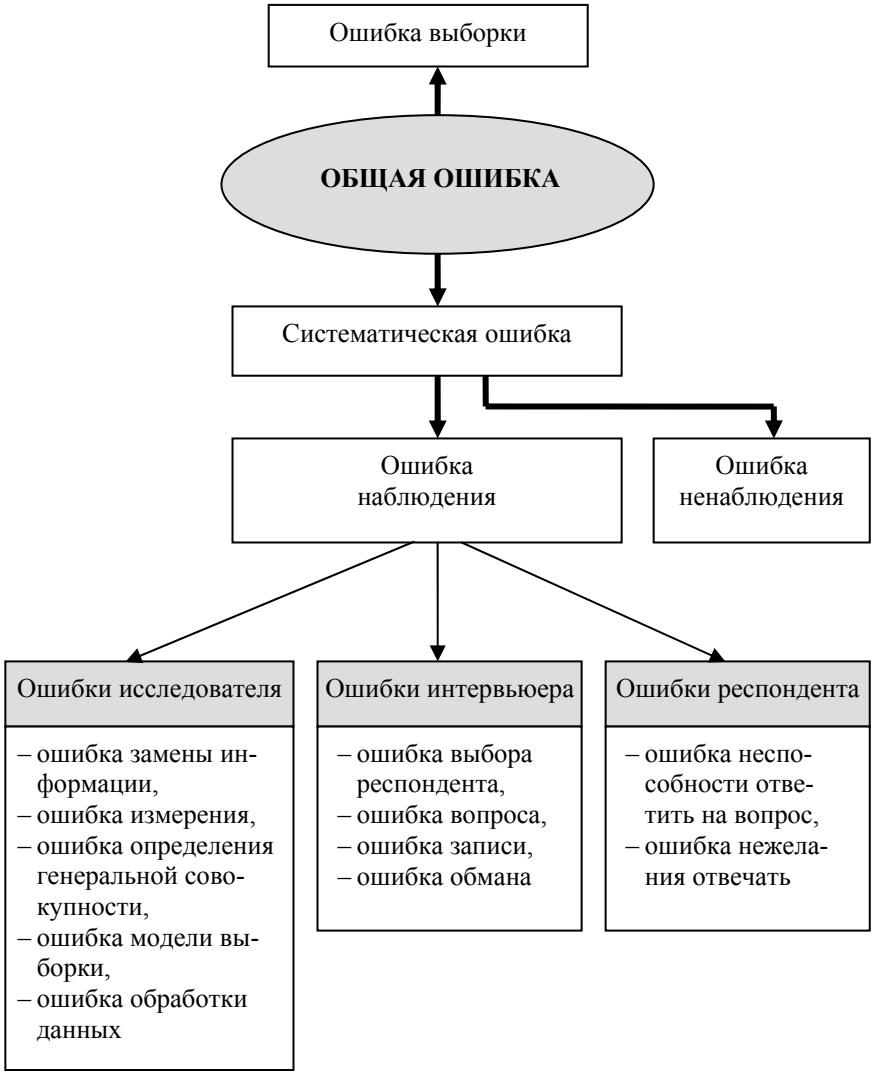

Рис. 1. Источники ошибок при проведении анкетного опроса

Характерная особенность масштабных исследований – необходимость привлекать большое количество интервьюеров. Как правило, это студенты, внештатные сотрудники, часто неквалифицированные, не всегда ответственно относящиеся к порученной им работе. Между тем многие авторы отмечают, что необходимо обращать внимание на личность интервьюера [6, 10]. Зарубежные исследователи, несмотря на удорожание услуг по сбору данных, постоянно повышают уровень требований к лицам, проводящим опросы. В этих условиях особое внимание должно быть уделено анализу достоверности информации, предоставляемой каждым отдельным интервьюером. Для решения этой задачи нужны специальные методы и программные средства, которые ни в один известный пакет по обработке данных не входят. Обработка данных может длиться

месяц и более. Большой объем результатов обработки требует систематизированного хранения.

Работа по организации маркетинговых исследований на основе анкетных опросов должна быть построена с соблюдением принципов системного подхода [4]. В этом случае систему маркетинговых исследований удобно рассматривать как совокупность двух подсистем: сбора первичных данных и анализа данных.

Анкета – один из основных элементов системы сбора первичных данных, в которую кроме методов составления анкет входят методы организации опросов, методы стимулирования респондентов и интервьюеров, методы кодирования и компьютерного представления данных.

Повышение эффективности обработки анкетной информации может быть достигнуто за счет использования новой компьютерной технологии обработки данных, действенность которой во многом определяется тем, насколько в ней выдержаны основные принципы системного подхода. Рассмотрим основные элементы компьютерной технологии с точки зрения соответствия этим принципам.

Любой компьютерный комплекс программ разрабатывается с целью повышения эффективности решения определенного класса задач. Совокупность методов и алгоритмов обработки информации еще не образует компьютерную технологию, при создании которой большое значение имеет проблема согласования элементов технологии по способу представления и передачи информации при переходе от одного этапа обработки к другому.

Рассмотрение компьютерной технологии анализа анкетных данных начнем с укрупненных блоков (рис. 2), исследование их взаимосвязи позволит определить логику построения или применения каждого из них.

Система задач, решаемых на основе материалов анкетных опросов, определяет систему сбора данных. Совместно они обусловливают требования к средствам обработки данных, но последние тоже оказывают влияние на систему сбора.

Специализированный комплекс программных средств, являющийся основным элементом новой компьютерной технологии, не решает всех задач, которые могут возникнуть при анализе анкетных данных, а только обеспечивает дополнительные возможности для этого.

Новый комплекс программных средств создается на базе существующего универсального программного средства, используемого для анализа данных. В качестве такой основы для встраиваемого комплекса выбрана программная среда EXCEL, что объясняется в первую очередь ее широким распространением среди пользователей. Кроме того, EXCEL постоянно пополняется новыми статистическими функциями и обеспечена доступной документацией на русском языке.

В то же время EXCEL содержит далеко не все методы анализа статистических данных. Более полно средства статистического анализа представлены в специализированных пакетах статистической обработки данных, таких как SPSS, STATISTICA, STATGRAPHICS, STADIA и др. Трудно отдать предпочтение тому или иному пакету, однако их исполь-

зование дает возможность решать многие дополнительные задачи, которые сегодня нельзя решить в EXCEL. Но и эти пакеты в силу своей универсальности не позволяют справиться с рядом специфических проблем, возникающих при обработке анкетных данных.

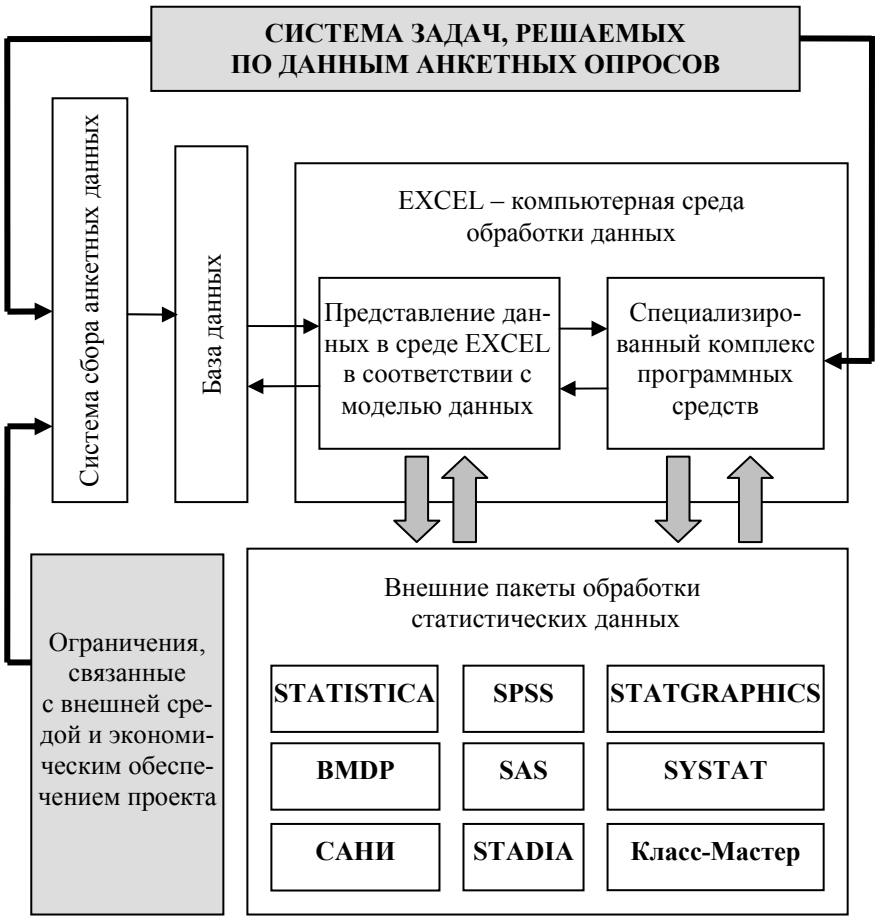

Рис. 2. Схема взаимосвязи основных блоков компьютерной технологии анализа анкетных данных

Специализированный комплекс повышает эффективность обработки данных с помощью программных средств, представленных в EXCEL, и облегчает взаимодействие с внешними пакетами статистического анализа. Его нельзя рассматривать только как набор некоторых дополнительных функций или инструментальных средств анализа данных в EXCEL. Эти средства связаны не только со средой, но и между собой некоторой общей концепцией.

Ряд средств специализированного комплекса предполагает соблюдение дополнительных условий представления данных, не предусмотренных в EXCEL. Эти условия связаны с понятиями «проект анкетного опроса» и «модель данных», которые в явной или неявной форме присущи любому специализированному пакету программных средств.

Подчинение данных некоторым дополнительным условиям позволяет существенно упростить обращение к программным средствам специализированного комплекса. Таким образом, если соблюдаются определенные правила в представлении данных, их не надо описывать при каждом обращении к программам, входящим в комплекс. С другой стороны, количество условий должно быть не очень велико, и они должны быть достаточно простыми и понятными, иначе будет затруднено пользование программами и согласование с программами среды и внешними пакетами.

Уточним смысл, вкладываемый в понятия «проект» и «модель данных» анкетного опроса при разработке специализированного программного комплекса.

Сбор анкетных данных, а затем их обработка могут проводиться достаточно длительно – от полугода до года, а при повторении опросов и несколько лет. Исходные данные и результаты, полученные на множестве различных этапов обработки, хранятся на нескольких страницах в одном файле EXCEL. Такой файл ассоциируется с проектом по анализу данных конкретного анкетного опроса. Одновременно один исследователь может сопровождать большое количество анкетных опросов и, соответственно, разрабатывать множество проектов.

В процессе обработки в данных могут происходить изменения (корректировка, пополнение выборки), что приводит к потере актуальности некоторых результатов (таблиц, признаков классификации). Определенная их часть пересчитывается автоматически, однако некоторые из них, например, результаты, полученные с помощью внешних программ, требуют пересчета (актуализации). Поэтому исследователю часто необходимо знать, когда и в какой последовательности он производил те или иные действия. В этом случае компьютерная программа будет представлять результаты по сопровождению проекта в систематизированном виде. Во избежание многих ошибок этот процесс в большой степени может быть автоматизирован.

Исследователю часто приходится разрабатывать некоторые новые инструментальные средства анализа информации, учитывающие специфику данных конкретного проекта. Все функции по сопровождению проекта может взять на себя специальная программа – мастер сопровождения.

Если исследователь работает с небольшими проектами, разрабатываемыми в сжатые сроки, ему нет необходимости применять все функции мастера, т.е. мастер предоставляет дополнительные возможности по обслуживанию, но использовать их необязательно.

Автоматическое сопровождение основано на определенной структуре проекта, под которой понимается создание определенного перечня элементов, оформленных в виде отдельных листов EXCEL. По мере развития проекта эти элементы заполняются информацией, которая может понадобиться исследователю в ходе работы. Структура проекта представлена на рис. 3.

Большинство параметров являются необязательными (некоторые из них необходимы только для использования определенных функций),

однако без установки обязательных параметров программный комплекс работать не будет.

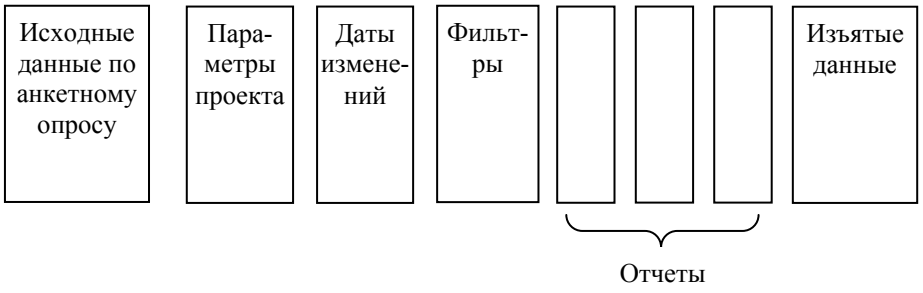

Рис. 3. Основные элементы структуры проекта

Рассмотрим параметры и содержание элементов проекта. На его первом листе должны быть размещены исходные данные анкетного опроса, оформленные в соответствии с моделью данных (она будет рассмотрена ниже). После ввода или экспорта исходных данных из внешней базы данных пользователь должен создать проект (функция СП). Программа выдаст запрос определить имя проекта и указать лист EXCEL, на котором размещена исходная информация. Это обязательные параметры (они выделяются красным цветом).

После инициализации проекта программа создаст следующие три элемента и внесет в них начальные сведения о проекте.

Рассмотрим элемент «Параметры проекта», включающий 4 раздела. Первый раздел «Заголовок проекта» расположен в верхней части листа. В нем содержатся следующие сведения: название проекта, число признаков в таблице данных анкетного опроса, количество анкет в выборке, дата создания проекта, дата последней корректировки данных.

Следующий раздел содержит характеристики данных (табл. 2.). При создании проекта в таблице автоматически заполняются первый, второй и четвертый столбец. В пятом столбце в исходном состоянии для всех строк установлен знак пропуска «–» . Остальную информацию должен ввести пользователь.

Таблица 2

Характеристики данных

Название признака	Шкала измерения	Представление	Признак отсутствия данных	Частота отсутствия данных	Зависимости
Признак 1	Номинальная	Числовое	-1	100	
Признак 2	– " –	Текстовое	Н/д	300	
Признак 3	– " –	Числовое	–	–	
Признак 4	Относительная	– " –	-1	–	

Признак отсутствия данных формализован в рамках пакета вполне конкретно. Для данных в числовом представлении признаком их отсутствия служит значение «-1». Его уникальность основана на том, что в анке-

тах оно принимает только положительные значения. Для текстовых данных признаком отсутствия служит обозначение «Н/д». Для тех признаков, значение которых невозможно определить, в графе проставляется символ «–». Если признак отсутствия данных определен, то частота встречаемости значения «отсутствие данных» рассчитывается автоматически. Столбец зависимости не формализован и заполняется произвольно. В нем полезно перечислить признаки, предопределяющие некоторые значения данного признака. Кроме указанных сведений таблица содержит даты последних корректировок признаков.

В третьем разделе листа «Параметры проекта» находится список зарегистрированных отчетов, содержащих результаты анализа данных, которые, как правило, имеют форму таблиц или графиков, отображающих представленные в них данные. Проект может содержать множество таблиц с результатами, однако регистрировать рекомендуется только те, которые предполагается включить в сводный отчет по проекту. Остальные можно считать промежуточными.

При регистрации таблицы в списке отчетов вводится название, используемые признаки и выделяется блок размещения данных таблицы. Даты создания и корректировки определяются системой автоматически. Поле актуализации выбирается пользователем в формате «требуется», «не требуется». Оно указывает, могут или не могут автоматически обновляться результаты таблицы при изменении данных.

Четвертый раздел листа «Параметры проекта» содержит список зарегистрированных вариантов классификации. Приемлемый вариант группировки объектов выборки может быть получен исследователем в результате длительной работы, возможно с использованием внешних программ.

Классификацию, или типизацию данных удобно представлять как дополнительный признак, который может содержать числовое значение – номер класса или текстовое имя, общее для всех объектов одного класса. Повторять типизацию необходимо в двух случаях: если количество объектов выборки возрастет и при корректировке данных. Но не все корректировки данных могут потребовать актуализации классификации или расчетных таблиц. Пересчет необходим только в том случае, если при расчетах использовались признаки, в данных которых произошли изменения.

Необходимость контроля актуальности классификаций возникает, когда проект содержит множество различных ее вариантов. Регистрация признака классификации осуществляется аналогично регистрации расчетных таблиц. Совместно хранятся название классификации, дата ее регистрации (создания), место размещения, состав признаков, используемых при классификации, средство классификации (в форме комментария). Если проект не очень масштабный, пользователь может не регистрировать ни отчеты, ни классификации.

Следующим (третьим) элементом проекта (рис. 3) является лист EXCEL, содержащий даты внесения изменений в таблицу основных данных. Размерность таблицы дат изменений совпадает с размерностью таблицы исходных данных. При завершении каждого сеанса работы с проек-

том происходит обновление дат корректировки. Информация по датам изменения данных может быть очень полезна для расчета статистики по ходу выполнения проекта.

Четвертый элемент проекта – «Фильтры». На этой странице хранятся все логические фильтры, связанные с проектом. Они позволяют существенно повысить достоверность данных. На разработку таких фильтров может потребоваться достаточно длительное время, их обслуживание и автоматизация разработки осуществляется с помощью специальной программы, входящей в состав специализированного комплекса.

Листы EXCEL, содержащие результаты обработки данных, составляют пятый элемент проекта – «Отчеты».

Шестой элемент – «Изъятые данные» – содержит анкеты, которые были по тем или иным причинам изъяты из выборки.

Теперь определим понятие другого компонента специализированного программного комплекса. Модель данных – это совокупность правил, регламентирующих представление основных данных на листе проекта «Исходные данные». Она предполагает, что измеренные характеристики анкет располагаются на листе по строкам. В первой строке находятся названия признаков (или вопросы анкеты), в первом столбце таблицы – номера анкет в опросе. Целесообразно (но не обязательно) последние два столбца таблицы данных отводить под дату опроса и фамилию интервьюера.

Таблица не должна содержать никаких дополнительных строк, связанных с обработкой данных, кроме основных (например, итогов и т.п.). На листе с данными кроме анкетных характеристик могут быть представлены дополнительные столбцы с расчетами и классификациями, которые могут быть созданы только после инициализации параметров проекта. Эта модель полностью согласуется с моделями данных, принятыми в большинстве статистических пакетов, что облегчает их экспорт и импорт во внешнюю программную среду.

Понятие проекта анкетного опроса и модели данных являются основой разрабатываемой компьютерной технологии обработки анкет. Расширение возможностей исследователя достигается за счет использования компьютерных программ, реализующих конкретные алгоритмы обработки данных. Алгоритмы нацелены на решение проблем, возникающих при анализе анкетных характеристик. Разрешение таких проблем либо невозможно с помощью известных программных средств, либо крайне затруднительно.

Эффективность обработки анкетных данных повышается не только за счет использования отдельных модулей, но и за счет совместного их использования. Программный комплекс охватывает все основные стадии обработки собранной информации – от подготовительного этапа до визуального представления результатов.

Каждый из четырех разделов комплекса связан с одной из проблем, возникающих в практической работе с анкетными данными.

Первая проблема – отсутствие ответов на конкретные вопросы анкеты – обсуждается в некоторых работах [6, 10]. Для ее успешного решения необходимо понять причину, которая могла вызвать отсутст-

вие ответа, а затем разрабатывать методы устранения ошибок. Методы восстановления пропущенных данных обсуждаются также в работах [2, 7].

На достоверность статистического вывода может повлиять проблема засоренности данных, которая проявляется в наличии явных выбросов и логически непоследовательных ответов.

Для решения двух указанных проблем (отсутствие данных и засоренность) существуют методы проверки состоятельности данных [7]. Они представлены блоком «Алгоритмы повышения достоверности анкетных данных» (рис. 4). Этот блок состоит из 2-х типов алгоритмов: статистических и логических.

Хотя обе выделенные проблемы в литературе отмечаются как важнейшие, их решение не нашло отражения в современных пакетах анализа статистических данных. Какие-либо методики, охватывающие проблему в целом, тоже отсутствуют. Между тем решение этих проблем не является тривиальным. Во-первых, речь идет о многомерных выборках, во-вторых, признаки могут быть представлены в различных шкалах. Сложность их решения многократно возрастает при большом количестве вопросов анкет и массовых опросах. В этом случае необходимо не просто разработать алгоритмы и программы, но и создать специальные технологии решения задач.

Рис. 4. Основные разделы специализированного программного комплекса

Следующая проблема, которая стоит перед исследователем анкетных данных, это создание типологий и классификация многомерных данных. Если модули по кластерному анализу присутствуют в специализированных статистических пакетах, то алгоритмы построения типологий в пакетах отсутствуют. Однако и содержащиеся в пакетах программы классификации имеют весьма ограниченное применение для обработки анкетных данных, поскольку они работают с числовыми признаками, измеренными в шкале отношений. В реальной анкетной информации такие признаки встречаются достаточно редко. Чаще всего пространство признаков представлено разнородными шкалами. В этом случае может оказаться весьма полезным непараметрический алгоритм классификации, рассмотренный и реализованный в таком комплексе.

Вспомогательные средства анализа данных имеют исключительно важное значение при решении практических задач. Не случайно большинство средств анализа, реализованных в статистических пакетах, относятся к этой категории. Различные формы представления инфомации весьма важны для формирования содержательных гипотез о структуре и взаимосвязи данных.

«Процедуры преобразования данных» и «Сервисные и визуальные средства анализа данных» – два блока вспомогательных средств (рис. 4). Ряд процедур преобразования можно найти в пакетах по обработке информации. Наш программный комплекс содержит ряд нестандартных преобразований. Те же, которые можно найти в пакетах, у нас имеют расширенные функции и обеспечивают расчет дополнительных характеристик. Кроме того, эффективность их работы в рамках проекта выше, чем стандартных средств.

Сервисные и визуальные средства, представленные в комплексе программ, являются проблемно-ориентированными и дают исследователю дополнительные возможности анализа специфических анкетных характеристик.

Комплекс программ включает также блок моделирования многомерных выборок, полезный при исследовании возможностей собственных и тестировании внешних программ. Большинство статистических пакетов содержат генераторы данных, однако все они рассчитаны на моделирование только одного признака и поэтому для тестирования программ многомерного анализа не пригодны.

Рассмотренная структурная схема является основой новой, более эффективной компьютерной технологии обработки анкетных данных, призванной расширить спектр решаемых задач.

Литература

1. Айвазян С.А. Прикладная статистика. Классификация и снижение размерности / С.А. Айвазян, В.М. Бухштабер, И.С. Енюков. – М.: Финансы и статистика, 1989. – 422 с.
2. Адамов С.Ю. Система анализа нечисловой информации «САНИ» / С.Ю. Адамов // Социология: 4М (методология, методика, математическое моделирование). 1991. № 2. С. 86–104.

3. Анализ нечисловой информации в социологических исследованиях / под ред. В.Г. Андреенкова, А.И. Орлова, Ю.Н. Толстовой. – М.: Наука, 1985. – 220 с.
4. Анфилатов В.С. Системный анализ в управлении / В.С. Анфилатов. – М.: Новый век, 2003. – 368 с.
5. Березин И.С. Проведение массовых опросов / И.С. Березин // Маркетинг и маркетинговые исследования в России. 1996. № 5.
6. Клюшина Н.А. Причины, вызывающие отказ от ответа / Н.А. Клюшина // Социол. исслед. 1990. № 1. С. 98–105.
7. Малхотра Н.К. Маркетинговые исследования / Н.К. Малхотра. – М.: Вильямс, 2002. – 960 с.
8. Мартышенко Н.С. Методическое обеспечение анализа поведения потребителей на региональном туристском рынке / Н.С. Мартышенко // Вестник ТГЭУ. 2005. № 4. С. 19–31.
9. Орлов А.И. Нечисловая статистика / А.И. Орлов. – М.: МЗ-Пресс, 2004. – 513 с.
10. Толстова Ю.Н. Анализ социологических данных. Методология, дескриптивная статистика, изучение связей между номинальными признаками / Ю.Н. Толстова. – М.: Научный мир, 2000. – 352 с.
11. Черчилль Г.А. Маркетинговые исследования / Г.А. Черчилль. – СПб.: Питер, 2002. – 752 с.