На правах рукописи

Кустов Дмитрий Александрович

РАЗРАБОТКА И АНАЛИЗ АЛГОРИТМОВ ОБРАБОТКИ АНКЕТНЫХ ДАННЫХ

Специальность 05.13.18 – Математическое моделирование,
численные методы и комплексы программ

Автореферат
диссертации на соискание ученой степени
кандидата технических наук

Владивосток
2007

Работа выполнена на кафедре математики и моделирования Владивостокского государственного университета экономики и сервиса

Научный руководитель: кандидат технических наук, профессор
Мартышенко Сергей Николаевич

Официальные оппоненты: доктор физико-математических наук, профессор
Цициашвили Гурам Шалвович

кандидат технических наук, доцент
Глушков Сергей Витальевич

Ведущая организация: Дальневосточный государственный технический университет (г. Владивосток)

Защита состоится «31» мая 2007 г. в 11^{30} часов на заседании диссертационного совета Д 005.007.01 в Институте автоматики и процессов управления ДВО РАН по адресу: 690041, г. Владивосток, ул. Радио, 5.

С диссертацией можно ознакомиться в библиотеке ИАПУ ДВО РАН.

Автореферат разослан «28» апреля 2007 г.

Ученый секретарь
диссертационного совета Д 005.007.01 А.В. Лебедев

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность исследования. Для повышения уровня обоснованности управленческих решений на всех уровнях экономики требуется качественная и достоверная информация. Одним из основных источников первичных данных в экономических и социологических исследованиях служат данные анкетных опросов.

Если для нас область исследования социально–экономических процессов на основе анкетного опроса является достаточно молодым направлением, то в странах с развитой рыночной экономикой это научное направление уже давно перешло в ранг классического знания. Здесь можно выделить таких известных зарубежных ученых как Ф. Котлер, Р. Блэкуэлл, Д. Хокинс, Р. Бест, Г. Ассэль, Х. Беркман, К. Хаксевер, Б. Рендер Ж.-Ж. Ламбен, Н. Малхотра, Дж. О′Шонесси и др.

В работах отечественных ученых, специализирующихся в области маркетинговых исследований, также уделяется внимание вопросам сбора и обработки первичных данных. Здесь можно назвать таких авторов как А.В. Алешина, Г.Л. Багиев, И.С. Белявский, Е.П. Голубков, Л.А. Козырев, С.Г. Светуньков, С.Г. Токарев. Однако большинство работ носит концептуальный теоретический характер, а не методический.

Вопросами изучения социально–экономических явлений методами анкетного опроса в нашей стране больше занимались ученые в области социологии. Среди них можно назвать таких исследователей, как И.С. Березин, С.Н. Григорьев, О.Ю. Ермолаев, А.Н. Кричевец, О.Н. Маслова, Ю.Н. Толстова, В.А. Ядов, Г.Г. Татарова, Г.И. Саганенко.

Следует отметить, что методы обработки данных не разрабатываются ни маркетологами, ни социологами, скорее всего этих специалистов можно отнести к заказчикам теоретических изысканий в области статистических методов и в особенности такого ее раздела как многомерный статистический анализ. В этой области давно и успешно работают такие известные отечественные ученые как С.А Айвазян, А.А. Боровков, И.И. Елесеева, И.С. Енюков, Б.Г. Миркина, Г.С. Лбов, Л.А. Сошникова, А.И. Орлов, Ю.Н. Тюрин.

Потребность широкого круга исследователей в результатах анализа данных и наличия методов еще не решает проблемы. Необходимы также и средства анализа, воплощенные в конкретных компьютерных технологиях. Здесь практика сталкивается с большим дефицитом таких средств.

Если ранее на рынке программных продуктов еще присутствовали некоторые отечественные пакеты, обрабатывающие статистические данные, то теперь они почти сошли со сцены, а новые не разрабатываются. Присутствующие же на рынке зарубежные пакеты по обработке информации не обеспечивают решение всего спектра задач анализа анкетных данных, так как они больше приспособлены для применения классических статистических методов анализа к данным числовой природы и требуют некоторых идеализированных данных. Данные анкетных опросов, как правило, не удовлетворяют этим требованиям.

Анкетные данные по своей природе содержат ошибку, которая складывается из множества составляющих. В отдельных наблюдениях уровень ошибки может быть не просто высок, но и достигать абсурдных значений. Поэтому прежде чем использовать анкетные данные для анализа исследуемых объектов и явлений, необходимо произвести серьезную подготовительную работу по оценке качества собранного материала.

Поэтому исследование, направленное на развитие методов повышения достоверности данных и разработку инструментальных средств обработки больших статистических выборок анкетных данных, является актуальным.

Актуальность проведенного исследования подтверждается и тем, что диссертационная работа выполнялась в рамках научно-исследовательской работы «Исследование взаимодействия в системе "биологический объект — внешняя среда" на основе моделирования и обработки данных статистики в условиях ограниченности и неопределенности исходной информации» (грант РФФИ — ДВО РАН № 06-05-96017) и научно-исследовательской работы «Построение математических моделей этнических миграций на примере переселения корейцев из районов Центральной Азии на Дальний Восток России в 90–е годы XX века» (грант РФФИ — ДВО РАН № 06-06-96004).

Целью диссертационной работы являются разработка и исследование методов и алгоритмов анализа многомерных статистических данных, полученных методом анкетного опроса и характеризующих состояние сложных социально–экономических систем, а также их реализация в виде комплекса программных средств.

В соответствии с поставленной целью в диссертации решались **следующие задачи:**

− обобщить существующие в отечественной и зарубежной теории и практике методические подходы и инструментальные средства анализа многомерных статистических данных;

− разработать методы и алгоритмы повышения достоверности анкетных данных;

− определить новые области приложения методов многомерной классификации признаков нечисловой природы;

− разработать компьютерную технологию анализа больших статистических выборок;

− реализовать предложенные в работе теоретические положения анализа данных в виде специализированного комплекса программных средств;

− исследовать эффективность разработанной системы анализа данных;

− разработать методику использования новых инструментальных средств для решения практических задач.

Объектом диссертационного исследования являются социально–экономические группы населения.

Предметом исследования являются многомерные статистические данные, характеризующие социально–экономические процессы и явления, полученные методом анкетного опроса.

Методы исследования. При выполнении диссертационной работы использовался системный анализ, общенаучные методы исследования (сравнение, анализ и синтез, индукция и дедукция, аналогия), методы многомерного анализа и моделирования, что позволило обеспечить достоверность результатов исследования и обоснованность выводов.

Информационной базой диссертационного исследования послужили материалы конференций и специальных периодических изданий, официальные документы и статистическая отчетность Комитета государственной статистики РФ, Приморского краевого комитета государственной статистики, данные анкетных опросов, предоставленные канд. экон. наук, доцентом кафедры маркетинга и коммерции ВГУЭС Н.С. Мартышенко, а также первичные данные, собранные и обработанные в процессе выполнения диссертационной работы.

Научная новизна проведенного исследования заключается в следующем:

− разработана и программно реализована серия статистических и логических алгоритмов повышения качества данных анкетных опросов;

− предложены и реализованы новые подходы использования алгоритмов многомерной классификации и распознавания нечисловых признаков в задачах восстановления данных и исследования структур данных в социально-экономических исследованиях;

− предложены и программно реализованы методы преобразования и обработки открытых вопросов анкетных данных;

− на основе системного анализа задач, решаемых по данным анкетных опросов, предложены новые подходы формализации и компьютерного представления пакетов анкетных данных, позволяющие разрабатывать компьютерные технологии их обработки.

Практическая ценность работы. Полученные в диссертации результаты составляют алгоритмическую и программную основу для создания нового класса систем обработки анкетных данных. Разработаны программные средства сопровождения крупных проектов по исследованию социально-экономических систем методом анкетного опроса. Программные средства и методика, полученные в результате проведения диссертационного исследования, могут быть использованы широким кругом исследователей-практиков, использующих данные анкетных опросов для обоснования управленческих решений.

Материалы диссертационной работы используются в учебном процессе Института международного бизнеса и экономики Владивостокского государственного университета экономики и сервиса. Комплекс программ на основе разработанных, программно реализованных и исследованных в работе алгоритмов обработки анкетных данных был внедрен в туристических компаниях города Владивостока, а также в научно–исследовательских лабораториях Владивостокского государственного университета экономии и сервиса. По фактам внедрения составлено четыре акта внедрения.

На защиту выносятся:

1. Концепция обработки анкетных данных в виде единого технологического проекта с определением собственной модели данных и заданной структурой хранения информации.
2. Статистические алгоритмы выявления грубых ошибок в многомерных анкетных данных, которые позволяют упорядочить их в соответствии с заданными критериями, полученными в результате обобщения и формализации действий экспертов по выявлению ошибок в анкетных данных.
3. Логические алгоритмы выявления грубых ошибок в многомерных анкетных данных.
4. Метод и реализующий его алгоритм обработки открытых и составных открытых вопросов, расширяющий пространство признаков, используемых для формирования статистических выводов при анализе анкетных данных.
5. Принципы решения задач повышения качества анкетных данных на основе применения непараметрического алгоритма интегральной диагностики.

Апробация работы. Основные результаты докладывались на научных конференциях: Всероссийская научно-практическая конференция «Информационные технологии в управлении и учебном процессе вуза» (Владивосток, 2002), Международная конференция студентов, аспирантов и молодых ученых (Владивосток, 2005, 2006), Региональная научно-техническая конференция «Молодежь и научно-технический прогресс» (Владивосток, 2006), Международная научно–практическая конференция «Компьютерные технологии в науке, производстве, социальных и экономических процессах» (Новочеркасск, 2006), Международная научно-практическая конференция «Управление в социальных и экономических системах» (Пенза, 2006), Международная открытая научная конференция «Современные проблемы информатизации» (Воронеж, 2007).

Публикации по теме диссертации. По основным результатам, полученным в диссертационной работе, опубликовано 12 печатных работ.

Структура и объем диссертации. Диссертационная работа состоит из введения, трех глав, заключения, списка литературы, включающего 136 наименований, и 9 приложений. Основной текст диссертации изложен на 146 страницах машинописного текста, включает 40 рисунков и 18 таблиц.

СОДЕРЖАНИЕ РАБОТЫ

Во введении обосновывается актуальность темы исследования, формулируются цель и задачи исследования, определяются объект, предмет и методы исследования, раскрываются новизна и практическая значимость работы, приводятся сведения об апробации и реализации основных положений диссертации.

В первой главе произведен анализ современных методов и средств обработки анкетных данных и рассмотрены специфические особенности дан-

ных анкетных опросов, являющихся основным источником первичной информации в экономических и социологических исследованиях.

Анкетный опрос производится с целью получения числовых характеристик, описывающих структуру и реакцию на сложившиеся внешние условия исследуемых совокупностей населения.

В России анкетные опросы пока не приобрели столь массового характера, как в странах с развитой рыночной экономикой. В работе представлен анализ причин, которые сдерживают более широкое использование данных опросов для решения практических задач.

Опрос представляет собой некоторый специфический способ измерения. Специфика этого способа измерения состоит в высокой степени неопределенности оценок достоверности данных. Неопределенность обусловлена тем, что данные имеют множество источников ошибки (рис. 1).

Повышение достоверности данных лежит на пути использования системного подхода при разработке методик анализа анкетных данных. Система сбора данных должна быть неотрывно связана с системой обработки и составлять единый технологический комплекс.

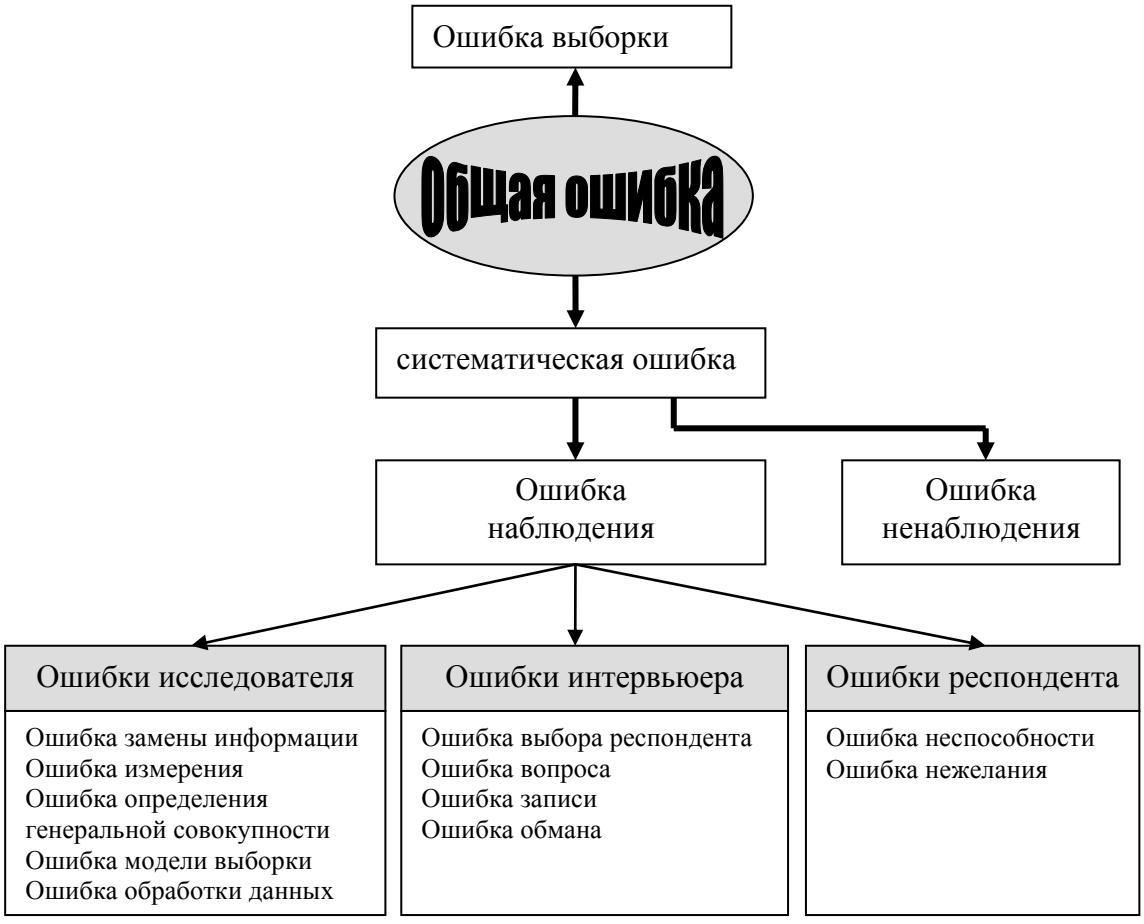

Рис. 1. Источники ошибок при проведении анкетного опроса

В связи с этим в работе приводится анализ современных пакетов по обработке статистических данных, из которого следует, что в настоящее время ощущается острый дефицит специальных программных средств, учитываю-

щих специфику анкетных данных. Снизить остроту проблемы и обеспечить в сжатый срок практиков столь необходимыми средствами можно расширением возможностей широко используемых средств обработки данных, таких, как EXCEL.

Дополнительные инструментальные средства должны в первую очередь обеспечить решение задач, которые не представлены в универсальных пакетах анализа данных. С целью выявления таких задач в работе производится анализ проблем, возникающих при обработке данных анкетных опросов в исследованиях социально–экономических процессов. Эти проблемы связаны с особенностями анкетных данных.

Анкетные данные включают признаки различной природы, в виде количественных и качественных характеристик изучаемых объектов и явлений, и могут содержать значительное количество пропусков. При масштабных опросах на точность данных может оказать большое влияние личность интервьюеров. Поэтому использование многомерных методов анализа данных требует предварительного исследования достоверности данных и разработки методик их восстановления. Для профессиональной обработки данных требуется разработка специальных компьютерных технологий.

Во второй главе рассматриваются методы и алгоритмы повышения достоверности анкетных данных.

Предложенные методы анализа качества данных были разработаны на основании обобщения и формализации процедур углубленного качественного анализа достоверности данных, производимого экспертами, обладающими обширными знаниями по исследуемым объектам и явлениям. Невозможно разработать единого метода выявления грубых ошибок, поскольку это понятие имеет нечеткое определение, зависящее от содержательного смысла признаков и целей решения задач анализа данных. Необходимость разработки набора инструментальных средств анализа качества данных обусловлена еще и тем, что признаки различной природы требуют своих методов обработки.

Предложенные методы объединяет то, что они используют единый подход к анализу и принятию решения о возможных действиях по снижению уровня выделенных ошибок. Алгоритмы работают по принципу многомерных фильтров, упорядочивающих многомерные данные в соответствии с установленными критериями.

В работе предложены две группы алгоритмов выявления грубых ошибок: статистические фильтры и логические фильтры. К группе статистических методов отнесены и алгоритмы выделения многомерных данных по количеству и качеству пропусков.

Далее в работе подробно рассмотрены семь **статистических алгоритмов выявления ошибок**:

1. **Фильтр отсутствия данных (ФОД),** выполняющий анализ анкетных данных на отсутствие данных и выделяющий анкеты, которые содержат наибольшее количество значений отсутствия данных.

2. **Фильтр экстремальных непрерывных значений (ФЭНЗ),** основанный на предположении о том, что если пакет данных (набор данных, представленный конкретным интервьюером) содержит недостоверную информацию, то большие отклонения от средних значений будут не только по одному признаку, но и по другим признакам.

3. **Фильтр ранжирования непрерывных значений (ФРНЗ),** основанный на ранжировании отклонений от среднего значения признака. Этот фильтр дает неплохие результаты, когда нет тесной связи между признаками, а при большом количестве признаков такой зависимости, как привило, не обнаруживается.

4. **Фильтр метрических непрерывных значений (ФМНЗ),** основанный на подсчете расстояния от объекта до центра выборки с использованием известных метрик Евклида, Хэмминга и Махаланобиса.

5. **Фильтр частот кодированных значений (ФЧКЗ)**, основанный на сравнении частотных рядов по каждому признаку. Этот фильтр используется для признаков, измеренных в ранговых и номинальных шкалах.

6. **Фильтр замены кодированных значений (ФЗКЗ),** основанный на том, что в номинальной шкале каждый признак может быть представлен некоторым ограниченным списком значений. Рассчитав частоту встречаемости каждого значения, можно произвести замену значений наблюдения на частоты встречаемости данного значения и рассчитать среднее значение частоты.

7. **Фильтр эталонных значений (ФЭЗ),** основанный на использовании идей распознавания образов. Пакет анкет, обеспечивающий наименьшую ошибку распознавания, можно считать более обособленным и, следовательно, он рассматривается как аномалия, требующая содержательного анализа.

В программной реализации каждый алгоритм представлен двумя модулями. Один модуль служит для обработки отдельных наблюдений, второй — для обработки пакетов анкет, представленных различными интервьюерами. При пакетной обработке в название фильтра добавляется буква «Г» (групповой).

В качестве примера статистического фильтра рассмотрим принцип работы «Фильтра экстремальных непрерывных значений группового» (ФЭНЗГ), предназначенного для работы с признаками, измеренными в шкале отношений. Этот фильтр создан для работы с пакетами данных различных интервьюеров.

Запишем одно наблюдение из r-го пакета набором m значений:

$$x_{i_r} = (x_{i_r 1}, x_{i_r 2}, \ldots, x_{i_r j}, \ldots, x_{i_r m}), \qquad (1)$$

считая, что все m признаков измерены в шкале отношений. Тогда задача состоит в том, чтобы из k пакетов анкет выделить пакет, который имеет наибольшие отличия от остальных пакетов.

Для этого последовательно для каждого пакета r (r=1, 2, …, k) повторяется следующая процедура: рассчитываются средние значения m признаков по выборке за исключением пакета с номером r:

$$\bar{X}^{-r} = (\bar{x}_1^{-r}, \bar{x}_2^{-r}, \ldots, \bar{x}_j^{-r}, \ldots, \bar{x}_m^{-r}), \qquad (2)$$

и средние значения признаков по пакету с номером *r*:

$$\overline{X}^{r} = (\overline{x}_1^{\,r}, \overline{x}_2^{\,r}, ..., \overline{x}_j^{\,r}, ..., \overline{x}_m^{\,r}). \quad (3)$$

Вычисляются поэлементные модули разностей двух векторов средних:

$$\lambda_r = \left| \overline{X}^{-r} - \overline{X}^{r} \right|. \quad (4)$$

Объединяем все отклонения λ_r в одну матрицу λ размерности $k \times m$. На основании матрицы λ рассчитаем матрицу *M* той же размерности. Вычисления производятся по схеме: определяется максимум в каждом столбце матрицы λ, затем элементу матрицы *M*, соответствующему значению максимума, присваивается значение единицы, всем остальным элементам матрицы *M* присваивается значение ноль. В результате построчного суммирования элементов матрицы *M* получим вектор оценок для каждого интервьюера:

$$\mu^{\text{м}} = (\mu_1^{\text{м}}, \mu_2^{\text{м}}, ..., \mu_r^{\text{м}}, ..., \mu_k^{\text{м}}). \quad (5)$$

Интервьюер с наибольшим значением $\mu^{\text{м}}$ будет иметь максимальный штраф, и поэтому его данные могут быть поставлены под сомнение. Теперь исследователь может сосредоточить свое внимание на отдельном пакете первичных данных, подвергнуть их дополнительному содержательному анализу, в результате которого он определяет, является отклонение допустимым или нет. Это достаточно грубый фильтр. Он основан на предположении о том, что если пакет анкет содержит недостоверную информацию, то большие отклонения от средних значений будут не только по одному признаку, но и по другим.

В программной реализации данный фильтр допускает применение двух вариантов весовых коэффициентов признаков. В первом случае учитывается наличие нескольких значений признака, соответствующих максимальному значению. В этом случае вначале рассчитываются коэффициенты q_j по формуле:

$$q_j = \frac{\sum_{j=1}^{m} \gamma_j}{\gamma_j}, \quad (6)$$

где γ_j – количество значений, равных максимальному значению для признака с номером j в исходной матрице данных.

Весовые коэффициенты Q_j получаются путем нормировки коэффициентов q_j:

$$Q_j = \frac{q_j}{\sum_{j=1}^{n} q_j}. \quad (7)$$

Вектор оценок интервьюеров с учетом весов, будет равен:

$$\mu^{\text{Ml}} = (\mu_1^{\text{Ml}}, \mu_2^{\text{Ml}}, ..., \mu_r^{\text{Ml}}, ..., \mu_k^{\text{Ml}}). \quad (8)$$

Второй вариант расчета весовых коэффициентов предполагает предварительную модульную нормализацию признаков. Другими словами, исходные значения признаков x_{ij} преобразуются к виду:

$$x'_{ij} = \frac{\left| x_{ij} - \overline{X}_j \right|}{\sqrt{S_j^{\,2}}}, \qquad (9)$$

где $\overline{X}_j$ – среднее значение признака с номером j;

$S^2{}_j$ – оценка дисперсии признака с номером j.

Тогда графическую интерпретацию вводимого коэффициента, на примере двух признаков, можно изобразить рис. 2.

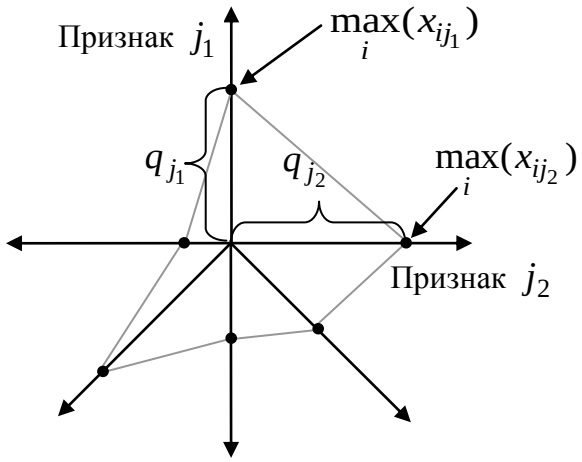

Рис. 2. Графическая интерпретация весового коэффициента

Далее в работе рассмотрены два **логических алгоритма повышения качества данных. Первый логический алгоритм** предназначен для разработки типологий по качественным признакам, полученным как ответы на открытый вопрос. Введено понятие составного признака, для которого определены три варианта расчета частотных рядов простых значений. Применение этого алгоритма позволяет расширить пространство признаков за счет новых признаков, полученных в результате обработки качественной информации. Такие признаки в анкетах, как правило, не обрабатываются ввиду отсутствия методики средств их обработки. Между тем эти признаки могут оказаться очень информативными, поскольку респондентам не навязывается жесткая схема ответа.

Второй логический алгоритм предназначен для выявления логических противоречий в многомерных данных, которые плохо улавливаются статистическими фильтрами. Алгоритм основан на содержательном анализе признаков.

Оба алгоритма позволяют аккумулировать знания и опыт, полученные в ходе работы над проектом анализа анкетного опроса. Отличие этих алгоритмов состоит в активном участии исследователя в процессе работы программ. Такие алгоритмы зависят от возможностей программной среды, в которой они реализованы. В нашем случае в процессе работы с программами

пользователь может использовать весь арсенал средств обработки данных, предоставляемых EXCEL.

Разработанные алгоритмы и реализующие их программные модули выявления грубых ошибок прошли апробацию на нескольких крупных проектах анкетных опросов и показали высокую эффективность по выявлению выбросов, которые исследователи не могли обнаружить при сверке данных и использовании традиционных одномерных методов анализа качества данных.

Далее в главе рассматриваются задачи обработки анкетных данных, которые могут быть решены с использованием **алгоритмов многомерной классификации и распознавания образов**. При этом выделяются четыре задачи:

– задача выявления выбросов или грубых ошибок;

– задача восстановления данных;

– задача выделения однородных групп объектов (классификация);

– задача прогнозирования признаков (распознавание по обучающей выборке).

Для их решения предлагается использовать непараметрический алгоритм интегральной диагностики, который ранее использовался только в технических системах. Преимущество алгоритма состоит в том, что он может работать с признаками различной природы. Рассматривается общая схема работы алгоритма. Принцип работы алгоритма состоит в разработке эталонов классов по многомерной обучающей выборке. При решении перечисленных задач используются свои способы формирования обучающей и контрольной выборок. Поэтому для каждой задачи разработаны свои программные модули, ориентированные на конкретные задачи. В работе рассмотрены особенности применения алгоритма при решении этих задач.

Все программы, реализующие алгоритмы, рассмотренные в главе, представлены в виде единого комплекса программных средств, выполненного в виде приложения EXCEL.

В третьей главе рассматриваются **структура и принципы построения специализированного комплекса программных средств обработки анкетных данных**. Компьютерная технология анализа данных строится на принципах системного подхода к анализу анкетных данных, который начинается от формулировки целей исследования и заканчивается формулировкой содержательных выводов (рис. 3).

Разработанный программный комплекс основан на определении понятий «проекта анкетного опроса» и «модель данных опроса», которые приводят к определенным правилам компьютерного представления информации и доступа к программам комплекса. Структура проекта включает семь элементов: исходные данные по анкетному опросу, параметры проекта, даты изменений, фильтры, словари замены, отчеты, изъятые данные. В работе обсуждается содержание и назначение этих элементов.

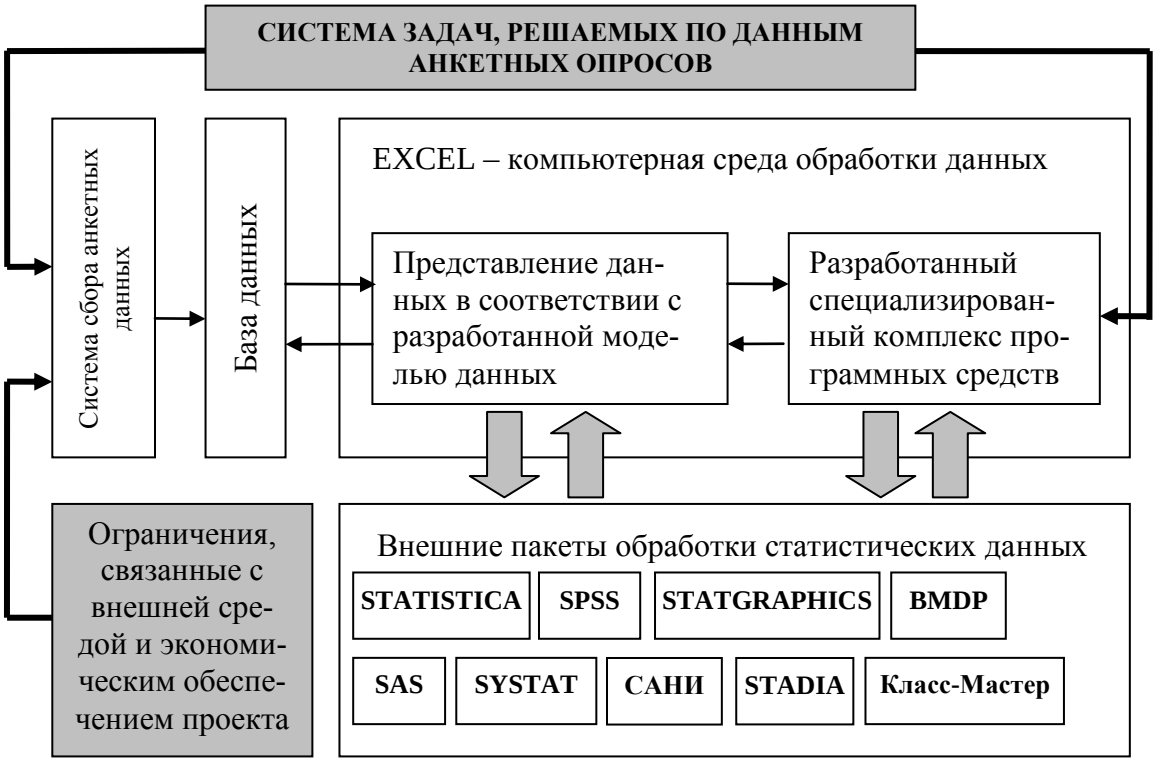

Рис. 3. Компьютерная технология анализа анкетных данных

Отдельные модули разработанного программного комплекса объединены в четыре раздела по функциональному признаку (рис. 4). В работе подробно рассмотрены функции каждого из разделов.

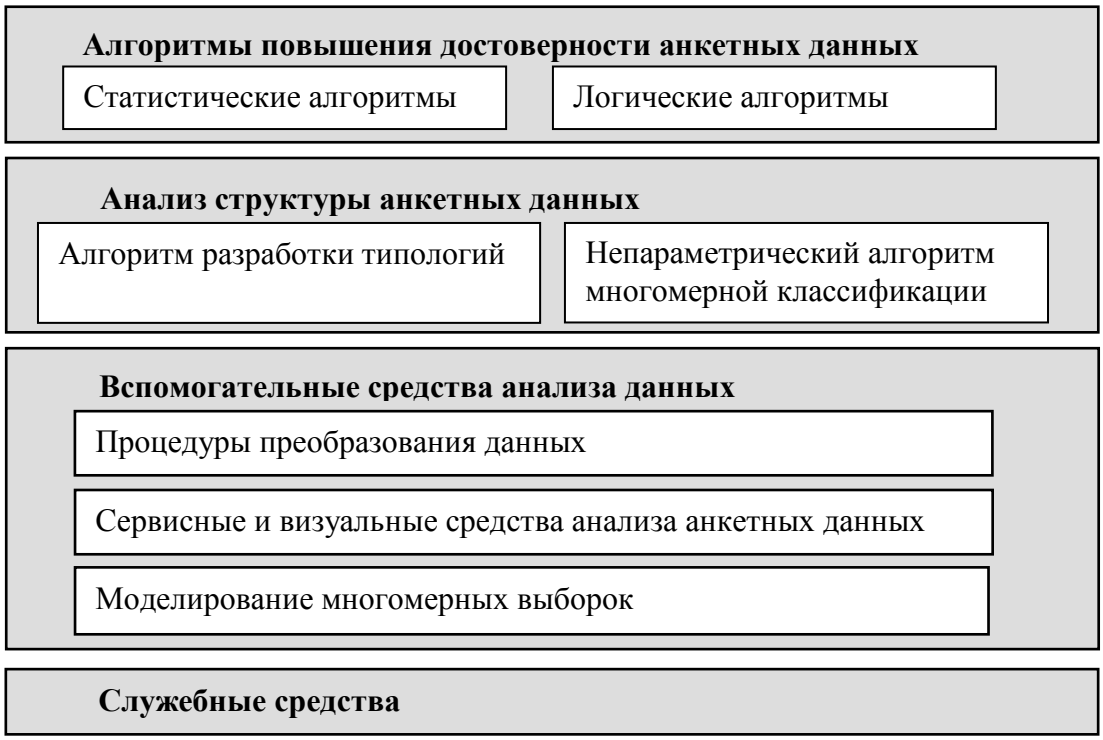

Рис. 4. Основные разделы программного комплекса

Далее в работе рассматриваются особенности работы с программами комплекса. Приводятся результаты расчетов и апробации программ на реальных данных. В частности, приводятся графики изменения критериев, полученные с помощью различных статистических фильтров, как, например, график расчета критериев ФОД (рис. 5).

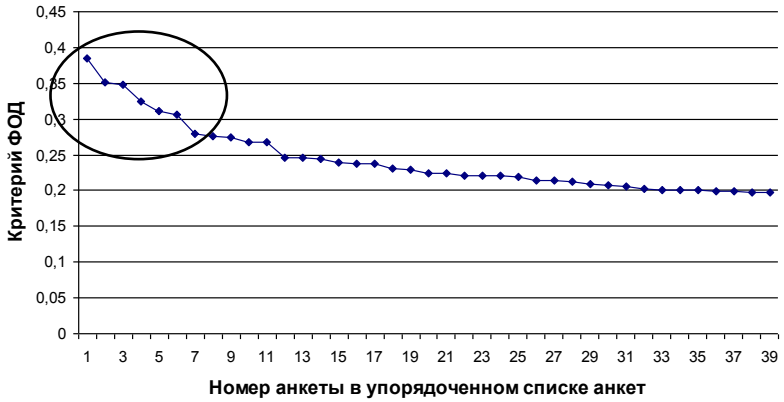

Рис. 5. Значения критерия ФОД

Приводится сравнительный анализ результатов полученных с помощью различных статистических фильтров на реальных данных (рис. 6).

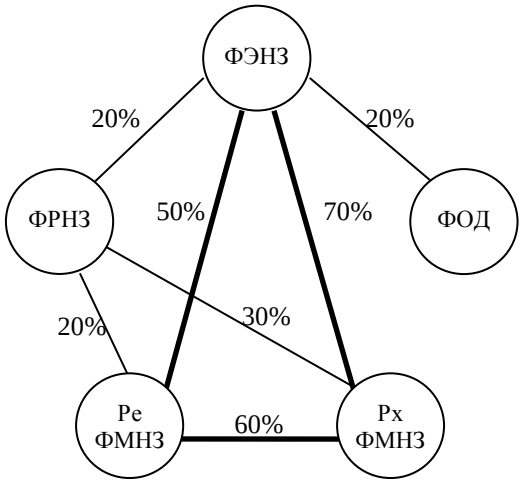

Рис. 6. Степень совпадения выявленных выбросов

Далее в работе рассматриваются вспомогательные и служебные программы разработанного программного комплекса.

Вспомогательные средства выполняют следующие задачи:

– автоматизируют работу пользователя при формировании элементов проекта в соответствии с устанавливаемыми правилами описания проекта;

– облегчают контроль целостности проекта, то есть определяют возможные нарушения в описаниях структуры проекта, ошибочно введенные пользователем;

– накапливают статистику о ходе выполнения проекта.

Вспомогательные средства объединены в три группы:

– операции преобразования данных;

– графические средства для визуализации данных;

– средства моделирования многомерных данных.

Предложенные в работе методы и программные средства являются основными элементами **новой методики анализа информации анкетных данных**. Методика охватывает все этапы исследований, начиная от составления текста анкеты до решения прикладных задач, и ориентирована на повышение качества конечного результата исследования. При этом можно выделить ряд направлений повышения качества.

Первое направление связано с постоянным совершенствованием системы сбора данных. Сбор и обработка анкетных данных рассматриваются не как разовая акция, а как многоэтапный процесс, в котором постоянно совершенствуется система сбора первичного материала. Иначе говоря после сбора порции данных в процессе обработки выясняются вопросы, ответы на которые оказываются менее достоверными или встретили затруднения при ответе. Далее необходимо определить причины, вызвавшие ошибки, и внести коррективы в анкету или систему организации сбора данных, после чего повторить опрос на новом качественном уровне. Целенаправленные действия исследователя не только подкрепляются характеристикам, но и обосновываются количественными оценками.

Второе направление связано с созданием системы описания (компьютерного представления) анкетных данных и результатов их обработки. Создание единых правил представления данных позволяет унифицировать разработку программного обеспечения и способов доступа пользователей к модулям программного комплекса. Данный подход позволяет выполнить сопровождение большого количества анкетных опросов в течение длительного времени, постоянно пополняя данные. Систематизация данных в рамках единой системы позволяет избежать многих ошибок технического и организационного характера.

Третье направление связано с повышением качества за счет обнаружения ошибок и восстановления данных. Эти функции выполняют основные модули программного комплекса. Программные средства кроме того что снижают уровень ошибки, создают предпосылки применения многомерных статистических методов (качество за счет методов).

Четвертое направление**.** Учет специфики данных позволяет выстроить действия в технологию, что в реальной ситуации существенно сокращает сроки обработки больших массивов информации.

Пятое направление. Накопление и хранение положительного опыта (память) в виде словарей корректировок данных и логических фильтров так же способствуют повышению технологичности и скорости выполнения многих действий над данными.

Шестое направление. Расширение спектра обрабатываемых вопросов за счет компьютерных технологий обработки открытых и составных открытых вопросов.

Седьмое направление. Автоматизация процедур преобразования данных и переход от одной шкалы измерения к другой. Расширение порядкового

пространства часто может позволить увеличить точность восстановления значений признака по многомерной выборке.

Совершенствование методики анализа данных заключается в изменении схемы анализа данных, что позволяет добиться новых результатов и существенно сократить сроки и трудозатраты на проведение исследований.

В заключении приводятся основные выводы и результаты, полученные в процессе диссертационного исследования.

В приложении 1 представлена сравнительная характеристика методов проведения опроса.

В приложении 2 представлены результаты сравнения имеющихся на рынке статистических пакетов прикладных программ.

В приложении 3 представлены элементы интерфейса основных модулей разработанного программного комплекса.

В приложении 4 представлены элементы интерфейса служебных и вспомогательных модулей разработанного программного комплекса.

В приложениях 5 и 6 приведены бланки анкет реально проведенных исследований, данные которых обрабатывались с помощью разработанного программного комплекса.

В приложении 7 приведены результаты работы программного модуля по обработке сложного составного признака на примере реальных анкетных данных.

В приложении 8 приведен пример использования разработанного логического алгоритма для некоторых реальных анкетных данных.

В приложение 9 включены документы по внедрению результатов.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ РАБОТЫ

1. Выявлены основные проблемы, сдерживающие применение для обработки анкетных данных методов многомерного статистического анализа. К числу таких проблем относятся: наличие пропусков в данных (отсутствие данных), наличие грубых ошибок (выбросов) и логических противоречий, разнообразие типов признаков и преобладание признаков нечисловой природы.
2. Разработана концепция обработки анкетных данных в виде единого технологического проекта заданной структуры с собственной моделью данных.
3. Разработаны, исследованы и программно реализованы статистические алгоритмы выявления грубых ошибок в многомерных анкетных данных, которые позволяют упорядочить их в соответствии с заданными критериями, полученными в результате обобщения и формализации действий экспертов по выявлению ошибок в анкетных данных.
4. Разработаны, исследованы и программно реализованы логические алгоритмы выявления грубых ошибок в многомерных анкетных данных. Они основаны на накоплении и формализации встречающихся или прогнози-

руемых логических противоречий для конкретных выборок анкетных данных.

5. Предложена и программно реализована методика обработки открытых и составных открытых вопросов, расширяющая пространство признаков, используемых для формирования статистических выводов при анализе анкетных данных.
6. Выработаны принципы решения задач повышения качества анкетных данных на основе применения непараметрического алгоритма интегральной диагностики, ранее не используемого для решения таких задач. На их основе разработан и исследован комплекс программных средств.

ПУБЛИКАЦИИ ПО ТЕМЕ ДИССЕРТАЦИИ

1. Мартышенко С.Н., Мартышенко Н.С., Кустов Д.А. Многомерные статистические методы повышения достоверности маркетинговых данных // Практический маркетинг. 2007. № 1 (119). С. 20–30.
2. Мартышенко С.Н., Мартышенко Н.С., Кустов Д.А. Применение алгоритмов многомерной классификации распознавания образов в решении задач анализа анкетных данных // Известия высших учебных заведений. Поволжский регион. Сер. Технические науки. 2006. № 6. С. 101–109.
3. Мартышенко С.Н., Мартышенко Н.С., Кустов Д.А. Цензурирование при обработке анкетных данных // Известия высших учебных заведений. Поволжский регион. Сер. Технические науки. 2006. № 6. С. 185–192.
4. Кустов Д.А., Мартышенко Н.С. Методика повышения достоверности анкетных данных // Интеллектуальный потенциал вузов – на развитие Дальневосточного региона России: Материалы VIII Международной конференции аспирантов и молодых ученых. 24–26 мая 2006 г.: В 6 кн.: Кн.2. – Владивосток: Изд-во ВГУЭС, 2006. С. 24–39.
5. Кустов Д.А., Мартышенко С.Н. Применение методов многомерной классификации для анализа данных анкетных опросов // Интеллектуальный потенциал вузов – на развитие Дальневосточного региона России: Материалы VIII Международной конференции аспирантов и молодых ученых. 24–26 мая 2006 г.: В 6 кн. Кн.2. – Владивосток: Изд-во ВГУЭС, 2006. С. 39–47.
6. Мартышенко С.Н., Мартышенко Н.С., Кустов Д.А. Совершенствование математического и программного обеспечения обработки первичных данных в экономических и социологических исследованиях // Вестник ТГЭУ. 2006. № 2. С. 91–103.
7. Кустов Д.А., Мартышенко С.Н., Мартышенко Н.С. Компьютерные технологии повышения качества первичных данных в социально-экономических исследованиях // Компьютерные технологии в науке, производстве, социальных и экономических процессах: Материалы VII науч.-практ. конф., г. Новочеркасск, 17 ноября 2006 г .: В 3 ч. /Юж.-Рос. гос. техн. ун-т (НПИ). – Новочеркасск: ООО НПО «Темп», 2006. С. 32–34.

8. Кустов Д.А., Мартышенко С.Н., Мартышенко Н.С. Компьютерные технологии анализа данных в социально-экономических системах. // Управление в социальных и экономических системах: сборник статей IV Международной научно-практической конференции. – Пенза: РИО ПГСХА, 2006. С. 77–78.
9. Мартышенко С.Н., Мартышенко Н.С., Кустов Д.А. Анализ проблем, возникающих при использовании данных анкетных опросов для исследования социально-экономических процессов // Журнал научных публикаций аспирантов и докторантов. 2007. № 2. С. 83–85
10. Мартышенко С.Н., Мартышенко Н.С., Кустов Д.А. Инструментальные средства обработки данных в EXCEL // Информационные технологии моделирования и управления: научно-технический журнал. 2007. № 1 (35). С.112–120
11. Мартышенко С.Н., Мартышенко Н.С., Кустов Д.А. Информационная технология обработки первичных данных в маркетинге. // Современные проблемы информатизации: Материалы науч.-практ. конф., г. Воронеж, 2006. С. 103–109
12. Мартышенко С.Н., Мартышенко Н.С., Кустов Д.А. Средства разработки типологий по данным анкетных опросов в среде EXCEL // Академический журнал Западной Сибири. 2007. № 1. С. 75–77

Личный вклад автора. Все результаты, составляющие основное содержание диссертационной работы, получены автором самостоятельно. В совместных работах автору принадлежат следующие научные и практические результаты: в работе [1] – многомерные статистические методы повышения достоверности маркетинговых данных; [2] – алгоритм многомерной классификации распознавания образов; [3] – анализ проблем цензурирования анкетных данных и методы их решения; [4, 7, 9] – изложены результаты анализа отдельных проблем, возникающих при использовании анкетных данных, предложены и исследованы методы их решения; [5, 6, 8, 10, 11] – постановка ряда задач, проведение исследований и интерпретация результатов; [12] – алгоритм разработки типологий по данным анкетных опросов.

Кустов Дмитрий Александрович

РАЗРАБОТКА И АНАЛИЗ АЛГОРИТМОВ ОБРАБОТКИ АНКЕТНЫХ ДАННЫХ

Автореферат

Лицензия на издательскую деятельность ИД № 03816 от 22.01.2001

Подписано в печать 22.04.2007. Формат 60×84 1/16.
Бумага писчая. Печать офсетная. Усл. печ. л. 1,2.
Уч.-изд. л. 1,5. Тираж 120 экз. Заказ

Издательство Владивостокского государственного университета
экономики и сервиса
690600, Владивосток, ул. Гоголя, 41
Отпечатано в типографии ВГУЭС
690600, Владивосток, ул. Державина, 57